SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 1–24

DEGENERATE VARIANCE CONTROL OF A ONE-DIMENSIONAL DIFFUSION*

DANIEL OCONE† AND ANANDA WEERASINGHE‡

Abstract. Consider an Itô equation for a scalar-valued process that is controlled through a dynamic and adaptive choice of its diffusion coefficient. Such a control is called a variance control and is said to degenerate when it becomes zero. We consider the problem of choosing a control to minimize a discounted, infinite-horizon cost that penalizes state values close to an equilibrium point of the drift and also imposes a control cost. Admissible controls are required to take values in the closed, bounded interval $[0, \sigma_0]$, where $\sigma_0 > 0$; in particular, the control can be degenerate. In general, there will be a bang-bang optimal control that takes the value σ_0 in some open set and is zero otherwise. We discuss the existence and properties of solutions to stochastic differential equations with such controls and characterize the value function and optimal control in more detail, in the case of both linear and nonlinear drift. Employing the Hamilton–Jacobi–Bellman equation and results of [N. V. Krylov, *Theory Probab. Appl.*, 17 (1973), pp. 114–131] and [P.-L. Lions, *Comm. Pure Appl. Math.*, 34 (1981), pp. 121–147], we derive sufficient conditions for the existence of single-region optimal controls, construct examples of multiple-region controls, and provide bounds on the number and size of the regions in which the optimal control is positive.

Key words. stochastic optimal control, degenerate diffusion, smooth fit

AMS subject classifications. 93E20, 60H10, 60H30, 60G35

PII. S0363012998347122

1. Introduction.

The simplest model for a scalar diffusion with variance control is the stochastic differential equation

$$X_x^u(t) = x + \int_0^t b(X_x^u(s))\,ds + \int_0^t u(s)\,dW(s). \tag{1.1}$$

This paper analyzes the problem of control to minimize the discounted cost

$$J(x,u) = E\left[\int_0^\infty e^{-\alpha t}\left[c(X_x^u(t)) + \lambda u^2(t)\right]\,dt\right], \tag{1.2}$$

where the state dynamics, the location cost c, and the controls have the following special structure. First, the differential equation $\dot{x} = b(x)$, corresponding to zero variance control, has a unique, global, asymptotically stable equilibrium point, set arbitrarily at $x = 0$. Second, c is a bounded function that achieves its unique maximum at the equilibrium point $x = 0$ and decreases as x moves away from $x = 0$. The precise definition of an admissible control is given at the beginning of section 3. The main point is that the control is bounded and allowed to degenerate to the value zero; that is,

$$0 \le u(t) \le \sigma_0 \qquad \text{for all } t, \tag{1.3}$$

where σ_0 is a given constant. Throughout, λ is a given positive constant. To lend the problem a simplifying symmetry, we shall also assume throughout that c is an

*Received by the editors November 13, 1998; accepted for publication (in revised form) March 6, 2000; published electronically August 17, 2000.
http://www.siam.org/journals/sicon/39-1/34712.html

†Department of Mathematics, Rutgers University, Piscataway, NJ 08854-8019 (ocone@math.rutgers.edu).

‡Department of Mathematics, Iowa State University, Ames, IA 50011 (weera@math.iastate.edu).

even function and that b is odd. Our work was motivated by the desire to understand qualitative properties of optimal controls when degeneration is allowed, and the special structure assumed here allows some intuition that gives insight into this question. A recent, specific example with similar cost and dynamic structure, but with a long run average criterion and potentially unbounded ($\sigma_0 = \infty$) controls, appears in a dynamic sampling application in Assaf [1].

Early specific examples of degenerate variance control problems appear in Genis and Krylov [7] and Krylov [12]. In particular, they find that the value function is C^1 and piecewise C^2. General (previscosity solution) theory for degenerate problems is established in Krylov [12] and Lions [13] and shows under what hypothesis the value function for degenerate control problems is C^1.

When the cost c and drift b have the structure specified above, there are two opposing tendencies. If zero variance control (u=0) is applied, no control cost is incurred, but the state moves monotonically to regions of higher and higher location cost. On the other hand, if positive control is exercised, it will spread the state out in an average sense and so tend to stall its progression toward the equilibrium. This effect will be of greatest advantage in a region about the equilibrium point, where c is concave. Conversely, when the state is far away from 0 in a region where c is reasonably flat or convex, one expects that the advantage of a positive diffusion control either does not exist or is outweighed by the control cost. Intuition thus suggests that *single-region bang-bang* feedback controls, which set $u = \sigma_0$ when $|x| < a$ and $u = 0$ when $x \geq a$ for some a, should provide good candidates for optimal controls. This control structure is evident in the example in Assaf [1]. The same intuition also applies in the example of Genis and Krylov [7].

This paper addresses the following questions. First, bang-bang, degenerate controls give rise to stochastic differential equations with discontinuous diffusion coefficients. Do solutions exist for such equations, and how do they behave? In section 2, it is shown that a weak solution will always exist in some probability space if the diffusion coefficient has the form $\mathbf{1}_G(x)$, where G is the indicator function of set G, as long as G is open, and the behavior of the solutions is described. Second, what conditions suffice to imply the existence of single-region optimal controls and how can the optimal region be characterized? More generally, what factors determine where positive diffusion control should be exercised? In this paper, we analyze in some detail how the C^1 condition and the Hamilton–Jacobi–Bellman (HJB) equation determine the optimal value function and how the structure of b and c determine where the regions of degeneracy for the optimal control occur. Section 3 recalls the theory of the HJB equation for the value function and develops the extensions and refinements which are needed to study problem (1.1)–(1.3). Section 4 analyzes this problem when the drift is linear, which is treated separately since our results in this case are more complete. It is shown that the decisive factor for turning on the diffusion is the degree of concavity of c. To illustrate, sufficient conditions are given for the optimal control to be single-region and a class of problems with multiple-region optimal controls are constructed. In section 5, sufficient conditions are derived for single-region control in the more delicate case of nonlinear drift b.

Several assumptions in (1.1)–(1.3) can be generalized without changing the results. For example, the constraint (1.3) is equivalent to $|u(t)| \leq \sigma_0$ because of the symmetry of Brownian motion. Also, the particular form $u^2(t)$ of the control cost could be replaced by $\ell(u)$, where $\ell(u)$ is any strictly increasing function satisfying $\ell(0) = 0$. Because the optimal controls turn out to be bang-bang, taking values either 0 or σ_0, the results in either case will be qualitatively the same and will depend quantitatively only on $\ell(\sigma_0)$.

Several authors have studied variance control problems with nondegenerate controls. McNamara [15] considers maximization of terminal reward for a drift-less diffusion with feedback controls u satisfying $\sigma_1 \le u(t) \le \sigma_2$ for strictly positive constants σ_1 and σ_2. In [16] he considers controls which switch between two drift-diffusion pairs, both nonsingular in the diffusion term. A simple variance control problem is treated as an example in Rogers and Williams [18]. Dorroh, Ferreyra, and Sundar [3] treat problems with control in both drift and noise when the control is allowed to be unbounded.

2. Stochastic differential equations with bang-bang diffusion. A formal analysis of the HJB equation (see section 3) for the problem (1.1)–(1.3) suggests that optimal controls will have a feedback form $\sigma_0 \mathbf{1}_G(x)$, where $\mathbf{1}_G(x)$ is the indicator function of a subset G in$R$. To make sense of such controls, it is necessary to show that (1.1) admits a solution when u is replaced by $\mathbf{1}_G(X(t))$. The assumption made in section 1 that $\dot{x} = b(x)$ has a unique stable equilibrium point is irrelevant to this problem; thus, in this section we consider the one-dimensional equation

$$dX(t) = f(X(t))\,dt + \sigma_0 \mathbf{1}_G(X(t))\,dW(t), \qquad X(0) = x_0, \tag{2.1}$$

where it is assumed only that

(2.2) f is locally Lipschitz; and

(2.3) $\sup_R x f(x) < K(1+|x|^2)$ for some constant $K < \infty$ and all x.

By a *solution* to (2.1), we always mean a weak solution with continuous paths and almost surely (a.s.) infinite lifetimes defined on some probability space $(\Omega, \mathcal{F}, P)$ with a filtration $\{\mathcal{F}_t\}$ and an $\{\mathcal{F}_t\}$-Wiener process W. The condition (2.3) implies that lifetimes of solutions to (2.1) are infinite, once existence is established locally; see Friedman [6, p. 125].

Solutions to (2.1) do not exist for arbitrary G. For example, the equation $dX(t) = \mathbf{1}_{[0,\infty)}(X)\,dW$, $X(0) = 0$, does not admit a solution even in a weak sense. If a solution did exist, it would clearly have to remain in $[0,\infty)$ for all time. This would imply that $dX = dW$, and hence that $X = W$. However, Brownian motion exits $[0,\infty)$ with probability one in any time interval $[0,\epsilon)$ for any $\epsilon > 0$, and so a contradiction is obtained. On the other hand, $dX = \mathbf{1}_{(0,\infty)}(X)\,dW$, $X(0) = x_0$, is easy to solve for any x_0; if $x_0 \le 0$, then $X(t) \equiv x_0$ is the solution; if $x_0 > 0$, a solution is $X(t) = x_0 + W(t \wedge \tau)$, where τ is the first time $x_0 + W(t)$ hits 0.

That the removal of the boundary point 0 from $[0,\infty)$ changes an unsolvable equation into a solvable one reflects the main result, Theorem 2.3, of this section: if G is open, then (2.1) admits a weak solution. This is a useful piece of information for the control problem. The HJB equation by itself does not indicate whether the boundary points should be included in the optimal G. The analysis of this section shows that they should be excluded in the construction of optimal feedback controls.

Solutions to (2.1) with open G are not unique in general. For example, both $X \equiv W$ and $X \equiv 0$ solve $dx = \mathbf{1}_G(X)dW$, $X(0) = 0$, where $G = R - \{0\}$, because the Lebesgue measure of the total time a Brownian motion spends at the origin is zero. For the control problem, solutions that are guided strictly by the drift when the diffusion is zero are preferred. We shall say that a continuous process X *passes through* *a point y in one direction* if the event that there are times $0 \leq t_1 < t_2 < t_3$ such that either $X(t_1) < y$, $X(t_2) > y$, $X(t_3) < y$, or $X(t_1) > y$, $X(t_2) < y$, $X(t_3) > y$, has probability zero. Theorem 2.3 shows that weak solutions to (2.1) for open G may be constructed to pass through points of G^c in one direction only.

The proof for general open G ultimately reduces to the case in which G is simply an open interval. This case is studied carefully in Lemma 2.1 and Theorem 2.2, which are stated separately from Theorem 2.3 because of their particular importance to the control problem. After deriving Theorem 2.3 in the first version of the paper, we learned of the recent book of Assing and Schmidt [2], which provides very general theorems about strong Markov processes on the real line from which the existence of solutions to (2.1) can be deduced. The direct proof here relies only on standard methods.

It will be convenient to let ϕ denote the flow associated to the drift f when no diffusion term is present; $\phi(t,x)$, $t \geq 0$, $x \in R$, solves $\frac{\partial\phi(t,x)}{\partial t} = f(\phi(t,x))$, $\phi(0,x) = x$.

LEMMA 2.1. *Let $G = (r,q)$. Then* (2.1) *admits a unique weak solution. If $f(r) > 0$ and $f(q) < 0$ and if the solution enters (r,q), it will remain in (r,q) for all future time. If $f(r) \leq 0$ (respectively, if $f(q) \geq 0$), the solution will exit (r,q) never to return, once it hits r (respectively, q).*

Proof. Without loss of generality, set $\sigma_0 = 1$. There are three cases to treat, depending on the signs of the drift at the endpoints of (r,q). In the first case, $f(r) \leq 0$ and $f(q) \geq 0$; we say that the drift points out of (r,q) at each endpoint. Then, if $x_0 \notin (r,q)$, the solution $\phi(t,x_0)$ to $\dot{z} = f(z)$, $z(0) = x_0$, never enters (r,q) and thus solves (2.1) for $G = (r,q)$. If $x_0 \in (r,q)$, let $\tilde{X}_{x_0}$ denote the solution starting at x_0 to $d\tilde{X} = f(\tilde{X})\,dt + dW$. Then a solution to (2.1) is constructed by first following $\tilde{X}(t)$ until the first time τ that it exits (r,q) and then following $\phi(t-\tau, \tilde{X}(\tau))$. Strong uniqueness of the solution is obvious, at least until the first time, if any, that X_{x_0} reaches either r or q. So it remains to establish uniqueness starting from the boundary points. For example, let $X_r(t)$ be any solution starting from r. We want to show $X_r(t) = \phi(t,r)$. For this, apply the generalized Itô rule of problem 7.3 on p. 209 in Karatzas and Shreve [11] to $(X_r(t)-r)^+$ and take expectations. Then, if $\tau_\epsilon = \inf\{t \mid X_r(t) \notin (r-\epsilon, r+\epsilon)\}$,

$$E\left[(X_r(t)-r)^+\right] = E\left[\int_0^{t\wedge\tau_\epsilon} f(X_r(s))\mathbf{1}_{(r,q)}(X_r(s))\,ds\right]. \tag{2.4}$$

By the local Lipschitz property of f, there is an $\epsilon > 0$ and $K < \infty$ so that for $s \leq \tau_\epsilon$, $(f(X_r(s)) - f(r))\,\mathbf{1}_{(r,q)}(X_r(s)) \leq K\,(X_r(s)-r)^+$. Thus, using the fact that $f(r) \leq 0$ and some rearrangement of terms, it follows that

$$E\left[(X_r(t)-r)^+\right] \leq KE\left[\int_0^{t\wedge\tau_\epsilon} (X_r(s)-r)^+\,ds\right]. \tag{2.5}$$

The Gronwall–Bellman inequality then implies that $0 \leq E\left[(X_r(t)-r)^+\right] \leq 0$ for all t. Hence the solution X does not enter (r,q) a.s., which means that $dX_r(t)/dt = f(X_r(t))$, $t \geq 0$, as desired.

To handle the remaining cases, it will be enough to consider $G = (r,\infty)$ when $f(r) > 0$; here, f points into G at r. The result is an immediate application of Theorem 7.1 in Chapter 4 of Ikeda and Watanabe [8], taking $d = 1$. We briefly sketch the construction for the insight it affords. Let B be a Brownian motion with filtration

$\{\widetilde{\mathcal{F}}_t\}$ and let $x_0 \geq r$. Then the diffusion process Z that reflects at r and has drift $f(\cdot)$ and diffusion coefficient σ_0^2 satisfies

$$Z(t) = x_0 + \int_0^t f(Z(s))\,ds + \int_0^t \sigma_0 \mathbf{1}_{(r,\infty)}(Z(s))\,dB(s) + L(t),$$

where L is the local time of the process Z at r. Now define $T(t) = t + (L(t)/f(r))$, and let B^* be a Brownian motion independent of B. Set $W(t) = B(T^{-1}(t)) + \int_0^t \mathbf{1}_{\{r\}}\left(Z(T^{-1}(s))\right) dB^*(s)$, $X(t) = Z(T^{-1}(t))$, and $\mathcal{F}_t := \widetilde{\mathcal{F}}_{T^{-1}(t)} \wedge \sigma\{B^*(s); s \leq t\}$. Then W is an $\{\mathcal{F}_t\}$-Brownian motion and (X, W) is a weak solution of (2.1). Indeed X is a Markov process with a sticky boundary at r; the set of times that X spends at r has positive Lebesgue measure but contains no open interval. If $x_0 \notin [r, \infty)$, then the solution is constructed by following the flow $\phi(t, x_0)$ until the time S when the flow hits r, if finite, and then following $X_r(t - S)$.

Now suppose that $G = (r, q)$, where $f(r) > 0$ and $f(q) \geq 0$. To construct the solution, simply follow the solution with the same drift f, but $G = (r, \infty)$ until it hits $[q, \infty)$ and from that time onward, follow the solution ϕ to the deterministic equation.

Finally, suppose that $G = (r, q)$ and that the drift points in at both boundaries, i.e., $f(r) > 0$ and $f(q) < 0$. If $x_0 < q$, construct X_{x_0} by first following the solution to (2.1) with G replaced by (r, ∞) until the first time it hits q, and then switch to an independent solution to (2.1) with G replaced by $(-\infty, q)$. Follow this solution until it hits r, then switch to an independent solution of (2.1) with $G = (r, \infty)$, and so forth. For $x_0 \geq q$, start instead with the solution to (2.1) for $G = (-\infty, q)$ and continue in the same way. Notice that once this solution enters $[r, q]$, it stays there forever. Weak uniqueness is a simple consequence of the analytic characterization of this process, which is given in the next result. □

THEOREM 2.2. *Let* $G= (r, q)$ *and assume* $f(r) > 0$ *and* $f(q) < 0$.

(a) *If* $r \leq x_0 \leq q$, *the solution* X_{x_0} *to* (2.1) *is a Markov diffusion in* $[r, q]$ *whose conservative, Feller transition semigroup is generated by the operator*

$$A = (\sigma_0^2/2)\frac{d^2}{dx^2} + f(x)\frac{d}{dx} \quad \text{on the domain}$$

$$\mathcal{D}(A) = \{\psi \in C^2[r, q]; \psi''(r) = 0, \psi''(q) = 0\}.$$

(b) *For any constants* $\alpha > 0$, $\lambda > 0$, *and function* $c \in C[r, q]$,

$$Az(x) - \alpha z(x) + \sigma_0^2\lambda + c(x) = 0, \qquad z \in C^2[r, q], \tag{2.6}$$

$$z''(r) = -2\lambda, \quad z''(q) = -2\lambda \tag{2.7}$$

has the unique solution

$$z(x) = E\left[\int_0^\infty e^{-\alpha t}\left(c(X_x(t)) + \lambda\sigma_0^2\mathbf{1}_{(r,q)}(X_x(t))\right) dt\right], \quad r \leq x \leq q. \tag{2.8}$$

Proof. To prove part (a), apply Theorem 4 on page 44 of Mandl [14] to conclude that the operator A on $\mathcal{D}(A)$ generates a unique, conservative, Feller transition semigroup on $C[r, q]$. Now let X_{x_0} solve (2.1) for $r \leq x_0 \leq q$. Then an application of Itô's lemma shows that X_{x_0} solves the martingale problem for $(A, \mathcal{D}(A))$. We leave the calculation to the reader; note only that, because $\psi''(r) = \psi''(q) = 0$ for $\psi \in \mathcal{D}(A)$,

$$\frac{\sigma_0^2}{2}\mathbf{1}_{(r,q)}(X_{x_0}(t))\psi''(X_x(t)) + f(X_{x_0}(t))\psi'(X_{x_0}(t)) = A\psi(X_{x_0}(t)) \qquad \text{for all } t.$$

It follows by a general theorem—see, for example, Theorem 4.1 in Ethier and Kurtz [4]—that X_{x_0} is a Markov process generated by $(A, \mathcal{D}(A))$, as claimed. This conclusion also proves the uniqueness in law of the solution.

Next, consider part (b). If $z \in C^2[r,q]$ satisfies (2.6)–(2.7), then the representation (2.8) is derived using Itô's rule in the usual way. To show the existence of a solution to (2.6)–(2.7), apply standard o.d.e. theory, as expressed, for example, in Mandl [14, Lemma 5, p. 43] for the case $\lambda = 0$. If $\lambda \neq 0$, let u solve $Au(x) - \alpha u(x) = -c(x) + 2\lambda(x-q)f(x) - \alpha\lambda(x-q)^2$, $u''(r) = 0$, $u''(q) = 0$; then $z(x) = u(x) - \lambda(x-q)^2$ solves (2.7). □

We turn now to the case of a general open set G. We intend to construct a weak solution to (2.1) in the path space $C[0,\infty)$.

THEOREM 2.3. *Let $G \subset R$ be an open set. Equation* (2.1) *admits a weak solution X_{x_0} with the property that it passes through points of G^c in one direction.*

Proof. Let $\{(r_i, q_i);\ i \in I\}$ denote the countable collection of connected components of G; thus $G = \cup_{i \in I}(r_i, q_i)$. When I is finite, it is easy to construct a solution to (2.1). Suppose the initial point x lies in (r_j, q_j). Take a solution $X^{(j)}$ to (2.1) when G is replaced by (r_j, q_j) and follow this solution until the first time, if any, that it enters a different interval $[r_i, q_i]$. Then switch to a solution $X^{(i)}$ of (2.1), initialized at the point of entry into $[r_i, q_i]$, when G is replaced by (r_i, q_i). Continue in this manner, switching to a solution $X^{(k)}$ corresponding to (r_k, q_k) each time a new interval $[r_k, q_k]$ is entered.

The same technique works easily when G has an infinite number of components as long as only a finite number of them intersect any compact set. For general open sets, a patching method can still be used, but care must be taken when X crosses an infinite number of intervals of G in finite time. Instead, we complete the proof by using martingale problem theory and taking weak limits.

For an open set G, let $A_G = \frac{1}{2}\sigma_0^2 \mathbf{1}_G(x)\frac{d^2}{dx^2} + f(x)\frac{d}{dx}$. Let ξ be the canonical process on the space $\Omega := C([0,\infty))$ of continuous, real-valued paths, let $\mathcal{F}_t$ denote the filtration generated by ξ, and set $\mathcal{F} := \mathcal{F}_\infty$. A solution to the martingale problem for A_G and initial value x_0 (in the sense of Stroock and Varadhan [19]) is a probability measure P_{x_0} on Ω such that $P(\xi(0) = x_0) = 1$ and $\psi(\xi(t)) - \int_0^t A_G\psi(\xi(s))\,ds$ is a P_{x_0}-martingale for any $\psi \in C_0^2$, the twice continuously differentiable functions with compact support. As is well known, the existence of a solution to the martingale problem for A_G is equivalent to the existence of a weak solution to (2.1).

Without loss of generality, we may assume that the initial value x_0 lies in G^c and satisfies $f(x_0) > 0$ and $(x_0, \infty) \cap G \neq \emptyset$. Indeed, if x_0 is in the component (r_j, q_j) of G, we may follow the solution $X^{(j)}$ until it exits (r_j, q_j) and then solve (2.1) from the exit point, which is not in G. If $f(x_0) > 0$ but $(x_0, \infty) \cap G = \emptyset$, then $X(t) = \phi(t, x_0)$ solves (2.1), while if $x_0 \in G^c$ and $f(x_0) = 0$, then $X(t) \equiv x_0$ solves (2.1), so we need not treat these cases. Finally, an argument analogous to the one below handles the case $f(x_0) < 0$. For convenience, we assume also that

$$x_1 := \inf\{y \in G^c; f(y) \leq 0\} < \infty. \tag{2.9}$$

This will be removed later. Notice that $x_1 \in G^c$ and $f(x_1) \leq 0$, because G^c is closed and f is continuous. Therefore, the solution we construct will not take values beyond x_1, and hence it can be assumed that $G \subset [x_0, x_1]$. Now define $G_n = \cup\left\{(r_i, q_i)\ ;\ q_i - r_i \geq \frac{1}{n}\right\}$. Since G is contained in a bounded set by our assumptions, G_n is a finite union of disjoint open intervals. Thus for each n, there is a solution to (2.1) with G replaced by G_n, and hence a solution Q_n to the martingale problem for A_{G_n} starting from x_0. Because $f(x_0) > 0$ and $f(x_1) \leq 0$, $Q_n(x_0 \leq \xi \leq x_1$, for all $t \geq 0) = 1$ for each n. Since $\sup_{[x_0,x_1]} |f(x)| < \infty$, it is an immediate consequence of Theorem 1.4.6 in [19], for example, that the sequence $\{Q_n\}$ is tight as a family of probability measures on Ω. Let Q denote a weak limit of some subsequence $Q_{n'}$. The aim is to show that Q solves the martingale problem for A_G and initial condition x_0.

We first show that for any $\psi \in W^{2,\infty}$ such that $\mathbf{1}_G(x)\psi''(x)$ has a continuous version, the process

$$\theta_\psi(t) := \psi(\xi(t)) - \int_0^t A_G\psi(\xi(s))\,ds \quad \text{is a martingale on } (\Omega, \mathcal{F}, Q). \tag{2.10}$$

For this, it suffices to show that

$$E^Q\left[H\left(\theta_\psi(t) - \theta_\psi(s)\right)\right] = 0 \tag{2.11}$$

for any $s < t$ and for any function H on Ω which is bounded, continuous (in the topology of uniform convergence on compact time intervals), and $\mathcal{F}_s$ measurable. Because ψ has been constrained so that $A_G\psi$ is continuous, the integrand $H(\theta_\psi(t) - \theta_\psi(s))$ is a continuous function of ξ, and thus the expectation in (2.11) is equal to the limit of $E^{Q_{n'}}[H(\theta_\psi(t) - \theta_\psi(s))]$ as $n' \to \infty$. However, since Itô's rule is valid for $\psi \in W^{2,\infty}(R)$ (see Karatzas and Shreve [11, p. 219]), its application shows that

$$E^{Q_{n'}}\left[H\left(\theta_\psi(t) - \theta_\psi(s)\right)\right] = E^{Q_{n'}}\left[H\int_s^t \frac{\sigma_0^2}{2}\mathbf{1}_{G-G_n}(\xi(v))\psi''(\xi(v))\,dv\right].$$

Under the measure $Q_{n'}$, ξ solves $\xi'(t) = f(\xi(t))$ when $\xi(t) \in G - G_{n'}$. Thus, $\int_{x_0}^{x_1-\delta} \mathbf{1}_{G-G_n}(z)\frac{1}{f(z)}\,dz$ represents the total time ξ spends in $G - G_n \cap [x_o, x_1 - \delta]$. Hence

$$E\left[\left|\int_s^t \mathbf{1}_{G-G_n}(\xi(v))\psi''(\xi(v))\,dv\right|\right] \leq \|\psi''\|_\infty \int_{x_0}^{x_1-\delta} \frac{\mathbf{1}_{G-G_n}(z)}{f(z)}\,dz. \tag{2.12}$$

However, the definition of x_1 and the continuity of f imply that the quantity $\triangle := \inf\{f(x);\ x \in [x_0, x_1 - \delta] \cap G^c\} > 0$. Since f is uniformly continuous on $[x_0, x_1]$, it follows that there is an N_δ such that

$$\inf\{f(x);\ x \in [x_0, x_1 - \delta] \cap (G - G_n)\} > \triangle/2 \qquad \text{for } n \geq N_\delta,$$

because any point in $[x_0, x_1 - \delta] \cap (G - G_n)$ is within a distance of $1/n$ from G^c. Therefore

$$\lim_{n\to\infty} \int_{x_0}^{x_1-\delta} \mathbf{1}_{G-G_n}(z)\frac{1}{f(z)}\,dz = 0,$$

which shows by (2.12) that $E^{Q_{n'}}\left[H\left(\theta_\psi(t) - \theta_\psi(s)\right)\right] \to 0$ as $n' \to \infty$, thus completing the proof of (2.11).

Now let $\psi \in C_0^2$. For any positive integer m, define

$$\rho_m(z) = \begin{cases} \psi''(z), & \text{if } z \in (G \cup [x_1 - m^{-1}, x_1)])^c, \\ \psi''(z)\left(m \cdot \operatorname{dist}\left(z, G^c \cup [x_1 - m^{-1}, x_1]\right) \wedge 1\right) & \text{otherwise.} \end{cases}$$

For a point a such that support$(\psi) \subset (a, \infty)$, define $\psi_m(z) = \int_a^x \int_a^y \rho_m(z)\,dz\,dy$. From the definition, it is evident that $\mathbf{1}_G(z)\rho_m(z) = \mathbf{1}_G(z)\psi_m''(z)$ is continuous and that $\psi_m''(z) = 0$ on $[x_1 - m^{-1}, x_1]$. Thus θ_{ψ_m} is a martingale on $(\Omega, \mathcal{F}, Q)$ for every m. Moreover, $\lim_{m\to\infty} \rho_m(z) = \psi''(z)$ for all $z \neq x_1$, and the functions $\rho_m(z)$, $m \geq 1$ are uniformly bounded. It follows that ψ is the bounded pointwise limit of the sequence $\{\psi_m\}$ and $A\psi$ is the bounded pointwise limit of $\{A\psi_m\}$. Taking limits as $m \to \infty$, θ_ψ is also a martingale on $(\Omega, \mathcal{F}, Q)$, as we needed to prove.

If (2.9) fails, if G is bounded, and if x_1 is an upper bound of G, the same construction works, but now with $\delta = 0$, to produce a solution up to the hitting time of x_1. If (2.9) fails and G contains a component (r, ∞), one can first solve the problem for $A_{G-(r,\infty)}$ and then patch to the solution of $dX = f(X)\,dt + \sigma_0\,dW$ when it hits r, because the solution will stay in (r, ∞) after this hitting time. If (2.9) fails, G is unbounded, and there is a sequence of points $\{a_i\}$, $x_0 = a_0 < a_1 < \cdots$ such that $a_i \in G^c$ for every i, the martingale problem for A_G can be solved by successively patching together the solutions for $A_{G\cap[a_{i-1},a_i]}$.

Fix a $y \in G^c$ and assume that $f(y) \geq 0$. To show that paths pass through y in the positive direction only with Q-probability one, it suffices to show that

$$E^Q[\phi(\xi(t) - y)\phi(y - \xi(s_1))\phi(y - \xi(s_2))] = 0$$

for any bounded, continuous function ϕ which is strictly positive on $(0, \infty)$ and 0 on $(-\infty, 0]$ and for any times $0 \leq s_1 < t < s_2$. But by construction, the set of paths that pass through y in one direction only has Q_n-probability one. Hence $E^{Q_n}[\phi(\xi(t) - y)\phi(y - \xi(s_1))\phi(y - s_2)] = 0$ for each n. Taking limits along a subsequence of Q_n converging weakly to Q gives the result. This has been done for a single $y \in G^c$. However, by taking a countable dense subset of G^c, it is true with probability one for all $y \in G^c$. □

3. The HJB equation for the value function. In this section, we discuss the HJB equation for the value function of the control problem (1.1)–(1.3). The results stated here are used in sections 4 and 5.

First, it is necessary to give a rigorous definition of an admissible control. An admissible control consists of a probability space $(\Omega, \mathcal{F}, P)$ endowed with a right continuous, complete filtration, $\{\mathcal{F}_t\}$, an $\{\mathcal{F}_t\}$-Wiener process W, and an $\{\mathcal{F}_t\}$-progressively measurable process u satisfying the constraint (1.3), where σ_0 is a fixed constant. The class of admissible controls is denoted by $\mathcal{U}$. We shall abuse terminology slightly by speaking of an admissible control u, without explicit mention of the underlying space or Brownian motion, and with the understanding that different spaces and Brownian motions may be attached to different u. A function $\alpha : [0, \infty) \times R \to [0, \sigma_0]$ is called an *admissible feedback control from* x if the equation

$$dX_x(t) = b(X(t))\,dt + \alpha(t, X_x(t))\,dW, \qquad X_x(0) = x,$$

admits a weak solution. In this case, $u(t) = \alpha(t, X_x(t))$ is an admissible control.

We assume without further mention that b is at least locally Lipschitz and satisfies (2.3). Additional hypotheses will be placed on b in the theorem statements. However, these minimal assumptions imply that, given any admissible control, (1.1) admits a unique, continuous, $\{\mathcal{F}_t\}_{t\geq 0}$-adapted solution $X_x^u(t)$, $t \geq 0$, for all $x \in R$; see Theorem 5.1 and problem 1 in Chapter 5 of Friedman [6]. This fact makes our definition of admissible control the same as that in Lions [13].

The value function for the control problem is $V(x) := \inf_{u\in\mathcal{U}} J(x, u)$, where $J(x, u)$ is defined in (1.2). The formal HJB equation for V is

$$\inf_{u\in[0,\sigma_0]} \frac{u^2}{2}\left(V''(x) + 2\lambda\right) + b(x)V'(x) - \alpha V(x) + c(x) = 0. \tag{3.1}$$

The rigorous interpretation of this equation requires specifying the appropriate class of functions to which the solution belongs. This issue was resolved by Krylov [12] for control problems with possibly degenerate diffusion coefficients; he showed in some generality that the value function is a solution in the Sobolev space $W^{2,p}$ for all $p \geq 1$ to its HJB equation. Lions [13] improved the regularity to $W^{2,\infty}$. In this theory, the value function satisfies the HJB equation for (Lebesgue) almost everywhere (a.e.) x. We apply Lions's theorem to (3.1) in Theorem 3.2.

The first result, Theorem 3.1, restates (3.1) in a convenient form, presents a verification theorem for suitably regular solutions, and derives an optimal control from this solution. Theorem 3.2 states the converse, namely that the value function is indeed a solution of (3.1). These two results are stated separately here because we wish to distinguish them when used in the subsequent analysis; many of the results in sections 4 and 5 are proved by direct construction and application of the verification theorem, which is elementary, while Theorem 3.2 requires a sophisticated theoretical result; see [13]. Theorem 3.3 elaborates on how C^1 and C^2 smooth fit conditions help determine V. The final lemmas establish some general qualitative facts about the optimal control.

THEOREM 3.1. *Assume that b is C^2 and that c is a bounded C^2 function. Suppose $\bar{V} \in L^\infty \cap C^1$ is piecewise C^2 and is a solution of*

$$\frac{\sigma_0^2}{2}\mathbf{1}_G(x)\left(\bar{V}''(x)+2\lambda\right)+b(x)\bar{V}'(x)-\alpha\bar{V}(x)+c(x)=0 \qquad \textit{for all } x, \tag{3.2}$$

and

$$\bar{V}''(x) \geq -2\lambda \qquad \textit{a.e. on } G^c, \tag{3.3}$$

where G is the open set $\{x\,;\, b(x)\bar{V}'(x)-\alpha\bar{V}(x)+c(x)>0\}$. Then $\bar{V}=V$ and the solution X_x^* of

$$X_x^*(t)=x+\int_0^t b(X_x^*(s))\,ds+\sigma_0\int_0^t \mathbf{1}_G(X_x^*(s))\,dW(s), \tag{3.4}$$

using the feedback control $u(t)=\sigma_0\mathbf{1}_G(X_x^*(s))$, is the optimal process.

Remark 3.1. On each bounded, connected component (r_i,q_i) of G, $\bar{V}$ is a C^4 solution of

$$\frac{\sigma_0^2}{2}\left(\bar{V}''(x)+2\lambda\right)+b(x)\bar{V}'(x)-\alpha\bar{V}(x)+c(x)=0, \quad r_i<x<q_i, \tag{3.5}$$

$$\bar{V}''(r_i+)=-2\lambda, \qquad \bar{V}''(q_i-)=-2\lambda. \tag{3.6}$$

This is an easy consequence of the assumption that $\bar{V} \in C^1$ and that $b \in C^2$ and $c \in C^2$.

Remark 3.2. The general theory (see Theorem 3.2) states that V solves the HJB equation in the sense that (3.1) holds for a.e. x. However, $V \in W^{2,\infty}_{\text{loc}}$ implies that $V \in C^1$ and V' is absolutely continuous, and once this is known it is not hard to show that V must satisfy (3.2) everywhere. The proof is omitted.

Proof of Theorem 3.1. The proof follows the usual method for verification lemmas. Because V is piecewise C^2, one can still apply Itô's lemma to $V(X_x^u(t))$; see [11, p. 209]. Let u be an admissible control such that $u(s) \geq \delta > 0$ for all s and some positive δ, and recall that $J(x,u)$ is its corresponding cost function. Then, if A is any set of zero Lebesgue measure, $E\left[\int_0^\infty \mathbf{1}_A(X_x^u(t))\,dt\right]=0$. (This is a consequence of the existence of a local time process $\{\Lambda_t(a)\,;\, t \geq 0, a \in R\}$ for X_x^u and the identity $\int_0^t g(X_x^u(s))u^2(s)\,ds = 2\int_{-\infty}^{\infty} g(a)\Lambda_t(a)\,da$, a.s. for bounded, measurable g; see [11, p. 218].) Then, by (3.3),

$$\inf_{u\in[0,\sigma_0]} \frac{u^2}{2}\left(\bar{V}''(X_x^u(t)) + 2\lambda\right) = \frac{\sigma_0^2}{2}\mathbf{1}_G(X_x^u(t))\left(\bar{V}''(X_x^u(t)) + 2\lambda\right) \quad \text{for a.e. } t.$$

Now apply Itô's lemma to $\bar{V}(X_x^u(t))e^{-\alpha t}$, use (3.2), take expectations, and let $t \to \infty$ in the usual way to conclude that $\bar{V}(x) \leq J(x,u)$.

If u is a given control, let $u_n(t) = n^{-1}\mathbf{1}_{\{u(t)\leq 1/n\}} + u(t)\mathbf{1}_{\{u(t)>1/n\}}$. Then $E[\int_0^T (u_n(t) - u(t))^2\,dt] \to 0$ as $n \to \infty$. From this, one can deduce by standard methods that $E[\sup_{[0,T]}|X_x^u(t) - X_x^{u_n}(t)|^2] \to 0$ as $n \to \infty$ for every $T > 0$. As a consequence, $J(x,u) = \lim_{n\to\infty} J(x,u_n) \geq \bar{V}(x)$. Since u was an arbitrary admissible control, $V(x) \geq \bar{V}(x)$.

Using Itô's rule and (3.2), one finds easily that

$$\bar{V}(x) = E\left[\int_0^\infty e^{-\alpha t}\left[c(X_x^*(t)) + \lambda\sigma_0^2\mathbf{1}_G(X_x^*(t))\right]\,dt\right],$$

and thus X_x^* is an optimal process, and $\bar{V} = V$. □

The next result states that the value function must indeed be a sufficiently regular solution of the HJB equation (3.1). It is a minor extension of a result of Lions [13] for multidimensional, degenerate control. Lions's theorem allows the drift b and the cost c to have linear growth. For the purposes of sections 4 and 5, we wish instead to assume

(B.1) $b \in C^3$, $xb(x) < 0$ for all $x \neq 0$, and b is decreasing,
(C.1) $c \in C_b^2$.

(C_b^2 is the set of C^2 functions whose derivatives of order 0 to 2 are bounded functions.) In condition (B.1), b may admit more than linear growth, but since (B.1) implies (2.3), the solutions to (1.1) have infinite lifetimes and bounded moments. One could relax assumption (B.1), but it allows a simple proof of the regularity of V.

Theorem 3.2. *Assume* (B.1) *and* (C.1). *Then the value function V is in $W^{1,\infty} \cap W_{loc}^{2,\infty}$ and is the unique such function solving*

$$\inf_{u\in[0,\sigma_0]} \frac{u^2}{2}\left(V''(x) + 2\lambda\right) + b(x)V'(x) - \alpha V(x) + c(x) = 0 \quad \textit{for a.e. } x. \tag{3.7}$$

Proof of Theorem 3.2. The uniqueness claim is proved in Theorem 3.1 because it is shown that if V does solve (3.7), it must be the value function.

For a given drift b, let V_b denote the associated value function. To prove that V_b is the solution of (3.7), we start with the case that $b \in C_b^2$. Without additional assumptions on b, Lions [13] establishes a general, multidimensional result which implies Theorem 3.2 for any $\alpha > \alpha_0$, where α_0 is a constant such that $J(x,u) \in W^{2,\infty}$ for any admissible control. To prove the theorem for $b \in C_b^2$ we need to show that $\alpha_0 = 0$ if (B.1) holds. This requires showing $J(\cdot,u) \in W^{2,\infty}$ for any $\alpha > 0$. To this end, let u be a fixed admissible control. First, elementary estimates show that

$$||V_b(x)||_\infty \leq \|J(x,u)\|_\infty \leq \frac{1}{\alpha}(\|c\|_\infty + \lambda\sigma_0^2).$$

Next, following a computation similar to [13], we observe that the solution $X_x^u(t)$ to (1.1) has a version which is a.s. $C^{0,2}([0,\infty)\times R)$ as a function of t and x. Furthermore, the first and second partial derivatives satisfy

$$\partial_x X_x^u(t) = 1 + \int_0^t b'(X_x^u(s))\partial_x X_x^u(s)\,ds, \tag{3.8}$$

$$\partial_x^2 X_x^u(t) = \int_0^t b''(X_x^u(s))(\partial_x X_x^u(s))^2 + b'(X_x^u(s))\partial_x^2 X_x^u(s)\,ds. \tag{3.9}$$

Since (B.1) requires b to be decreasing, $b' \le 0$, and thus (3.8) implies $\partial_x X_x^u(t) \le 1$ for all t a.s. Using (B.1), the assumption that b is decreasing, and the boundedness of b'', there is a constant K such that $E[\partial_x^2 X_x^u(t)] \le Kt$. In view of these estimates, it is easy to prove that $J(x,u)$ is twice continuously differentiable for any $\alpha > 0$. For example, $\partial_x J(x,u) = E\left[\int_0^\infty e^{-\alpha t} c'(X_x^u(t))\partial_x X_x^u(t)\,dt\right]$, from which follows

$$\|\partial_x J(\cdot,u)\|_\infty \le \frac{\|c'\|_\infty}{\alpha}. \tag{3.10}$$

Similarly,

$$\partial_x^2 J(x,u) = E\left[\int_0^\infty e^{-\alpha t}\left(c'(X_x^u(t))\partial_x^2 X_x^u(t) + c''(X_x^u(t))(\partial_x X_x^u(t))^2\right)\,dt\right],$$

from which follows the estimate

$$|\partial_x^2 J(x,u)| \le \alpha^{-1}\|c''\|_\infty + K\alpha^{-2}\|c'\|_\infty.$$

Thus $\alpha_0 = 0$ and Theorem 3.2 is valid when $b \in C_b^2$. In addition, from (3.10), as in Lions [13],

$$\|V_b'\|_\infty \le \frac{\|c'\|_\infty}{\alpha}. \tag{3.11}$$

Now let b satisfy only (B.1). For each positive integer n, choose a $b_n \in C_b^2$ such that b_n also satisfies (B.1) and agrees with b on $[-n,n]$. Let $J_n(x,u)$ and $V_n(x)$ be the cost and optimal value functions when the drift is b_n, and let $J_b(x,u)$ and $V_b(x)$ be the corresponding functions for drift b. Let $G_n = \{x\ ;\ b_n(x)V_n'(x) - \alpha V_n(x) + c(x) > 0\}$. Then we know from Theorems 3.1 and 3.2 for $b \in C^2$ (see Remark 3.2) that $\mathbf{1}_{G_n}(x)$ is the optimal feedback control, and that V_n solves the HJB equation with b replaced by b_n. We show that for every compact $K \subset R$, there is an N such that $V_b(x) = V_n(x)$ on K for all $n \ge N$. Theorem 3.2 for b then follows easily.

The following facts will be used:

(a) For any admissible control u and any x, $J(x,u) = \lim_{n\to\infty} J_n(x,u)$.

(b) There is a positive constant M (independent of n) such that any interval of length M contains a point z_n in G_n^c for every n.

We first complete the proof assuming these facts. Let $n > M$, where M is as in (b). Then for each $m \ge n$ and $x \in [-n+M, n-M]$, let $X_{x,m}^*$ be the optimal process when the drift coefficient is b_m. The corresponding optimal control is given by $u_m^*(t) = \mathbf{1}_{G_m}(X_{x,m}^*(t))$. From (b) the set G_m^c has nonempty intersections with $[-n,-n+M]$ and $[n-M,n]$, and therefore, by Theorem 2.3, the process $X_{x,m}^*(t)$ stays in the set $[-n,n]$ for all time. But on the set $[-n,n]$, $b_m \equiv b_n \equiv b$, $u_m^*(t) = \mathbf{1}_{G_m\cap[-n,n]}(X_{x,m}^*(t))$ and u_m^* is an admissible control for b_n as well as b. Therefore

$V_m(x) \geq V_n(x)$ and $V_m(x) \geq V_b(x)$ for x in $[-n+M, n-M]$. Similarly, u_n^* is admissible for b_m and $V_n(x) \geq V_m(x)$ on $[-n+M, n-M]$. Consequently, for all $m \geq n$, $V_m(x) = V_n(x) \geq V_b(x)$ for all x in $[-n+M, n-M]$. However, from (a), $J_b(x,u) = \lim_{n\to\infty} J_n(x,u) \geq \limsup_{n\to\infty} V_n(x)$. Thus, taking an infimum over admissible u gives $V_b(x) \geq \limsup_{n\to\infty} V_n(x)$. In conclusion, $V_b(x) = V_m(x)$ for all $m \geq n$ and x in $[-n+M, n-M]$.

It remains to prove (a) and (b). Fact (a) is a straightforward consequence of the following: for any x, admissible control u, and $T > 0$, $\lim_{n\to\infty} P(\tau_n \leq T) = 0$, where τ_n is the first time that X_x^u exits $[-n, n]$.

To prove fact (b), let $b \in C_b^2$ and let (r, q) be any connected component of G_b, where G_b is as in Theorem 3.1. Because $V_b''(x) < -2\lambda$ on G_b, $V_b(q) - V_b(r) \leq V_b'(r)(q-r) - \lambda(q-r)^2$. Since $|V_b'(r)(q-r)| \leq (\lambda/4)(q-r)^2 + |V_b'(r)|^2/\lambda$, it follows, using (3.11), that

$$\frac{3}{4}\lambda(q-r)^2 \leq V_b(r) - V_b(q) + |V_b'(r)|^2/\lambda \leq \frac{2}{\alpha}(\|c\|_\infty + \lambda\sigma_0^2) + \frac{\|c'\|_\infty^2}{\alpha^2\lambda}.$$

Thus the length of any component of G is bounded by a constant independent of b, and this establishes (b). □

Although the value function V is a C^1 function, as shown by Theorems 3.1 and 3.2, it is not in general C^2; the second derivative may have jumps at boundary points of G. However, the fact that V is C^1 and satisfies (3.2) determines how the second derivative jumps at boundary points of G. In particular, the next theorem will show that C^2 smooth fit must hold at an endpoint of a component of G if the drift points into G at that endpoint. As above, $G = \cup_{i\in I}(r_i, q_i)$ is the decomposition of G into its connected components.

THEOREM 3.3. *Assume $b \in C^2$ satisfies $xb(x) < 0$ for all $x \neq 0$, let $c \in C_b^2$, and suppose that $\bar{V}$ and G are as in Theorem* 3.1. *For each $i \in I$ such that $q_i > 0$, $\bar{V}'$ is differentiable at q_i and $\bar{V}'''(q_i-) = 0$. For each $i \in I$ such that $r_i < 0$, $\bar{V}'$ is differentiable at r_i and $\bar{V}'''(r_i+) = 0$.*

Remark 3.3. The assumption that $xb(x) < 0$ is made for convenience only. The extension to more general b follows from the same arguments.

Proof. Recall that $\bar{V}$ satisfies (3.5) on (r_i, q_i) and $\bar{V}''(r_i+) = \bar{V}''(q_i-) = -2\lambda$. Suppose that q_i is an isolated point of G^c, and, equivalently, that $q_i = r_j$ for some j. Then $\bar{V}''(q_i-) = -2\lambda = \bar{V}''(q_i+)$, and so $\bar{V}''$ is defined and continuous at q_i. Since $\bar{V}$ satisfies (3.5) on both sides of q_i, it is actually a solution of (3.5) in a neighborhood of q_i, and hence, because of the smoothness of b and c, $\bar{V}$ is C^4 in a neighborhood of q_i. Since $\bar{V}''$ has a local maximum at q_i, it follows that $\bar{V}'''(q_i) = 0$.

Suppose now $q_i > 0$, so that $b(q_i) < 0$. We need some preliminary observations. If $x \in G^c$ is not an isolated point of G^c, and if $\bar{V}'$ is differentiable at x, then

$$(b'(x) - \alpha)\,\bar{V}'(x) + c'(x) = -b(x)\bar{V}''(x), \tag{3.12}$$

because on G^c, $b\bar{V}' - \alpha\bar{V} + c = 0$.

Suppose q_i is not an isolated point of G^c. Observe first that $\bar{V}'''(q_i-) \geq 0$, because $\bar{V}'' < -2\lambda$ on (r_i, q_i) and $\bar{V}''(q_i-) = -2\lambda$. By differentiation of (3.5) and then evaluation at q_i-,

$$(b'(q_i) - \alpha)\,\bar{V}'(q_i) + c'(q_i) = b(q_i)2\lambda - \frac{\sigma_0^2}{2}\bar{V}'''(q_i-) \leq b(q_i)2\lambda. \tag{3.13}$$

By comparing (3.12) to (3.13), it is clear that if $\bar{V}'$ is differentiable at q_i, then $\bar{V}'''(q_i-) = 0$.

Let η be a version of the distributional second derivative $\bar{V}''$, which, because $\bar{V} \in W^{2,\infty}_{\text{loc}}$, is a locally bounded function. To prove that $\bar{V}'$ is differentiable at q_i it suffices to show that there is a set A of zero Lebesgue measure such that

$$\lim_{x_n \downarrow q_i, x_n \notin A} \eta(x_n) = -2\lambda. \tag{3.14}$$

Let A consist of those points x at which either $\bar{V}'$ is not differentiable, $\bar{V}''$ and η are not equal, x is an isolated point of G^c or $x \in G^c$, but $\bar{V}''(x) < -2\lambda$. Since $\bar{V}'$ is absolutely continuous, it is differentiable a.e., and its derivative coincides a.e. with η. Also we know $\bar{V}''(x) \geq -2\lambda$ a.e. on G^c, and isolated points of G are in the countable set $\{r_i, q_i\,;\, i \in I\}$. Thus A has measure zero. Now, because $b(q_i) < 0$,

$$-b(q_i)2\lambda \geq \lim_{x_n \downarrow q_i, x_n \in G^c \cap A^c} b(x_n)\eta(x_n) = -(b'(q_i) - \alpha)\bar{V}'(q_i) - c(q_i) \geq -b(q_i)2\lambda,$$

where we have used (3.12) and the continuity of $(b' - \alpha)\bar{V}' - c$, and the last inequality comes from (3.13). Thus $\lim_{x_n \downarrow q_i, x_n \in G^c \cap A^c} \eta(x_n) = -2\lambda$. On the other hand, if $x_n \in G$, $\bar{V}''(x_n) = \frac{2}{\sigma_0^2}\left[-b(x_n)\bar{V}'(x_n) + \alpha\bar{V}(x_n) - c(x_n) + \sigma_0^2\lambda\right]$, and, since the right-hand side is continuous and equal to $\bar{V}''(q_i-) = -2\lambda$ at q_i, $\lim_{x_n \downarrow q_i, x_n \in G \cap A^c} \bar{V}''(x_n) = -2\lambda$ also. Thus, we have shown (3.14). □

Remark 3.4. Suppose $0 < r_i < q_i$ for some component (r_i, q_i) of G. The value function V is determined on (r_i, q_i) by the differential equation (3.5) and the values of $V(r_i)$ and $V'(r_i)$. The values of $V(r_i)$ and $V'(r_i)$ are constrained first by the requirement that $r_i \in G^c$, which implies $b(r_i)V'(r_i) - \alpha V(r_i) + c(r_i) = 0$. The smooth fit condition of V' at q_i imposes a second rigid constraint; $V(r_i)$ and $V'(r_i)$ must be so related that V''' equals 0 at the first q after r at which $V''(q) = -2\lambda$. This fact is very useful in understanding how to piece together a solution of the HJB equation. In fact, the third derivative condition is sufficient as well as necessary for a C^2 smooth fit. We state this explicitly for later use. Suppose that $q \neq 0$. Let S be a function on a neighborhood of $(q - \delta, q + \delta)$ satisfying

$$\begin{aligned} &\tfrac{\sigma_0^2}{2}(S'' + 2\lambda) + bS' - \alpha S + c = 0 \quad &&\text{for } x \in (q-\delta, q) \text{ (resp., } x \in (q, q+\delta)),\\ &bS' - \alpha S + c = 0 &&\text{for } x \in (q, q+\delta) \text{ (resp., } x \in (q-\delta, q)). \end{aligned} \tag{3.15}$$

Assume that S is C^1 on $(q - \delta, q + \delta)$ and that S'' is continuous except possibly at q. Then

$$S'' \text{ is continuous at } q \text{ iff } S'''(q-) = 0 \text{ (resp., } S'''(q+) = 0). \tag{3.16}$$

Indeed, by differentiating (3.15) on both sides of q and using the continuity of S and S',

$$\begin{aligned} &b(q)[S''(q-) - S''(q+)] = -\tfrac{\sigma_0^2}{2}S'''(q-)\\ &\Big(\text{resp., } \; b(q)[S''(q+) - S''(q-)] = -\tfrac{\sigma_0^2}{2}S'''(q+)\Big). \end{aligned} \tag{3.17}$$

Hence (3.16) follows.

In the sections that follow, the cost function c will be even and decreasing on $(0, \infty)$ in addition to belonging to C_b^2, and b will be odd. Here are some general facts concerning V and G under these assumptions.

LEMMA 3.4. *Assume that c is a bounded, even function decreasing on* $(0, \infty)$. *Assume that b is an odd function satisfying* (B.1). *Then V is even and decreasing on* $(0, \infty)$.

Proof. Let u be an admissible control. Suppose $0 < y < z$. Let X_y and X_z be the solutions to (1.1) using the same control u, and define $\tau = \inf\{t > 0\,;\, X_y = -X_z\}$. The difference $Z = X_z - X_y$ satisfies $\dot{Z} = b(X_z) - b(X_y)$ and hence remains nonnegative for all time. Thus, using the evenness and unimodality of c, $c(X_z) \leq c(X_y)$ if $t \leq \tau$. Define a new process $\tilde{X}_z(t) = X_z(t)$ if $t \leq \tau$ and $\tilde{X}_z(t) = -X_y(t)$ if $t > \tau$. Then, using the oddness of b, $\tilde{X}_z$ solves (1.1) with Brownian motion $\tilde{W}(t) := W(t)\mathbf{1}_{\{t\leq\tau\}} + (W(\tau) - (W(t) - W(\tau)))\,\mathbf{1}_{\{t>\tau\}}$ and the same control u. To emphasize that the Brownian motion is a different one in $\tilde{X}_z$ than in X_y, let us denote the control for $\tilde{X}_z$ by $\tilde{u}$. Clearly $c(\tilde{X}_z(t)) \leq c(X_y(t))$ for all t. Hence $J(z, \tilde{u}) \leq J(y, u)$. As this construction is valid for any admissible u, it follows that $V(z) \leq V(y)$. □

LEMMA 3.5. *Let b be an odd function satisfying* (B.1).

(a) *If $c \in C_b^2$ is even and positive, and if G contains a connected component of the form $(-a, a)$, then $a < \sqrt{c(0)/\lambda\alpha}$.*

(b) *If $c \in C_b^2$ is even, positive, and decreasing on $(0, \infty)$, and if $G = \cup_i(r_i, q_i)$ is the decomposition of G into disjoint connected components, then $\sum_i (q_i - r_i)^2 < 4c(0)/\lambda\alpha$.*

Proof. V is even, positive, and C^1. Thus $V'(0) = 0$. Clearly $V(0) \leq c(0)/\alpha$, since if zero control is applied starting from $x = 0$, the solution of (1.1) remains at 0, and then $J(0, 0) = c(0)/\alpha$. Because $V'' < -2\lambda$ on $(-a, a)$,

$$0 < V(a) = V(0) + \int_0^a \int_0^y V''(z)\,dz\,dy < \frac{c(0)}{\alpha} - \lambda a^2.$$

The inequality $a^2 < c(0)/\lambda\alpha$ is immediate.

Lemma 3.4 implies that V decreases on $(0, \infty)$. If (r, q) is a connected component of $(0, \infty) \cap G$, it follows that $V' \leq 0$ and $V'' < -2\lambda$ on (r, q). Hence $V(q) - V(r) < -\lambda(q - r)^2$. Thus if $q_i > 0$,

$$0 \leq V(q_i) \leq V(0) + \sum_{j, 0\leq q_j \leq q_i} V(q_j) - V(r_j \vee 0) < \frac{c(0)}{\alpha} - \lambda \sum_{j, 0\leq q_j\leq q_i} (q_j - (r_j \vee 0))^2.$$

Combining this with a similar inequality for negative q_i and letting $|q_i| \to \sup_j |q_j|$ gives the result. □

4. Linear drift.

In this section we consider the stochastic control problem (1.1)–(1.3) when the drift is a linear function $b(x) = -\theta x$, with $\theta > 0$. Thus b satisfies (B.1) and there is a unique equilibrium point at $x = 0$ for (1.1) when $u \equiv 0$. The class of admissible controls is defined precisely at the beginning of section 3. For convenience, we set $\lambda = 1$; all results can be restated for a general λ by rescaling the cost c.

The work of this section was motivated by a conjecture that the optimal control will be a single-region feedback control of the form $u(t) = \mathbf{1}_{(-a,a)}(X(t))$, if the cost c satisfies (C.1) and the following additional hypotheses.

(C.2) c is even.

(C.3) c is continuous, positive, and decreasing on $(0, \infty)$.

We discovered instead that the location of regions of positive control action is determined by the behavior of the function

$$\hat{c}(x) := c''(x) + 2(2\theta + \alpha), \tag{4.1}$$

and hence by convexity properties of c. We know from section 3 that the optimal feedback control has the form $\mathbf{1}_G(x)$, where G is open. The main point of this section is that the number and location of the components of G are constrained by the set $\{x; \hat{c}(x) < 0\}$. Theorem 4.1 shows that positivity of $\hat{c}$ implies optimality of $u \equiv 0$. Theorem 4.2 establishes a sufficient condition for the optimality of single-region feedback controls in terms of $\hat{c}$. Lemma 4.5 shows in general how the number of components of G can be bounded using the number of component intervals of the set $\{\hat{c} < 0\}$. Theorem 4.7 provides a class of examples for which c satisfies (C.1), (C.2), and (C.3), but for which the optimal control is *not* of single-region form, contradicting the original conjecture. In Theorem 4.2, an interesting connection is made between the optimal control problem (1.1)–(1.3) and an optimal stopping problem.

THEOREM 4.1. *Assume that $c \in C^2$, and, for some positive K and m, $|c(x)| \leq K(1+|x|^m)$. If $\hat{c}(x) \geq 0$ for all x, then the control $u \equiv 0$ is optimal, and $Z^*(t) = xe^{-\theta t}$ is the optimal process.*

Remark 4.1. A particular case occurs when $c''(x) \geq 0$ for all x, so that c is convex. In fact, if c is convex, whether it is differentiable or not, $u \equiv 0$ is optimal. Indeed, for any admissible control $E[X_x^u(t)] = Z^*(t)$, and so, for convex c, $E[c(X_x^u(t))] \geq c(Z^*(t))$ by Jensen's inequality for any $t \geq 0$. Hence Z^* is optimal.

Proof of Theorem 4.1. When $u \equiv 0$, $X_x^u(t) = Z^*(t) = xe^{-\theta t}$, and the corresponding payoff is $Q(x) = \int_0^\infty e^{-\alpha t} c(xe^{-\theta t})\, dt$. The following properties are immediate.

(a) Q is a C^2 function, and $|Q(x)| \leq \tilde{K}(1+|x|^m)$ for some $\tilde{K} < \infty$.

(b) $\theta x Q'(x) + \alpha Q(x) = c(x)$.

Moreover, nonnegativity of $\hat{c}$ implies $Q''(x) \geq -2$ for all x. Thus Q is a C^2 solution to the HJB equation (3.1). Next, one can apply Itô's lemma together with standard verification-theorem arguments from stochastic control [5, pp. 145–146] to verify that Q is the value function. This proves Theorem 4.1. □

The next theorem discusses optimality of single-region feedback controls. For its statement, it is useful to define $\tilde{V}_a$ to be the solution to

$$\text{(4.2)} \qquad \begin{cases} \frac{\sigma_0^2}{2}(\tilde{V}_a''(x) + 2) - \theta x \tilde{V}_a'(x) - \alpha \tilde{V}_a(x) + c(x) = 0, \\ \tilde{V}_a''(-a) = \tilde{V}_a''(a) = -2. \end{cases}$$

THEOREM 4.2. *Assume* (C.1)–(C.2). *Suppose that there exists an open interval $(-\ell, \ell)$ such that*

$$\text{(4.3)} \qquad \{x; \hat{c}(x) < 0\} = (-\ell, \ell).$$

Let

$$\text{(4.4)} \qquad a^* = \sup\left\{a\,;\, \tilde{V}_a''(x) < -2 \textit{ for } -a < x < a\right\}.$$

Then $a^* > \ell$ and the solution to $Z_x^*(t) = x - \int_0^t \theta Z_x^*(s)\, ds + \int_0^t \sigma_0 \mathbf{1}_{(-a^*,a^*)}(Z_x^*(s))\, dW(s)$ is an optimal process for the control problem (1.1)–(1.2) *for every x.*

Examples. (a) Let $c(x) = (1+x^2)^{-1}$. Note that c also satisfies (C.3). If $(2\theta+\alpha) \geq 1$, then $\hat{c}(x) \geq 0$ for all x and $u \equiv 0$ is optimal. If $2\theta + \alpha < 1$, then there is an $\ell > 0$ so that (4.3) holds, and hence it will be optimal to turn on the diffusion in an interval about 0.

(b) More generally, (4.3) will hold for a function c satisfying (C.1)–(C.2) if c admits only one positive inflection point z and c'' is increasing on $(0, z)$.

We first prove Theorem 4.2 by direct construction of a solution V^* to (3.2) and (3.3) with $G = (-a^*, a^*)$, where G is defined as in Theorem 3.1. Later, we give an alternate derivation using developments based on Theorem 3.2.

To present the construction of V^*, it is convenient and interesting to convert the original problem of minimizing $J(x,u)$ to an optimal stopping problem. Given a Brownian motion W on any probability space, let Y_x denote the Ornstein–Uhlenbeck process solving

$$Y_x(t) = x - \int_0^t \theta Y_x(s)\,ds + \sigma_0 W(t). \tag{4.5}$$

Let

$$U(x) = \inf_\tau E\left[\int_0^\tau e^{-(2\theta+\alpha)t}\hat{c}(Y_x(t))\,dt\right], \tag{4.6}$$

where the infimum is taken over all $\{\mathcal{F}_t^W\}$ stopping times, $\{\mathcal{F}_t^W\}$ being the filtration generated by W, and $\hat{c}$ is given by (4.1). It will turn out that U is related to the optimal value function V by $U(x) = (V''(x)+2)\mathbf{1}_{\{V''(x)+2<0\}}$. The connection between value functions of optimal control and optimal stopping problems is a common theme of stochastic control. For example, see [9] and [10] and references therein.

Consider next a particular class of stopping times for the optimal stopping problem. For $a \geq |x|$, define

$$\tau_a^x := \inf\{t \geq 0\,;\, |Y_x(t)| \geq a\}. \tag{4.7}$$

If $|x| > a$, set $\tau_a^x = 0$. For any a, τ_a^x is finite a.s. since Y_x satisfies (4.5). Define the function U_a on $[-a,a]$ by

$$U_a(x) = E\left[\int_0^{\tau_a^x} e^{-(2\theta+\alpha)t}\hat{c}(Y_x(t))\,dt\right], \quad |x| \leq a. \tag{4.8}$$

Then U_a satisfies the differential equation

$$\left.\begin{array}{c} \frac{\sigma_0^2}{2}U_a'' - \theta x U_a' - (2\theta+\alpha)U_a + \hat{c} = 0 \;\text{ for } |x| < a, \\ U_a(-a) = U_a(a) = 0. \end{array}\right\} \tag{4.9}$$

For each positive a, U_a can be extended to R so that it satisfies the differential equation everywhere, and henceforth we represent this extension by U_a. Notice that each U_a is an even function, because $\hat{c}$ is even and the boundary conditions are symmetrical. Thus $U_a'(0) = 0$.

The next result contains the technical work necessary for constructing an optimal policy for the stopping problem.

LEMMA 4.3. *Assume* (C.1) and (C.2). *Then*

(a) *There exists a point $a^* > \ell$ such that U_{a^*}, the solution of* (4.9) *for $a = a^*$, also satisfies*

$$U_{a^*}'(a^*) = U_{a^*}'(-a^*) = 0. \tag{4.10}$$

(b) *The point $a^* > 0$, and hence the solution U_{a^*} satisfying* (4.10)*, are unique. Also, if $A := \{a > 0\,;\, U_a(x) < 0, \quad x \text{ in } (-a,a)\}$, then A is nonempty and $a^* = \sup A$.*

Proof. We indicate the idea of the proof and omit elementary details.

Let ℓ be given by (4.5); then $\ell \in A$ and $U'(\ell) > 0$. This follows from the boundary point lemma [17]. Consider the solution ϕ to the homogeneous equation

$$\frac{\sigma_0^2}{2}\phi'' - \theta x\phi' - (2\theta+\alpha)\phi = 0, \qquad \phi(0) = 1, \;\; \phi'(0) = 0. \tag{4.11}$$

Then ϕ is a nonnegative, even function. Also, $U(x)$ defined in (4.6) satisfies

$$U(x) \geq \frac{\inf_y \hat{c}(y)}{2\theta + \alpha}. \tag{4.12}$$

In particular, $U_a(x)$ also satisfies (4.12) for $|x| \leq a$, for each $a \in A$. For each $t > 0$, we consider $U_\ell(x) - t\phi(x)$, which solves (4.9). Introduce the set $T := \{t > 0\,;\, \exists a > \ell,\ U_\ell(a) - t\phi(a) = 0\}$. Then T is nonempty since $U_\ell'(\ell) > 0$, and (4.12) implies that T is bounded. Let $t^* := \sup T$, and introduce $a^* = \inf\{a > 0\,;\, U_\ell(a) - t^*\phi(a) = 0\}$. Then $a^* \in A$ and elementary arguments show that $A = (0, a^*]$, $U_{a^*}(a^*) = 0$, $U_{a^*}'(a^*) = 0$, and $U(x) < 0$ for $a^* > x \geq 0$. Hence, a^* satisfies the conditions in the lemma. Since $A = (0, a^*]$, uniqueness of a^* also follows. □

Lemma 4.3 allows us to solve the optimal stopping problem and calculate U.

Theorem 4.4. *Assume* (C.1) *and* (C.2). *Let a^* be defined as in Lemma* 4.3, *and τ_a^x as in* (4.7). *Then $\tau_{a^*}^x$ is the optimal stopping policy for the problem* (4.5)–(4.6) *and*

$$U(x) = \begin{cases} U_{a^*}(x) & \text{if } |x| \leq a^*, \\ 0 & \text{if } |x| > a^*. \end{cases}$$

Proof. Let $U^*(x) = U_{a^*}(x)\mathbf{1}_{[-a^*,a^*]}(x)$. Certainly, $U^*(x) \geq U(x)$, so we need only to verify the opposite inequality. By Lemma 4.3, U^* is a C^1 function and the second derivative of U^* is continuous everywhere except at $\pm a^*$. But we may apply Itô's rule to $U^*(Y_x(t))e^{-(2\theta+\alpha)t}$ (see [11, p. 219]) and use (4.9) to verify that $U^*(x) \leq U(x)$. □

Proof of Theorem 4.2. We will show that Theorem 4.2 holds when a^* is defined as in Lemma 4.3. For this a^*, let Z_x^* be the process defined in the statement of Theorem 4.2. The existence of Z_x^* and its uniqueness in law are treated in Theorem 2.2. Let V^* be the payoff from Z^*;

$$V^*(x) = E\left[\int_0^\infty e^{-\alpha t}\left(c(Z_x^*(t)) + \sigma_0^2 \mathbf{1}_{(-a^*,a^*)}(Z_x^*(t))\right) dt\right]. \tag{4.13}$$

We intend to employ Theorem 3.1. Observe that V^* satisfies (3.2) for all $x \neq \pm a^*$, with $G = (-a^*, a^*)$ and $\lambda = 1$. This follows from Theorem 2.2 and the fact that $Z_x^*(t)$ is the solution of the deterministic equation $\dot{x} = b(x)$, as long as it remains in $(-\infty, -a^*) \cup (a^*, \infty)$. In particular, Theorem 2.2 shows that

$$(V^*)''(-a^*+) = (V^*)''(a^*-) = -2. \tag{4.14}$$

Now let $\hat{U} := (V^*)'' + 2$. By differentiating the differential equation for V^* on $(-a^*, a^*)$ and using the boundary conditions (4.14), it follows that $\hat{U}$ satisfies (4.9) on $(-a^*, a^*)$. Thus, $(V^*)''(x) + 2 = U_{a^*}(x)$ on $(-a^*, a^*)$, and since $a^* \in A$, by Lemma 4.3, $(V^*)''(x) + 2 < 0$ on $(-a^*, a^*)$. By Lemma 4.3, it follows that, $(V^*)'''(-a^*+) = (V^*)'''(a^*-) = U_{a^*}'(\pm a^*) = 0$. Hence, condition (3.16) applies, implying that V^* is a C^2 function. Direct differentiation and the continuity of $(V^*)''$ imply

$$\theta x \hat{U}' + (2\theta + \alpha)\hat{U} = \hat{c} \quad \text{for } x > a^*, \quad \hat{U}(a^*) = (V^*)''(a^*) + 2 = 0. \tag{4.15}$$

By directly solving this equation and using the fact that $\hat{c} > 0$ on (a^*, ∞), $(V^*)'' + 2 = \hat{U} > 0$ on (a^*, ∞). By symmetry, $(V^*)'' + 2 > 0$ for $|x| > a^*$. Hence all the conditions of Theorem 3.1 are fulfilled, which completes the proof. □

The proof of Theorem 4.2 relied on direct construction and an application of the verification lemma of Theorem 3.1 We now invoke Theorem 3.2 to show that the characterization of a^* in Theorem 4.2 and the relation of the set $\{\hat{c} < 0\}$ to $(-a^*, a^*)$ reflects deeper and more general facts about the optimal value function. Let $G := \{x\,;\, -\theta x V'(x) - \alpha V(x) + c(x) > 0\}$ as in Theorem 3.1. General bounds on the size of the components of G are presented in Lemmas 3.4 and 3.5. The following result bounds the number of connected components of G. Note that it implies that G consists of at most one interval, symmetric about zero, under the assumptions of Theorem 4.2.

LEMMA 4.5. *Let c satisfy* (C.1) *and* (C.2), *and suppose the open set $\{x\,;\,\hat{c}(x) < 0\}$ has N connected components. Then G contains at most $N+1$ connected components. If G contains a connected component of the form $(-\delta, \delta)$, then G has at most N connected components.*

Proof. We intend to show that

(a) each connected component of G intersects with a connected component of the set $\{x\,;\,\hat{c}(x) < 0\}$, and that

(b) each connected component of $\{x\,;\,\hat{c}(x) < 0\}$ intersects with at most one connected component of G.

The proof of the lemma is an easy consequence of these facts.

Let (r, q) be a connected component of G with $q > 0$. Then V is a C^4 solution of (3.5) and (3.6) on (r, q) by Theorems 3.1 and 3.2. Let $U := V'' + 2$. Then $U < 0$ on (r, q) and there is an x_0 in (r, q) such that $U(x_0) = \min_{(r,q)} U$. By differentiating (3.5) twice and applying (3.6), U is a solution of

$$\left.\begin{array}{r} \frac{\sigma_0^2}{2}U'' - \theta x U' - (2\theta + \alpha)U + \hat{c} = 0 \quad \text{for } x \in (r, q), \\ U(r) = U(q) = 0. \end{array}\right\} \tag{4.16}$$

From (4.16), it follows that $\hat{c}(x_0) < 0$. This proves (a). Also, $\hat{c}(q)$ cannot be negative by (4.16) and the boundary point lemma. If (a_1, a_2) is a connected component of $(0, \infty) \cap \{\hat{c} < 0\}$ such that $(a_1, a_2) \cap (r, q) \neq \emptyset$, then $q \notin (a_1, a_2)$, and (b) follows.

If G contains zero, then the above argument shows that G contains at most N connected components. □

The next result shows that the characterization of a^* in Theorem 4.2 defines the limits of the connected component of G containing 0, in the general case.

THEOREM 4.6. *Let c satisfy* (C.1) *and* (C.2), *and define a^* by* (4.6) *and the set A as in Lemma* 4.3. *Assume that A is nonempty. Then $(-a^*, a^*)$ is one of the connected components of G.*

Proof. Let $a \in A$ and let $\tilde{V}_a$ be the payoff function associated with the feedback control $\mathbf{1}_{(-a,a)}(X(t))$. By Theorem 2.2, $\tilde{V}_a(x)$ solves (4.2). Because $\tilde{V}_a$ is even, $\tilde{V}_a'(0) = 0$, and by definition of A, $\tilde{V}_a'' + 2 < 0$ on $(-a, a)$. Hence, evaluation of (4.2) at 0 gives

$$\alpha V(0) \leq \alpha V_a(0) = c(0) + \sigma_O^2(V_a''(0) + 2) < c(0).$$

Since $V'(0) = 0$ and V is even, it follows that $0 \in G$. Thus there is a connected component $(-\bar{a}, \bar{a})$ of G. Consequently, by (4.2) and Theorem 3.1, we deduce $V(x) = V_{\bar{a}}(x)$ for $|x| \leq \bar{a}$. Therefore $\bar{a} \in A$ and $\bar{a} \leq a^*$. To show $a^* \leq \bar{a}$, let $a \in A$. If $x \in [-a, a]$ and if u is an admissible control that keeps the process X_x^u in $[-a, a]$ for all time, then a verification-lemma-type argument shows that $\tilde{V}_a(x) \leq J(x, u)$ for $|x| \leq a$. In particular, $\tilde{V}_{a_2}(x) \leq \tilde{V}_{a_1}(x)$ for all x in $[-a_1, a_1]$ whenever $0 < a_1 < a_2$ and a_1 and a_2 belong to A. If u^* denotes the optimal feedback control $\mathbf{1}_G(X(t))$ and

$(-\bar{a}, \bar{a})$ is the connected component of G containing 0, $V(x) = V_{\bar{a}}(x)$ for $|x| \leq \bar{a}$. Thus, if $\bar{a} \leq a$ and $a \in A$, then $\tilde{V}_a(x) = V(x)$ for $|x| \leq a$, and it follows that $a = \bar{a}$. In summary, $a \in A$ and $\bar{a} \leq a$ imply $\bar{a} = a$. Consequently, $\bar{a} \geq \sup A = a^*$, which completes the proof. □

Remark 4.2. Lemma 4.5 and Theorem 4.6 together show that if $\{\hat{c} < 0\}$ is an interval $(-\ell, \ell)$ and A is nonempty, then $G = (-a^*, a^*)$ with $a^* > 0$. Using the simple proof of the nonemptiness of A derived above in Lemma 4.3 yields an alternative proof of Theorem 4.2.

In the rest of this section, we consider cost functions c which satisfy (C.1), (C.2), and (C.3), and positive constants θ and α such that the set $\{x\,;\,\hat{c}(x) < 0\}$ is the union of three disjoint intervals. Under the additional assumption (C.4), (C.5), and (C.6) listed below, it will be shown that the optimal strategy is bang-bang and consists of multiple switchings of the control. This is rather technical but the intuition is simple. Since c satisfies (C.3), it is decreasing on $(0, \infty)$. Now suppose that at some very large value of x_0, c suddenly decreases to a fraction of the value $c(x_0)$ in a short interval beyond x_0. If the state process starts above x_0 and no control is exercised, it decreases, passing quickly through x_0 into a region of much higher cost. However, using positive diffusion control in a region near or above x_0 will delay arrival at x_0 and, despite the control cost, this might be worthwhile.

Because of Theorem 4.3, the value function is C^2 in a problem with a single-region optimal control. For multiple-region optimal controls, such as those constructed here, C^2 smoothness will fail at some of the switching points. See Remark 4.3 below.

Here are the precise hypotheses placed on c, θ, and α, in addition to (C.1)–(C.3).

(C.4) $\{x\,;\,\hat{c}(x) < 0\} = (-\delta_0, \delta_0) \cup (\delta_1, \delta_2) I_2 \cup (-\delta_2, -\delta_1)$, where $0 < \delta_0 < \delta_1 < \delta_2$.

(C.5) There exists a z_0, with $\sqrt{c(0)/\alpha} \leq \delta_0 < z_0 < \delta_1$, such that $\int_{\delta_0}^{x} \hat{c}(u)\,du > 0$ if $\delta_0 < x < z_0$.

(C.6) $\int_{\delta_0}^{\delta_2} u^{1+\alpha/\theta} \hat{c}(u)\,du < 0$.

Assumptions (C.4)–(C.6) guarantee the multiple switching of the optimal control.

Example. The idea is to start with a function c satisfying the assumptions of Theorem 4.2 and then to modify it far away from 0 by a sudden dip. The location and width of the dip are controlled by defining a "dip" function ρ and scaling it appropriately. Choose a function ρ such that (a) ρ is C^∞, decreasing, and odd; (b) $\rho(x) = 1$ if $x \leq -1$, and $\rho(x) = -1$ if $x \geq 1$; (c) $\rho''(x) \leq 0$ if $x < 0$, and $\rho''(x) \geq 0$ if $x \geq 0$; and (d) ρ'' admits a single local maximum and a single local minimum. Notice that $\rho''(x) = 0$ at $x = 0$ and for $|x| \geq 1$. Condition (d) on ρ is imposed so that if $\max_{[0,1]} \rho'' = -\min_{[-1,0]} \rho'' = k > 0$, then $\{\rho'' < 0\}$ is an interval (s_1, s_2) with $-1 < s_1 < s_2 < 0$. For $\epsilon > 0$ define scaled versions ρ_ϵ of ρ by $\rho_\epsilon(x) := \rho(x/\epsilon)$. Assume that $\gamma := 2\theta + \alpha < 1$. Fix an $R > \sqrt{c(0)/\alpha}$, and let c_ϵ be an even, nonnegative, C^3 function which satisfies

$$\begin{array}{ll} c_\epsilon(x) = 1 - x^2 & \text{if } |x| \leq 1/2, \\ c_\epsilon(x) = 1/2 & \text{if } 1 \leq |x| \leq R, \\ c_\epsilon(x) = 1/4 + (1/4)\rho_\epsilon\left(|x| - (R{+}\epsilon)\right) & \text{if } |x| > R. \end{array}$$

Assume also that c_ϵ is decreasing on $(0, \infty)$ and that $c \vee (1/2)$ satisfies the condition (4.3) of Theorem 4.2. Then $(-R, R) \cap \{\hat{c} < 0\}$ is a single interval of the form $(-\delta_0, \delta_0)$ contained in $(-1, 1)$. Clearly, (C.5) is satisfied for this construction.

We shall show that ϵ may be chosen to satisfy (C.4) and (C.6) as well. Observe that $c''(x) = \epsilon^{-2}\rho((x - R - \epsilon)/\epsilon)$ when $x > R$. Therefore, when $\max_{[0,1]} \rho'' > \epsilon^2 2\gamma$, it follows from the shape of ρ'' that $(R, \infty) \cap \{\hat{c} < 0\} = (\delta_1(\epsilon), \delta_2(\epsilon))$, where $R < \delta_1(\epsilon) <$

$\delta_2(\epsilon) < R + \epsilon$. In fact $(\delta_i - R - \epsilon)/\epsilon = z_i(\epsilon)$, where $z_1(\epsilon)$ and $z_2(\epsilon)$ are the solutions of $\rho''(x) = -2\epsilon^2\gamma$. This verifies (C.4). It is easy to see that $z_1(\epsilon) \downarrow -1$ and $z_2(\epsilon) \uparrow 0$ as $\epsilon \downarrow 0$. A calculation and using change of variables shows that

$$\int_{\delta_0}^{\delta_2} u^{1+\alpha/\theta}\hat{c}(u)\,du \le K + \epsilon^{-1}\int_{z_1(\epsilon)}^{z_2(\epsilon)} \rho''(u)(R+\epsilon(u+1))^{1+\alpha/\theta}\,du,$$

where K is a positive constant independent of ϵ. Therefore, since ρ'' is negative on $(-1,0)$, the last integral can be made as negative as desired by choosing ϵ small enough. Hence (C.6) will be satisfied for small enough ϵ.

THEOREM 4.7. *Assume* (C.1)–(C.6). *Then there exist* $0 < a_0^* < a_1^* < a_2^*$ *such that the optimal control is determined by the feedback function* $\sigma_0\mathbf{1}_G(x)$, *where*

$$G := (-a_2^*, -a_1^*) \cup (-a_0^*, a_0^*) \cup (a_1^*, a_2^*).$$

Moreover, $\delta_0 < a_0^* < z_0$ *and* $a_1^* < \delta_2 < a_2^*$, *where* δ_0, z_0, *and* δ_2 *are as defined in* (C.4)–(C.6).

Proof. Let $a_0^* = \sup A$, where A is defined as in Theorem 4.2. Then we know from Theorem 4.6 that $(-a_0^*, a_0^*)$ is a connected component of G. By Lemma 3.5 and the proof of Lemma 4.5, we know that $\delta_0 < a_0^* < z_0$. Suppose that G contains no more components. Then on $|x| > a_0^*$, V satisfies (3.7) and hence $U = V'' + 2$ satisfies (4.15). The solution of (4.15) is

$$U(x) = \frac{1}{\theta}\frac{1}{(a_0^*)^{2+\alpha/\theta}}\int_{a_0}^{x} u^{1+\alpha/\theta}\hat{c}(u)\,du.$$

However, this becomes negative by (C.6) in some interval about δ_2. Since $U(x) \ge 0$ a.e. on G^c, we obtain a contradiction to the assumption that $(-a_0^*, a_0^*)$ is the single component of G. Therefore G must contain at least two other components, symmetrically placed about 0. By Lemma 4.5 it can contain no more components. By the proof of Lemma 4.5, the additional positive component (a_1^*, a_2^*) must intersect (δ_1, δ_2), and it must be true that $\delta_2 < a_2^*$. □

Remark 4.3. In this example, the value function V is an even function which is C^2 everywhere except at $\pm a_1^*$. At $x = a_1$, $(\sigma_0^2/2)V'''(a_1^*+) + \theta a_1^*(V''(a_1^*-) + 2) = 0$. This condition can be easily verified by differentiating (3.2) near a_1^* and using the continuity of V'.

5. Nonlinear drift.

In this section we consider problem (1.1)–(1.3) when the drift b is nonlinear. In the treatment of linear drift, the function $\hat{c}$ defined in (4.1) played a crucial role. The generalization of this function, which we continue to denote by $\hat{c}$, is

$$\hat{c}(x) = c''(x) + 2\left(\alpha - 2b'(x)\right), \tag{5.1}$$

and this coincides with $\hat{c}$ defined in (4.1) when the drift is linear. The nonlinearity of b will make itself felt through a second auxiliary function g_c defined as

$$g_c(x) = \hat{c}(x) + \frac{b''(x)\left(c'(x) - 2b(x)\right)}{\alpha - b'(x)}. \tag{5.2}$$

The first result is a generalization of Theorem 4.1. It is proved in the same way, and so the proof is omitted.

THEOREM 5.1. *Assume that $b \in C^2$ is decreasing and $xb(x) < 0$ if $x \neq 0$. Assume that $c \in C^2$ and there exist constants K and m such that $0 \leq c(x) \leq K(1 + |x|^m)$ for all x. If*

$$g_c(x) \geq 0 \qquad \text{for all } x,$$

then $u \equiv 0$ is the optimal control for problem (1.1)–(1.3).

For the rest of the section it is a standard assumption that c satisfies (C.1)–(C.3), b satisfies (B.1), and

$$\text{(B.2)} \qquad b \text{ is odd.}$$

Again, by adding a constant to c, we may and do make the harmless assumptions that c is nonnegative and, in fact, $\lim_{|x|\to\infty} c(x) = 0$. Also, without loss of generality, $\lambda = 1$.

For each $a > 0$, let V_a denote the cost from the control strategy defined by $u(t) = \sigma_0 \mathbf{1}_{(-a,a)}(X_x(t))$. Then by Theorem 2.2, on $(-a,a)$, the function V_a satisfies

$$\text{(5.3)} \quad \frac{\sigma_0^2}{2}\left(V_a'' + 2\right) + b(x)V_a' - \alpha V_a + c(x) = 0, \quad V''(-a+) = -2, \ V''(a-) = -2,$$

and for $|x| > a$, $b(x)V_a'(x) - \alpha V_a(x) + c(x) = 0$. Hence V_a solves (3.2) with $G = (-a,a)$ for $x \neq \pm a$, but it does not in general satisfy (3.3). Our next concern is to find conditions under which a single-region control is optimal. In this section we take a direct approach. We verify directly that V_a is a solution to (3.2)–(3.3) for a suitable choice of a when b and c satisfy certain conditions; then application of Theorem 3.1 proves optimality of the single-region control.

Define the set A exactly as in section 4: $A := \{a > 0\,;\, V_a'' + 2 < 0 \text{ on } (-a,a)\}$.

LEMMA 5.2. *Assume that A is nonempty. Let $a \in A$ and assume that*

$$\text{(5.4)} \qquad V_a'''(a-) = 0 \qquad \text{and} \qquad V_a''(x) + 2 > 0 \qquad \text{for } |x| > a.$$

Then V_a is a C^2 solution to (3.2)–(3.3) *and $\sigma_0 \mathbf{1}_{(-a,a)}(x)$ defines an optimal feedback control.*

Proof. The argument is very similar to the proof of Theorem 4.2. It was shown in (5.3) that V_a satisfies (3.2) with $G = (-a,a)$ for all $x \neq \pm a$. Condition (5.4) asserts both that (3.3) is satisfied and that V_a is a C^2 function (see Remark 3.4 in section 3). Hence, Theorem 3.2 can be applied. □

To establish when the conditions of 5.2 are true, define again $W_a = V_a'' + 2$. This function is not necessarily defined at $\pm a$, but it will have limits from the left and right at these points. By differentiating (5.3) twice, we find

$$\text{(5.5)} \qquad LW_a = -g_c(x), \qquad W_a(-a+) = W_a(a-) = 0,$$

where L is the differential operator

$$L = \frac{\sigma_0^2}{2}\frac{d^2}{dx^2} + \left(b(x) + \frac{\sigma_0^2 b''(x)}{2(\alpha - b'(x))}\right)\frac{d}{dx} - \left[(\alpha - 2b'(x)) - \frac{b(x)b''(x)}{\alpha - b'(x)}\right].$$

On $|x| > a$, a similar derivation shows

$$\text{(5.6)} \qquad b(x)W_a' - \left(\alpha - 2b'(x) - \frac{b(x)b''(x)}{\alpha - b'(x)}\right)W_a = -g_c(x).$$

The following facts are needed. First, for any $a > 0$, any continuous function g, and any constant η,

$$\begin{cases} \frac{\sigma_0^2}{2}(Z'' + 2) + b(x)Z' - \alpha Z + g(x) = 0 \;\; \text{for } x \in (-a, a), \\ Z''(-a) = Z''(a) = \eta \end{cases} \tag{5.7}$$

has a unique solution. This is an immediate consequence of part (b) of Theorem 2.2. Second, $V_a''(x)$ is continuous in a for $|x| < a$. This can be deduced from elementary o.d.e. arguments. The next lemma generalizes the characterization of a^* in section 4.

LEMMA 5.3. (a) *Assume that* $\hat{c}(0) < 0$*. Then the set* A *is nonempty.*

(b) *Assume that* A *is nonempty. Then* A *is a closed, bounded set. If* $a^* = \sup A$*, then* $V_{a^*}'''(a^*-) = 0$.

Proof. Observe that $g_c(0) = \hat{c}(0) < 0$ by assumption. We pick ϵ small enough so that $g_c(x) < 0$ on $(-\epsilon, \epsilon)$ and $\alpha - 2b'(x) > b(x)b''(x)/(\alpha - b'(x))$ for all x in $(-\epsilon, \epsilon)$. Now we pick any a such that $0 < a < \epsilon$. Then $W_a(\pm a) = 0$ and from (5.5) we have $LW_a > 0$ on $(-a, a)$. Thus, $W_a < 0$ on $(-a, a)$ by applying the maximum principle to (5.5). Hence $a \in A$.

Consider now part (b). To show A is bounded we follow the proof of Lemma 3.5.

To show that $a^* \in A$, first let $a_1 < a_2$, where both $a_1, a_2 \in A$. Then, since W_{a_1} and W_{a_2} both solve the same differential equation in (5.5) on $[-a_1, a_1]$, it follows that $V_{a_2}''(x) < V_{a_1}''(x)$ on $[-a_1, a_1]$. Now let $a_n \to a^*$, where all the a_n are in A. Then if $|x| < a^*$, $V_{a^*}''(x) = \lim_{n\to\infty} V_{a_n}''(x) < -2$, where we have used the continuity in a noted above. Thus $a^* \in A$.

Since $a^* \in A$, $V_{a^*}'''(a^*-) \geq 0$. We show that if $a \in A$ and $V_a'''(a) > 0$, then $a < a^*$. This will complete the proof of part (b). Let $\tilde{V}_a$ denote the function which satisfies (5.3) for all x and coincides with V_a on $[-a, a]$. If we assume that $V_a'''(a-) > 0$, then $V_a''(x) > -2$ in an interval $(a, a + \epsilon)$ for some positive ϵ. Consider V_{a_1} for some $a < a_1 < a + \epsilon$. Then $V_{a_1}''(a_1) = -2 < \tilde{V}_a(a_1)$. By uniqueness of solutions to (5.7), it then follows that $V_{a_1}''(x) < \tilde{V}_a''(x)$ for all x. In particular $V_{a_1}''(x) < -2$ on $[-a, a]$. Let $a_2 = \inf\{x\,;\, V_{a_1}''(x) \geq -2\}$. Then $a_1 \geq a_2 > a$ and $a_2 \in A$ because V_{a_1} and V_{a_2} will coincide on $[-a_2, a_2]$. Thus $a < a^*$, as claimed. □

According to Lemma 5.2, V_{a^*} is the optimal value function if the second condition in (5.4) holds. This requires further conditions on b and c. We state and prove two theorems in this vein. Define the conditions: for some $\delta_0 > 0$ and some $x_0 > 0$,

(H.1) $\{\hat{c} < 0\} = (-\delta_0, \delta_0)$ and $\hat{c}$ increases on $[0, \delta_0]$,

(H.2) $c'(x) \geq 2b(x)$ for $x \geq x_0$,

(H.3) $b'''(x) < 0$ on $[0, \delta_0]$ and $b''(x) \leq 0$ on $[0, x_0]$, and

(H.4) g_c has exactly one zero in $[0, \delta_0]$.

Remark 5.1. The assumption (H.2) is not very restrictive when b is negative and decreasing on $(0, \infty)$, since, typically, $\lim_{|x|\to\infty} c'(x) = 0$.

In regard to (H.4), observe that (H.1) implies $g_c(0) = \hat{c}(0) < 0$. If we assume $c'(x) - 2b'(x)$ is negative on $[0, \delta_0]$, it follows from (H.1) and (H.3) that $g_c(\delta_0) > 0$, and hence g_c has at least one zero on $(0, \delta_0)$.

THEOREM 5.4. *Assume* (H.1)–(H.4) *in addition to* (B.1)–(B.2) *and* (C.1)–(C.3). *Then* A *is nonempty,* V_{a^*} *is the optimal value function, and* $u(x) = \sigma_0 \mathbf{1}_{(-a^*, a^*)}(x)$ *is an optimal feedback control.*

The proof is established through several lemmas. Using the assumptions (C.1) and (C.2) and applying the maximum principle [17], one can easily show that for each $a > 0$, $V_a'(x) \leq 0$ for all $x > 0$. This will be used in the following proofs.

Lemma 5.5. *Assume* (H.1), (H.2), *and* (H.3) *as well as* (B.1)–(B.2) *and* (C.1)–(C.3). *Then*

(a) $W_{a^*}(x) \geq 0$ *on* $[x_0, \infty)$ *where* x_0 *is defined in* (H.2),

(b) A *is nonempty and* $a^* = \sup A \leq x_0$, *and*

(c) $V'''_{a^*}(x) \geq 0$ *on* $[0, a^*)$.

Proof. On $|x| > a$, differentiate the equality $b(x)V'_a(x) - \alpha V_a(x) + c(x) = 0$ to derive

$$b(x)W_{a^*}(x) = (\alpha - b'(x))V'_{a^*}(x) - (c'(x) - 2b(x)), \qquad |x| > a. \tag{5.8}$$

Since $V'_a(x) \leq 0$ for $x \geq 0$, and, by (B.2), $b(x) < 0$ for $x > 0$, and, by (H.2), $c'(x) - 2b(x) \geq 0$ for $|x| \geq x_0$, conclusion (a) follows. The assumption $\hat{c}(0) < 0$ in (H.1) implies that A is nonempty, as in Lemma 5.3. Also, $W_{a^*}(x) < 0$ for $|x| < a^*$, and hence $a^* \leq x_0$. The proof of part (c) follows from the maximum principle applied to $LW'_{a^*} = -[\hat{c}' + 3b''W_{a^*} + b'''V_{a^*}]$ in $[0, a^*)$ and $W'_{a^*}(0) = W'_{a^*}(a^*-) = 0$. □

The final lemma finishes the proof of Theorem 5.4 by verifying (5.4).

Lemma 5.6. *Let the assumptions of Theorem* 5.4 *hold. Then* $V''_{a^*} + 2 \geq 0$ *on* $[a^*, \infty)$.

Proof. Because of the previous lemma, it suffices to show that $W_{a^*} \geq 0$ on $[a^*, x_0]$. Introduce ξ in $[a^*, x_0]$ by $W_{a^*}(\xi) = \min_{[a^*, x_0]} W_{a^*}(x)$. Assume that $W_{a^*}(\xi) < 0$. Then $a^* < \xi < x_0$ since $W_{a^*}(x_0) \geq W_{a^*}(a^*) = 0$. By differentiation of (5.8),

$$bW'_{a^*} - (\alpha - 2b')\,W_{a^*} = -\hat{c} - b''V'_{a^*} \qquad \text{for } |x| > a^*. \tag{5.9}$$

Evaluating at $x = \xi$ implies $\hat{c}(\xi) + b''(\xi)V'_{a^*}(\xi) < 0$. But since $V'_{a^*} \leq 0$ on $(0, \infty)$ and $b''(x) \leq 0$ on $[0, x_0]$, we must have $\hat{c}(\xi) < 0$, and therefore $a^* < \xi < x_0 \leq \delta_0$. However, this is not possible. By (H.4), g_c has exactly one zero in $[0, \delta_0]$, and $g_c(0) = \hat{c}(0) < 0$, $g_c(a^*) \geq 0$. Therefore, $g_c > 0$ on $(a^*, x_0]$. For all $x > a$, W_{a^*} satisfies (5.6) and $W_{a^*}(a^*) = 0$. Also, $b(x) < 0$ for $x > 0$, and hence (5.6) implies that $W_{a^*} > 0$ on $(a^*, \delta_0]$. This contradicts $W_{a^*}(\xi) < 0$ for $a^* < \xi < \delta_0$. Hence $W_{a^*} \geq 0$ on $[a^*, x_0]$. □

In the next theorem, we show optimality of one-region control under hypotheses that require $b'' \geq 0$ on $(0, x_0)$, rather than (H.3).

Theorem 5.7. *Assume that* (B.1), (B.2), (C.1)–(C.3), *and* (H.2) *hold and that* $\hat{c}(0) < 0$. *Assume also that* $b''' \geq 0$, $b'' \geq c'''/6$, *and* $b'' > 0$ *on* $(0, x_0)$, *where* x_0 *is as in* (H.2). *Then the conclusions of Theorem* 5.4 *hold.*

Proof. Because we assume $\hat{c}(0) < 0$, we know that A is nonempty, and we conclude from Lemma 5.1 and the smooth fit condition (3.16) that V_{a^*} is a bounded, C^2 function. Thus, we need only to verify $W_{a^*} \geq 0$ on (a^*, ∞). Using (H.2), we can follow the proof of Lemma 5.5 to conclude that $W_{a^*} \geq 0$ on $[x_0, \infty)$. Hence it remains to show that W_{a^*} is nonnegative on $[a^*, x_0]$. By differentiating (5.9) on the interval (a^*, x_0) once more, we obtain

$$-bW''_{a^*} + (\alpha - 3b')\,W'_{a^*} - 3b''W_{a^*} = c''' - 6b'' + b'''V'_{a^*}. \tag{5.10}$$

We also know $W_{a^*}(a^*) = 0$, and $W_{a^*}(x_0) \geq 0$. The maximum principle applied to (5.10) on the interval $[a^*, x_0]$, using the assumptions placed on b'' and b''', implies that W_{a^*} cannot admit a strictly negative minimum in (a^*, x_0). This completes the proof. □

6. Conclusion. We have derived some explicit characterizations of the optimal control and optimal value functions in scalar problems with possibly degenerate variance control in which the variance is used to keep the diffusion away from regions of high running cost.

Throughout our analysis the discount rate α and the control cost multiplier λ were fixed. It is interesting to ask how the solutions behave as either $\lambda \to \infty$ or $\alpha \to 0$, with appropriate scaling of the cost. When $\alpha \to 0$ and the cost is scaled by α, one expects a limit which corresponds to an average cost control problem. The paper of Assaf [1] treats an example of average cost minimization using unbounded variance control directly.

REFERENCES

[1] D. Assaf, *Estimating the state of a noisy continuous time Markov chain when dynamic sampling is feasible,* Ann. Appl. Probab., 3 (1997), pp. 822–836.

[2] S. Assing and W. Schmidt, *Continuous Strong Markov Processes in Dimension One*, Springer-Verlag, New York, 1991.

[3] J.R. Dorroh, G. Ferreyra, and P. Sundar, *A technique for stochastic control problems with unbounded control set,* J. Theoret. Probab., 12 (1999), pp. 255–270.

[4] S. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[5] W.H. Fleming and H. Mete Soner, *Controlled Markov Process and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[6] A. Friedman, *Stochastic Differential Equations and Applications*, Vol. 1, Academic Press, New York, 1975.

[7] I.L. Genis and N.V. Krylov, *An example of a one-dimensional control process,* Theory Probab. Appl., 21 (1976), pp. 148–152.

[8] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, New York, 1989.

[9] I. Karatzas and S.E. Shreve, *Connections between optimal stopping and singular stochastic control,* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.

[10] I. Karatzas and S. Shreve, *Equivalent models for finite-fuel stochastic control,* Stochastics, 18 (1986), pp. 245–276.

[11] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Lecture Notes in Math. 1688, Springer-Verlag, New York, 1998.

[12] N.V. Krylov, *Control of a solution of a stochastic integral equation with degeneration*, Theory Probab. Appl., 17 (1973), pp. 114–131.

[13] P.-L. Lions, *Control of diffusions in R^n,* Comm. Pure Appl. Math., 34 (1981), pp. 121–147.

[14] P. Mandl, *Analytical Treatment of One-Dimensional Markov Processes*, Springer-Verlag, New York, 1968.

[15] J.M. McNamara, *Optimal control of the diffusion coefficient of a simple diffusion process,* Math. Oper. Res., 8 (1983), pp. 373–380.

[16] J.M. McNamara, *Control of a diffusion by switching between two drift-diffusion coefficient pairs,* SIAM J. Control Optim., 22 (1984), pp. 87–94.

[17] M.H. Protter and H.F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984.

[18] L.C.G. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales*, 2, Wiley, New York, 1987.

[19] D.W. Stroock and S.R.S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

SIAM J. Control Optim.
Vol. 39, No. 1, pp. 25–48

FEEDBACK STABILIZATION AND LYAPUNOV FUNCTIONS[*]

F. H. CLARKE[†], YU. S. LEDYAEV[‡], L. RIFFORD[†], AND R. J. STERN[§]

Abstract. Given a locally defined, nondifferentiable but Lipschitz Lyapunov function, we employ it in order to construct a (discontinuous) feedback law which stabilizes the underlying system to any given tolerance. A converse result shows that suitable Lyapunov functions of this type exist under mild assumptions. We also establish that the feedback in question possesses a robustness property relative to measurement error, despite the fact that it may not be continuous.

Key words. asymptotic stabilizability, discontinuous feedback law, system sampling, locally Lipschitz Lyapunov function, nonsmooth analysis, robustness

AMS subject classifications. 93D05, 93D20, 34D20, 26B05

PII. S0363012999352297

1. Introduction.

Consider a standard control system of the form

$$\dot{x}(t) = f(x(t), u(t)) \quad \text{almost everywhere (a.e.)}, \qquad u(t) \in \mathcal{U}, \tag{1.1}$$

and let V be a smooth Lyapunov function for the system; thus we have

$$V(x) \geq 0, \qquad V(x) = 0 \text{ iff } x = 0, \qquad V(x) \to \infty \text{ as } \|x\| \to \infty$$

and (for some function W) the infinitesimal decrease condition

$$\min_{u \in \mathcal{U}} \langle \nabla V(x), f(x,u) \rangle \leq -W(x) < 0, \qquad x \neq 0. \tag{1.2}$$

It is well known (but true) that the existence of such a "control-Lyapunov function" (V, W) (a framework introduced by Eduardo Sontag) implies (open-loop) asymptotic controllability to the origin: for every $\alpha \in \mathbb{R}^n$, there is a control $u(t)$ such that the solution $x(\cdot)$ of (1.1) with initial condition $x(0) = \alpha$ satisfies $x(t) \to 0$ as $t \to \infty$. (In addition, convergence to 0 takes place in a certain uniform and stable manner that we will not dwell upon here.) A related and important goal in many situations is to produce a state feedback $k(\cdot) : \mathbb{R}^n \to \mathcal{U}$ which stabilizes the system, i.e., such that the system $\dot{x} = f(x, k(x))$ is globally asymptotically stable. This article explores the question of how to define such a feedback law through the use of a given Lyapunov function V.

The ideal case, a well-known heuristic useful for motivational purposes, is the one in which we can find a continuous function $k(x)$ that selects a value of $u \in \mathcal{U}$ attaining

[*] Received by the editors February 1, 1999; accepted for publication (in revised form) December 17, 1999; published electronically August 17, 2000.
http://www.siam.org/journals/sicon/39-1/35229.html

[†] Institut Girard Desargues (Bât. 101), Université Claude Bernard Lyon I (La Doua), 69622 Villeurbanne, France (clarke@desargues.univ-lyon1.fr, rifford@desargues.univ-lyon1.fr).

[‡] Steklov Institute of Mathematics, Moscow 117966, Russia (ledyaev@math-stat.wmich.edu). The research of this author was supported in part by Russian Fund for Fundamental Research grant 96-01-00219. Current address: Department of Mathematics, Western Michigan University, Kalamazoo, MI 49008.

[§] Department of Mathematics and Statistics, Concordia University, Montreal, Quebec H4B 1R6, Canada (stern@vax2.concordia.ca). The research of this author was supported by the Natural Sciences Engineering Research Council of Canada and Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec.

(or almost) the minimum in (1.2):

$$\langle \nabla V(x), f(x, k(x))\rangle \le -W(x) \quad \forall x \neq 0.$$

Then any solution of $\dot{x} = f(x, k(x))$ is such that

$$\frac{d}{dt} V(x(t)) = \langle \nabla V(x(t)), \dot{x}(t)\rangle \le -W(x(t)) < 0,$$

a monotonicity conclusion that, together with the growth property of V, assures that $x(t) \to 0$ as $t \to \infty$.

There are two fundamental difficulties with this ideal picture, and both concern regularity issues. The first is that a differentiable Lyapunov function may not exist, and the second is that even when a smooth V exists, the continuous selection $k(\cdot)$ does not generally exist. If we have recourse to a discontinuous feedback $k(\cdot)$, then the issue arises of how to interpret the discontinuous differential equation $\dot{x} = f(x, k(x))$.

The primary goal of this article is to give a general answer to the problem of defining a (discontinuous) stabilizing feedback based upon a given (nondifferentiable) Lyapunov function, one for which a nonsmooth version of infinitesimal decrease is known to hold only on a restricted set. The construction is described in section 1, while section 2 establishes the converse result that under mild conditions, a Lyapunov function of the type required in the previous section always exists. In the final section, we address the issue of robustness of the feedback with respect to measurement error and small perturbations of the dynamics, a particularly important issue when discontinuity is present. Some works and general references related to this article include [2, 3, 8, 11, 12, 13, 15, 17, 18, 20, 23, 24, 26, 27, 28, 29, 31, 32]. We proceed now to situate our results with respect to the literature.

The possible nonexistence of continuous stabilizing feedback was brought to light in the seminal work of Sontag and Sussmann [30] and of Brockett [4]. The latter developed a necessary condition for continuous stabilizability and adduced the following example, the "nonholonomic integrator":

$$\begin{cases} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= u_2, \qquad\qquad (u_1, u_2) \in \bar{B} =: \mathcal{U}. \\ \dot{x}_3 &= x_1 u_2 - x_2 u_1, \end{cases}$$

This system is globally asymptotically controllable yet fails to admit a continuous stabilizing feedback (by Brockett's condition). In considering the use of discontinuous feedback laws $k(\cdot)$, one could have recourse to the Filippov solution concept [14]: x is said to be a solution of $\dot{x} = f(x, k(x)) =: g(x)$ provided that we have

$$\dot{x} \in \bigcap_{\substack{\delta > 0 \\ \mathrm{meas}(\Omega) = 0}} \mathrm{cl}\,\mathrm{co}(g([x + \delta B] \setminus \Omega)).$$

However, as shown by Ryan [25] and by Coron and Rosier [12], Brockett's condition continues to hold for this solution concept, so that the nonholonomic integrator (for example) cannot be stabilized by a discontinuous feedback in the Filippov sense.

In [6] it was shown that any globally asymptotically controllable system is stabilizable by a (possibly discontinuous) feedback if the trajectory $x(\cdot)$ associated to the feedback is defined in a natural way that involves discretizing the control law (closed-loop system sampling) in a manner similar to that used in differential games by Krasovskii and Subbotin [19]. We proceed now to describe this concept, which is the one used in this article.

Let $\pi = \{t_i\}_{i\geq 0}$ be a partition of $[0, \infty)$, by which we mean a countable, strictly increasing sequence t_i with $t_0 = 0$ such that $t_i \to \infty$ as $i \to \infty$. The diameter of π, denoted $\operatorname{diam}(\pi)$, is defined as $\sup_{i\geq 0}(t_{i+1} - t_i)$. Given an initial condition x_0, the π-trajectory $x(\cdot)$ corresponding to π and an arbitrary feedback law $k : \mathbb{R}^n \to \mathcal{U}$ are defined in a step-by-step fashion as follows. Between t_0 and t_1, x is a classical solution of the differential equation

$$\dot{x}(t) = f(x(t), k(x_0)), \qquad x(0) = x_0, \qquad t_0 \leq t \leq t_1.$$

(Of course in general we do not have uniqueness of the solution, nor is there necessarily even one solution, although nonexistence will be ruled out by the feedback constructed in section 1, which will preclude blow-up of the solution in finite time.) We then set $x_1 := x(t_1)$ and restart the system with control value $k(x_1)$:

$$\dot{x}(t) = f(x(t), k(x_1)), \qquad x(t_1) = x_1, \qquad t_1 \leq t \leq t_2,$$

and so on in this fashion. The resulting trajectory x is a physically meaningful one that corresponds to a natural sampling procedure and piecewise constant controls; the smaller $\operatorname{diam}(\pi)$, the greater the sampling rate. Since our results are couched in terms of π-trajectories, the issue of defining a solution concept for discontinuous differential equations is effectively sidestepped. Our approach will lead to precise estimates of how small the step size $\operatorname{diam}(\pi)$ must be for a prescribed stabilization tolerance to ensue, and of the resulting stabilization time, in terms of the given data.

The next major point to address concerns the nonsmoothness of the Lyapunov function V. An early and important result of Artstein [1] implies in particular that the nonholonomic integrator fails to admit a smooth V (see [7] for related results). It has been shown by Sontag [26], however, that globally asymptotically controllable systems always admit a continuous Lyapunov function V satisfying the following nonsmooth version of the infinitesimal decrease condition:

$$\inf_{u\in\mathcal{U}} DV(x; f(x,u)) \leq -W(x) < 0, \qquad x \neq 0, \tag{1.3}$$

where the lower Dini derivate DV is defined by

$$DV(x; v) := \liminf_{\substack{t\downarrow 0 \\ v'\to v}} \frac{V(x + tv') - V(x)}{t}. \tag{1.4}$$

Among the several important ways in which the theory of nonsmooth analysis intervenes in this article is that of asserting the equivalence to (1.3) of another, and for our purposes more useful, form of the infinitesimal decrease condition:

$$\inf_{u\in\mathcal{U}} \langle f(x,u), \zeta\rangle \leq -W(x) < 0 \quad \forall x \neq 0, \qquad \forall \zeta \in \partial_P V(x). \tag{1.5}$$

Here $\partial_P V(x)$ refers to the proximal subdifferential of V at x (which may very well be empty); ζ belongs to $\partial_P V(x)$ iff there exist σ and $\eta > 0$ such that

$$V(y) - V(x) + \sigma\|y - x\|^2 \geq \langle \zeta, y - x\rangle \qquad \forall y \in x + \eta B.$$

(B denotes the open unit ball, and the open ball of radius r centered at x is written as either $x + rB$ or $B(x, r)$.) The equivalence of (1.3) and (1.5) is a consequence of Subbotin's theorem, which links Dini derivates to proximal calculus (see for example [6] or [9], our principal sources for the theory of nonsmooth analysis).

The essential reason for which proximal calculus is well suited to our approach has to do with its relation to metric projection onto sets, upon which is based the "proximal aiming" method that we employ. The crux is this: when $x(t_i) = x$ lies outside a level set $S = S(c) := \{V \leq c\}$ and admits closest point (or projection) s in S, then $x - s$ is a "proximal normal" to S at s, and for some $\lambda > 0$ we have $\lambda(x - s) \in \partial_P V(s)$. Then (1.5) can be invoked at s to find a suitable value of the control u which moves the state toward S, in the sense that the Euclidean distance d_S decreases at a certain positive rate Δ:

$$d_S(x(t)) - d_S(x(t_i)) \leq -\Delta(t - t_i), \qquad t_i \leq t \leq t_{i+1},$$

provided $x(t_i)$ is close enough to S to start with and provided $\operatorname{diam}(\pi)$ is small enough. A sequence of such feedbacks is amalgamated in the first section to produce the stabilizing feedback $k(\cdot)$ that is sought.

Our approach requires the Lyapunov function V to be Lipschitz (in the zone under consideration). In the case of a system which is globally asymptotically controllable to the origin, which has received considerable attention, it is not known whether a suitable V exists that is Lipschitz near 0, but that is somewhat beside the point in the setting in which we work. Theorem 2.1 derives finite-time stabilizability, to a close approximation of some level set $S(a)$, as a consequence of the supposed existence of a Lipschitz Lyapunov function. When applied to the special case of global asymptotic controllability, this requires only a Lipschitz V defined on the complement of a small ball around the origin, and the stabilization takes place not asymptotically to the origin, but rather in finite time to a small neighborhood of it. (This has been called "practical" stabilization.) In contrast, [6] obtains asymptotic stabilizability to the origin (the case $S(a) = \{0\}$); the proof uses the Moreau–Yosida inf convolution to make a continuous Lyapunov function Lipschitz as an intermediate step. This methodology was also employed earlier [10], in a differential game setting. The direct use of a Lipschitz Lyapunov function, when it is possible, leads to a far more transparent feedback construction with direct stabilization estimates and has the important consequence of yielding robustness, as we discuss presently. The fact that under mild assumptions suitable Lipschitz Lyapunov functions do exist leading to practical stabilization to any required tolerance is proven in section 2.

Ledyaev and Sontag [22] have recently proved that there is a close relationship between the issues of "how regular a Lyapunov function does the system admit" and "how robust a stabilizing feedback does the system admit." Consider, for example, a perturbed equation $\dot{x} = f(x, k(x + p))$, where p represents a measurement error. Full robustness of the feedback k is taken to mean that for any ϵ, there is a $\delta > 0$ such that whenever the perturbation $p(t)$ satisfies $\|p(t)\| \leq \delta$ for all t, then stabilization to the ϵ-ball takes place. Then [22] asserts that the system admits a fully robust stabilizing feedback iff it admits a smooth (C^1 or C^∞) Lyapunov function. Thus the nonholonomic integrator, which *can* be stabilized by a discontinuous feedback (in view of [6]), does *not* admit a fully robust stabilizing feedback. It is possible to recover robustness, however, through the use of a dynamic feedback; see [21]. The issue of the robustness of discontinuous feedbacks with respect to measurement error seems to have been raised first by Hermes [16].

The concept of full robustness, unrelated as it is to the system sampling method that we employ, is not the one discussed in this article. Instead, we introduce a type of relative robustness in which we require the size of the measurement error to be limited as a function of the maximum step size δ of the underlying partition. This step size δ must still be small enough (for stabilization), but at the same time the individual steps must be big enough to preclude a possible chattering phenomenon, even in the presence of small errors. This consideration, which leads us to specify "reasonably uniform" sampling in section 3, appears to be new in this context. The term "reasonably uniform" is taken here to mean that the following holds:

$$\frac{\delta}{2} \leq t_{i+1} - t_i \leq \delta \quad \forall i \geq 0,$$

although it is possible to replace the factor 1/2 by any constant in $(0, 1)$.

To conclude with the nonholonomic integrator, then, it turns out that the system does admit a relatively robust stabilizing feedback to within any prescribed tolerance r, in the sense that for all initial conditions in a bounded set, we will have $\|x(t)\| \leq r$ for all $t \geq T$, whenever x is a π-trajectory, if π is a reasonably uniform partition whose diameter is sufficiently small, and whenever measurement and external error do not exceed a critical level related to the sampling rate. It is possible to exhibit a Lipschitz Lyapunov function for the system, and to make explicit the resulting feedback, as will be shown in a forthcoming article.

2. A feedback construction. For a given function $V : \mathbb{R}^n \longrightarrow (-\infty, \infty]$, we shall deal frequently with the sublevel sets $S(r)$ defined as follows:

$$S(r) := \{x \in \mathbb{R}^n : V(r) \leq r\}.$$

In addition, the following sets are considered:

$$S(a, b) := \{x \in \mathbb{R}^n : a \leq V(x) \leq b\}.$$

Let a and b be two given numbers with $a < b$. The following hypotheses are made concerning the function V and the system function f.

(H1) V is lower semicontinuous, $S(b) \neq \emptyset$, and for some $\eta > 0$, V is Lipschitz of rank L_V on $S(a, b) + \eta B$:

$$|V(x) - V(y)| \leq L_V \|x - y\| \quad \forall (x, y) \in S(a, b) + \eta B.$$

(H2) There exists $\delta_1 \in (0, b - a)$ and $\delta_2 > 0$ such that

$$S(a + \delta_1) + \delta_2 B \subset S(b).$$

(H3) $f(x, u)$ is continuous on $S(b) + \eta B$ as a function of x for each $u \in \mathcal{U}$, and there exists $m > 0$ such that

$$\|f(x, u)\| \leq m \quad \forall x \in S(b) + \eta B, \qquad \forall u \in \mathcal{U}.$$

(H4) f is Lipschitz in x of rank L_f on $S(a, b) + \eta B$:

$$\|f(x, u) - f(y, u)\| \leq L_f \|x - y\| \quad \forall (x, y) \in S(a, b) + \eta B, \qquad \forall u \in \mathcal{U}.$$

(H5) There exists $\omega > 0$ such that, for every $x \in S(a, b) + \eta B$, we have

$$\inf_{v \in \operatorname{co} f(x, \mathcal{U})} DV(x; v) \leq -\omega.$$

Remark 2.1. We do not require that f and V be defined except on $S(b)+\eta B$; the Lipschitz conditions on these functions, as well as the infinitesimal decrease condition (H5), are posited only on a neighborhood of $S(a,b)$. No hypotheses are made concerning the abstract set $\mathcal{U}$, nor on the nature of the dependence of f on the control variable. It is shown in section 2 that (H2) automatically holds when $S(b)$ is compact and V is continuous on $S(b)+\eta B$; see Lemma 2.9. The set $S(a)$ is not assumed to be nonempty a priori, but that fact is a consequence of the following theorem.

THEOREM 2.1. *For any $\gamma > 0$ sufficiently small, there exist positive numbers δ, T, and a feedback $k : S(b)+\eta B \longrightarrow \mathcal{U}$ such that whenever a partition π satisfies* $\operatorname{diam}(\pi) < \delta$, *then any π-trajectory $x(\cdot)$ having $x(0) \in S(b)+\gamma B$ satisfies*

$$x(t) \in S(b)+\gamma B \quad \forall t \geq 0,$$

$$x(t) \in S(a)+\gamma B \quad \forall t \geq T.$$

Remark 2.2. Thus we almost recover the conclusion of the "ideal case" discussed in the Introduction (with $\{0\}$ replaced by the more general $S(a)$), but in approximate terms (to tolerance γ), with a discontinuous feedback, and for a nonsmooth Lyapunov function satisfying only local hypotheses. The proof is constructive and provides explicit estimates of γ, δ, and T in terms of the given data.

Remark 2.3. Taking $W(x) = \omega$ for purposes of comparison, note that (H5) is an apparently weaker hypothesis than (1.3); in fact, each is equivalent to (1.5). Because V is Lipschitz and f continuous in x near the points in question, this in turn is equivalent to the following:

$$\inf_{u \in \mathcal{U}} \langle \zeta, f(x,u) \rangle \leq -\omega \quad \forall x \in S(a,b)+\eta B, \qquad \forall \zeta \in \partial_L V(x),$$

where $\partial_L V$ is the limiting subdifferential of V (see [9]).

Proof of Theorem 2.1. The proof of Theorem 2.1 is based upon defining a feedback control via projections. The first two lemmas below guarantee that the projections lie in the set where the hypotheses are active.

LEMMA 2.2. *Let ϵ lie in $[0, \delta_1]$ and suppose that x is a point in the set*

$$[S(a+\epsilon) + \min\{\delta_2, \eta\} B] \setminus S(a+\epsilon).$$

Then $x \in S(a,b)$, and if $s \in \operatorname{proj}(x, S(a+\epsilon))$, then $s \in S(a,b)+\eta B$.

Proof. Since we have $S(a+\delta_1)+\delta_2 B$ contained in $S(b)$ by hypothesis, it follows that x lies in $S(b)$. Since x does not belong to $S(a+\epsilon)$, we deduce $x \in S(a,b)$. Finally, we have

$$\|s-x\| < \min\{\delta_2, \eta\} \leq \eta,$$

whence $s \in S(a,b)+\eta B$. □

LEMMA 2.3. *Let $0 < \gamma < \eta/2$, and suppose that for some r' and r with $a \leq r' < r \leq b$ we have $x \in [S(r)+\gamma B] \setminus [S(r')+\gamma B]$. Then $x \in S(a,b)+\gamma B$, and if $s \in \operatorname{proj}(x, S(r))$, then $s \in S(a,b)+\eta B$.*

Proof. There exists $y \in S(r)$ having $\|y-x\| < \gamma$. Since x does not belong to $S(r')+\gamma B$, we have $V(y) > r'$ necessarily. Thus $y \in S(a,b)$ and $x \in S(a,b)+\gamma B$. Finally, we note $\|x-s\| < \gamma$, whence $\|y-s\| < 2\gamma$ and s lies in $S(a,b)+\eta B$. □

The next "solvability" result is central to our approach. The notation u_+ stands for $\max\{u, 0\}$.

LEMMA 2.4. *For any $r \in [a,b]$, for any $x \in S(a,b)$, we have*

$$d(x, S(r)) \leq \frac{m}{\omega}(V(x) - r)_+.$$

Proof. We shall invoke results (and terminology) from [9] to give a short proof of this result, whose proof from first principles would be lengthy.

We define a lower semicontinuous function $g : \mathbb{R}^n \longrightarrow [0, \infty]$ as follows:

$$g(x) := (V(x) - r)_+ + I_{S(b)}(x),$$

where $I_{S(b)}(\cdot)$ is the indicator function of the set $S(b)$. At any point x in the open set $C := \{y : g(y) > 0\}$ at which g is finite, we have $x \in S(a,b)$, and the infinitesimal decrease condition implies that

$$\inf\{Dg(x; v) : v \in \mathrm{co} f(x, \mathcal{U})\} \leq -\omega.$$

It follows from this that for any $\epsilon > 0$, for any $x \in C$, for any $\zeta \in \partial_P g(x)$, there exists $u \in \mathcal{U}$ such that

$$\langle \zeta, f(x,u) \rangle \leq -\omega + \epsilon.$$

Since $\|f(x,u)\| \leq m$ and since $\epsilon > 0$ is arbitrary, we derive $\|\zeta\| \geq \omega/m$. This verifies the hypothesis of the solvability theorem [9, Theorem 3.3.1] (with the sets labeled there as V and Ω both taken to be $\mathbb{R}^n$), whose conclusion is precisely the desired one since $S(r) = \{x : g(x) = 0\}$. We remark that an alternate proof can be based upon weak monotonicity: the infinitesimal decrease condition implies the existence of a trajectory x with $x(0) = x$ and along which $V(x(t)) + t\omega$ is decreasing (see [9, Theorem 4.5.7]), which implies the result. □

We now proceed to fix $\gamma > 0$ such that

$$\gamma < \min\left\{\delta_1, \frac{\eta}{2}, \frac{\omega}{12L_f L_V}\right\}, \tag{2.1}$$

and we define

$$\beta := \min\left\{\delta_1, \frac{(b-a)}{2}, \frac{\gamma\omega}{4m}\right\}. \tag{2.2}$$

Let N be the first integer such that

$$b - N\beta > a \geq b - (N+1)\beta.$$

Note that $N \geq 1$ since $\beta < b - a$. We proceed to define certain sets $\Omega_i (i = 0, 1, \ldots, N+1)$ that lie at the heart of our construction.

For $0 \leq i \leq N-1$, we set

$$\Omega_i := [S(b - i\beta) + \gamma B] \setminus [S(b - (i+1)\beta) + \gamma B];$$

for $i = N$ we set

$$\Omega_N := [S(b - N\beta) + \gamma B] \setminus \left[S(b - N\beta) + \frac{\gamma}{4}B\right];$$

and finally, we define

$$\Omega_{N+1} := S(b - N\beta) + \frac{\gamma}{4}B.$$

We now gather some facts about these sets.

LEMMA 2.5.

(a) *The Ω_i are disjoint, and Ω_i is contained in $S(a,b) + \gamma B$ for $i \leq N$.*

(b) $\bigcup_{i=0}^{N+1} \Omega_i = S(b) + \gamma B$.

(c) *If $x \in \Omega_i$ for some $i \in \{0, 1, \ldots, N\}$ and $s \in \mathrm{proj}(x, S(b - i\beta))$, then $s \in S(a,b) + \eta B$.*

(d) $S(b - i\beta) + \frac{\gamma}{4}B \subset S(b - (i+1)\beta) + \gamma B (i = 0, 1, \ldots, N-1)$.

(e) *For every $i \in \{0, 1, \ldots, N\}, \forall x \in \Omega_i$, we have $\frac{\gamma}{4} \leq d(x, S(b - i\beta)) < \gamma$.*

(f) $S(b - N\beta) + \frac{\gamma}{2}B \subset S(a) + \gamma B$, *so that $\Omega_{N+1} \subset S(a) + \gamma B$.*

Proof.

(a) That the Ω_i are disjoint is evident; that they lie in $S(a,b) + \gamma B$ for $i \leq N$ follows from Lemma 2.3 for $i < N$ and from Lemma 2.2 for $i = N$ (recall that $b - N\beta - a \leq \beta \leq \delta_1$ and $\gamma < \delta_2$).

(b) Evident.

(c) Direct from Lemma 2.3 ($i < N$) or Lemma 2.2 ($i = N$).

(d) Let x lie in $S(b - i\beta) + \gamma/4B$, and let $s \in S(b - i\beta)$ satisfy $\|x - s\| < \gamma/4$. Then $V(s) \leq b - i\beta$, and if $V(s) \leq b - (i+1)\beta$ the conclusion is immediate. Otherwise we have

$$V(s) > b - (i+1)\beta > a,$$

so that $s \in S(a,b)$. By Lemma 2.4 there exists $y \in S(b - (i+1)\beta)$ such that

$$\|s - y\| \leq \frac{m}{\omega}[V(s) - b + (i+1)\beta] \leq \frac{m\beta}{\omega} \leq \frac{\gamma}{2}$$

in view of (2.2). Then

$$\|x - y\| \leq \|x - s\| + \|s - y\| < \frac{\gamma}{4} + \frac{\gamma}{2} < \gamma,$$

which establishes the desired conclusion.

(e) For $i = N$, this is immediate from the definition of Ω_N; for $i < N$, it is a consequence of (d).

(f) Let x belong to $S(b - N\beta) + \gamma/2B$, and let $s \in S(b - N\beta)$ satisfy $\|x - s\| < \gamma/2$. If $V(s) \leq a$, then $x \in S(a) + \gamma B$. Otherwise, s belongs to $S(a,b)$, and Lemma 2.4 implies the existence of $y \in S(a)$ such that

$$\|y - s\| \leq \frac{m}{\omega}[V(s) - a] \leq \frac{m}{\omega}[b - N\beta - a] \leq \frac{m\beta}{\omega} \leq \frac{\gamma}{2}.$$

But then $\|x - y\| < \gamma$, so again $x \in S(a) + \gamma B$. □

LEMMA 2.6. *Let $x \in \Omega_i$ ($i = 0, 1, \ldots, N$), and let $s \in \mathrm{proj}(x, S(b - i\beta))$. Then there exists $u \in \mathcal{U}$ such that*

$$\langle x - s, f(s,u) \rangle \leq \frac{-\omega}{2L_V}\|x - s\|.$$

Proof. By definition, $x - s$ lies in the proximal normal cone $N^P(s, S(b - i\beta))$. Note that s lies in $S(a,b) + \eta B$ (by Lemma 2.3 for $i < N$, by Lemma 2.2 for $i = N$), so that V is Lipschitz of rank L_V in a neighborhood of s. A basic calculus result [9, 1.11.26] yields the existence of $\lambda > 0$ such that $\lambda(x-s) \in \partial_L V(s)$ and necessarily $\lambda\|x-s\| \le L_V$. In accord with Remark 2.1, there exists $u \in \mathcal{U}$ such that

$$\langle \lambda(x-s), f(s,u)\rangle \le \frac{-\omega}{2}.$$

The result follows. □

Defining the feedback. We now define a feedback $k(\cdot)$ on $S(b)+\gamma B$ as follows. If $x \in \Omega_i$ for some $i \in \{0,1,\ldots,N\}$, then we set $k(x) = u$, where u is one of the points corresponding to x (and a projection s) as in Lemma 2.6. There remain the points x in Ω_{N+1} to consider (see Lemma 2.5(b)). For such x, we define $k(x)$ to be any point in $\mathcal{U}$. (We remark that we have phrased the definition of $k(x)$ in such a way that the choice of u as indicated above is made once and for all, but in fact a different choice could be made if the same state x recurred subsequently, without at all affecting what follows; this fact is relevant for real-time control.)

The remainder of the proof consists in establishing that for suitably small mesh size, any π-trajectory generated by $k(\cdot)$ with initial condition in $S(b)+\gamma B$ remains in $S(b)+\gamma B$, enters $S(a)+\gamma B$ within a certain (uniform) time and then remains in that set subsequently.

We consider countable partitions $\{t_j\}$ such that $t_0 = 0, \lim_{j\to\infty} t_j = \infty$, and such that $0 < t_{j+1} - t_j \le \delta$ for all $j \ge 0$, where δ is any positive number satisfying

$$\delta < \min\left\{\frac{\gamma}{4m}, \frac{\omega}{6mL_fL_V}, \frac{\gamma\omega}{48m^2L_V}, 1\right\}. \tag{2.3}$$

For such a partition, let x_0 be any point in $S(b)+\gamma B$, and let $x(\cdot)$ be a π-trajectory with $x(0) = x_0$. We denote $x(t_j)$ by x_j, and we set

$$\Delta := \frac{\omega}{60L_V}.$$

Lemma 2.7. *For some $t_j \in \pi$, suppose that $x_j \in \Omega_i, i \in \{0,1,\ldots,N\}$. Then*

$$x(t) \in S(b)+\gamma B \qquad \forall t \in [t_j,t_{j+1}], \text{ and}$$

$$d(x(t), S(b-i\beta)) \le d(x_j, S(b-i\beta)) - \Delta(t-t_j) \qquad \forall t \in [t_j, t_{j+1}].$$

Proof. We have $x_j \in S(a,b)+\gamma B$ by Lemma 2.5(a) and $\|\dot{x}(t)\| \le m$ while $x(t)$ lies in $S(b)+\eta B$. Since $\delta m < \gamma/4$ by (2.3) and $\gamma < \eta/2$, it follows that $x(t)$ lies in $S(a,b)+\eta B$ for $t \in [t_j, t_{j+1}]$, as does the point s that figures in the definition of $k(x_j)$; this was pointed out in the proof of Lemma 2.6, where we also deduced the inequality

$$\langle x_j - s, f(s,k(x_j))\rangle \le \frac{-\omega}{2L_V}\|x_j - s\|. \tag{2.4}$$

We fix $t \in (t_j, t_{j+1})$ and set

$$\Psi := \frac{x(t)-s}{\|x(t)-s\|}.$$

Note that $x(t) \ne s$, since $\|x_j - s\| \ge \gamma/4$ by Lemma 2.5(e) and since

$$\|x(t) - x_j\| < \delta m < \frac{\gamma}{4}.$$

We now observe two inequalities:

$$d(x(t), S(b - i\beta)) \leq \|x(t) - s\| = \langle \Psi, x(t) - s \rangle,$$

$$d(x_j, S(b - i\beta)) = \|x_j - s\| \geq \langle \Psi, x_j - s \rangle.$$

These together imply

$$\begin{aligned} d(x(t), S(b - i\beta)) - d(x_j, S(b - i\beta)) &\leq \langle \Psi, x(t) - x_j \rangle \\ &= \tau \langle \Psi, f_j \rangle, \end{aligned} \tag{2.5}$$

where we introduce the notation $\tau := t - t_j$,

$$x(t) = x_j + \tau f_j, \qquad f_j := \frac{1}{\tau} \int_{t_j}^{t} f(x(r), k(x_j)) dr.$$

We also set

$$\hat{f}_j := f(s, k(x_j)) = \frac{1}{\tau} \int_{t_j}^{t} f(s, k(x_j)) dr.$$

Note that

$$\begin{aligned} \|f_j - \hat{f}_j\| &\leq \frac{1}{\tau} \int_{t_j}^{t} L_f \|x(r) - s\| dr \\ &\qquad \text{(the Lipschitz condition holds since we are in } S(a, b) + \eta B) \\ &\leq \frac{1}{\tau} \int_{t_j}^{t} L_f (\|x(r) - x_j\| + \|x_j - s\|) dr \\ &\leq L_f (\tau m + d(x_j, S(b - i\beta))). \end{aligned}$$

It follows from this and (2.4) that we have

$$\begin{aligned} \langle x_j - s, f_j \rangle &= \langle x_j - s, \hat{f}_j + f_j - \hat{f}_j \rangle \\ &\leq \frac{-\omega}{2L_V} \|x_j - s\| + L_f d(x_j, S(b - i\beta)) \{\tau m + d(x_j, S(b - i\beta))\} \\ &\leq d(x_j, S(b - i\beta)) \left[\frac{-\omega}{2L_V} + L_f \delta m + \gamma L_f \right] \\ &\qquad \text{(since } \tau < \delta \text{ and } d(x_j, S(b - i\beta)) \leq \gamma) \\ &\leq d(x_j, S(b - i\beta)) \left[\frac{-\omega}{2L_V} + \frac{\omega}{6L_V} + \frac{\omega}{6L_V} \right] \\ &\qquad \text{(we have invoked (2.3) and (2.1))} \\ &\leq -\frac{\gamma\omega}{24L_V} \\ &\qquad \text{(since } d(x_j, S(b - i\beta)) \geq \gamma/4 \text{ by Lemma 2.5(e)).} \end{aligned}$$

We shall use this bound on $\langle x_j - s, f_j \rangle$ to derive one on $\langle x(t) - s, f_j \rangle$ as follows:

$$\begin{aligned} \langle x(t) - s, f_j \rangle = \langle x_j + \tau f_j - s, f_j \rangle &\leq \langle x_j - s, f_j \rangle + \tau \|f_j\|^2 \\ &\leq \frac{-\gamma\omega}{24L_V} + \delta m^2 \leq \frac{-\gamma\omega}{48L_V} \end{aligned} \tag{2.6}$$

(in light of (2.3)). We also have

$$\begin{aligned}\|x(t) - s\| = \|x_j + \tau f_j - s\| \le \|x_j - s\| + \tau\|f_j\| &\le \gamma + \delta m \\ &< \frac{5\gamma}{4} \text{ (by (2.3))}.\end{aligned}$$

Combining this with (2.6) we arrive at

$$\langle \Psi, f_j \rangle = \left\langle \frac{x(t) - s}{\|x(t) - s\|}, f_j \right\rangle \le \frac{-\gamma\omega}{48L_V} \Big/ \frac{5\gamma}{4} = -\Delta.$$

Together with (2.5), this gives the inequality asserted by the lemma. Since this inequality evidently implies

$$d(x(t), S(b - i\beta)) < \gamma,$$

it also follows that $x(t) \in S(b) + \gamma B$. □

LEMMA 2.8. *If $x_j \in \Omega_i$ where $0 \le i \le N$, then x_{j+1} lies in $\Omega_{i'}$ for some $i' \ge i$.*

Proof. Since $x_j \in \Omega_i$, we have $d(x_j, S(b - i\beta)) < \gamma$, and (by Lemma 2.7) $d(x_{j+1}, S(b - i\beta)) < \gamma$. Now let $1 \le k < i$. Since $S(b - i\beta) \subset S(b - (k+1)\beta)$, we deduce $d(x_{j+1}, S(b - (k+1)\beta)) < \gamma$. But then $x_{j+1} \notin \Omega_k$ by definition of Ω_k. Since $x_{j+1} \in S(b) + \gamma B$ by Lemma 2.7, we must have $x_{j+1} \in \Omega_{i'}$ for some $i' \ge i$, in view of Lemma 2.5(b). □

LEMMA 2.9. *If $x(\tau) \in \Omega_{N+1}$ for some $\tau \in \pi$, then $x(t) \in S(a) + \gamma B$ $\forall t \ge \tau$.*

Proof. We know that $x(\tau)$ lies in the interior of $S(a) + \gamma B$ by Lemma 2.5(f). For $t > \tau$, as long as $d(x(t), S(b - N\beta))$ does not attain or exceed $\gamma/2$, then $x(t)$ remains in $S(a) + \gamma B$. Thus $\|\dot{x}(t)\|$ remains bounded by m and no blow-up occurs (i.e., $x(t)$ is well defined).

It suffices therefore to prove that the continuous function

$$g(t) := d(x(t), S(b - N\beta))$$

does not become greater than or equal to $\gamma/2$ for some $t_0 > \tau$. We have $g(\tau) < \gamma/4$.

Since we have chosen δ to satisfy $\delta m < \gamma/4$, at the next node τ_1 following τ we have $g(\tau_1) < \gamma/2$, and two cases arise. The first is when $\gamma/4 \le g(\tau_1) < \gamma/2$; in that case, $x(\tau_1)$ belongs to Ω_N, and Lemma 2.7 shows that at the next node $\tau_2, g(\tau_2)$ will have decreased relative to $g(\tau_1)$. The other case is when $g(\tau_1) < \gamma/4$; but then we are in the same situation as we were with τ. The conclusion is that $g(t)$ never exceeds $\gamma/2$, as required. □

One last lemma and the proof is complete. We set

$$T := \left(1 + \frac{b - a}{\beta}\right)\left(1 + \frac{45\gamma L_V}{\omega}\right).$$

LEMMA 2.10.

$$x(t) \in S(a) + \gamma B \qquad \forall t \ge T.$$

Proof. In view of Lemma 2.9 and Lemma 2.5(f), it suffices to prove that there is a node $\tau \in \pi$ with $\tau \le T$ for which $x(\tau) \in \Omega_{N+1}$. Note that $x(0)$ belongs to some $\Omega_i (0 \le i \le N + 1)$ by Lemma 2.5(b); if $i = N + 1$ we are done, so assume $i \le N$. Since $\delta < 1$ by (2.3), there is a node τ_1 lying in the open interval $(\sigma, \sigma + 1)$, where

$\sigma := (3\gamma)/(4\Delta)$. By Lemma 2.8, $x(\tau_1)$ belongs either to Ω_i or to $\Omega_{i'}$ for some $i' > i$. In the former case, it follows that $x(t)$ lies in Ω_i for every node $t \in \pi$ lying between 0 and τ_1, and the inequality of Lemma 2.7 applies to give

$$\begin{aligned} d(x(\tau_1), S(b - i\beta)) &\leq d(x(0), S(b - i\beta)) - \Delta\tau_1 \\ &< \gamma - \frac{3\gamma}{4} = \frac{\gamma}{4}. \end{aligned}$$

However, the left side is no less than $\gamma/4$ by Lemma 2.5(e). This contradiction shows that, in fact, $x(\tau_1)$ must belong to some $\Omega_{i'}$ for an index $i' > i$. If $i' = N + 1$, we are done; otherwise, the same argument, beginning now at $(\tau_1, x(\tau_1))$, yields the existence of a node $\tau_2 \in \pi$ with $\tau_2 \leq 2\sigma + 2$ such that $x(\tau_2)$ belongs to $\Omega_{i''}$, where $i'' > i'$. Continuing in this manner, we find that (since there are at most $N + 1$ steps as above prior to landing in Ω_{N+1}), there is a node $\tau \in \pi$ with $\tau \leq (N + 1)(\sigma + 1)$ such that $x(\tau) \in \Omega_{N+1}$. But $N < (b - a)/\beta$ implies that T as defined above is greater than $(N + 1)(\sigma + 1)$. □

Remark 2.4. It is a consequence of the construction that for suitably small $\delta > 0$, the set $A := S(a + \delta)$ is attained in a time that approaches 0 as $x(0)$ approaches A. This is the key property that is needed in the converse Lyapunov theorem that we now proceed to develop.

3. Construction of a Lyapunov function. We show in this section that under reasonable assumptions, there always exist Lyapunov functions having the properties required for the feedback construction of the preceding section and giving rise to practical feedback stabilization of an arbitrarily prescribed range. While the result below appears to be new and the approach to proving it has some novel features, there is a familiar heuristic at work: the Lyapunov function is constructed as the value function associated with a parametrized family of optimal control problems.

The function $f(x, u)$ describing the dynamics is supposed in this section to satisfy much the same regularity conditions as before. Specifically, we require that for any bounded subset S of $\mathbb{R}^n$, there exist constants $m = m(S)$ and $L = L(S)$ such that

$$\|f(x, u)\| \leq m \quad \forall x \in S, \qquad \forall u \in \mathcal{U},$$

$$\|f(x, u) - f(y, u)\| \leq L\|x - y\| \quad \forall x, y \in S, \qquad \forall u \in \mathcal{U}.$$

(As before, $\mathcal{U}$ is just an abstract set, and no hypotheses are made concerning the nature of the dependence of f on u.)

In addition, we require "nice" controllability to a given compact set A via relaxed trajectories. Let us now proceed to make this precise. We define a multifunction Γ on $\mathbb{R}^n$ by

$$\Gamma(x) := \text{cl co}\{f(x, u) : u \in \mathcal{U}\}.$$

By "trajectory" (or "Γ-trajectory") we mean an absolutely continuous function $x(\cdot)$ on an interval $[0, T]$ such that

$$\dot{x}(t) \in \Gamma(x(t)) \quad \text{a.e.} \quad t \in [0, T].$$

Given $\alpha \in \mathbb{R}^n$, we define $T_A(\alpha)$ as the least time required for a trajectory to go from α to the set A:

$$T_A(\alpha) := \inf\{T \geq 0 : x(\cdot) \text{ is a trajectory on } [0, T], x(0) = \alpha, x(T) \in A\}.$$

The controllability hypothesis that we make is that every α admits a trajectory steering it to A in finite time, a time which goes to 0 as α approaches A.
Equivalently,

$$\text{(CH)} \qquad T_A(\alpha) < \infty \quad \forall \alpha \in \mathbb{R}^n, \qquad \text{and} \lim_{d(\alpha,A)\downarrow 0} T_A(\alpha) = 0.$$

We remark that proximal criteria exist ensuring that A satisfies (CH); see [9, 4.6.7]. Now fix a point $a_0 \in A$. Below, $\operatorname{diam}(A)$ refers to the usual diameter of A as a subset of $\mathbb{R}^n$.

THEOREM 3.1. *For any $r > 0$ and $R > \operatorname{diam}(A)+r$, there exist numbers a, b, γ, η, and a function $V : \mathbb{R}^n \longrightarrow \mathbb{R}$ such that*

$$S(a) + \gamma B \subset A + rB \subset \bar{B}(a_0, R) \subset S(b) + \gamma B$$

and such that all the hypotheses of Theorem 2.1 *are satisfied and, in addition, $S(b)$ is compact. The feedback defined in Theorem* 2.1 *stabilizes $S(b) + \gamma B$ to $A + rB$.*

The following addresses the issue of practical semiglobal stabilization of globally asymptotically controllable systems.

COROLLARY 3.2. *Let the system be globally asymptotically controllable to the origin. Then, for any $0 < r < R$, there exists a feedback of the type constructed in Theorem* 2.1 *which is defined on a neighborhood of $\bar{B}(0, R)$, and which stabilizes every initial point in $B(0, R)$ to the ball $B(0, r)$.*

Proof. It is known that a continuous global Lyapunov function exists for the problem of stabilization to the origin [26], and every level set $S(a)$ for $a > 0$ of that function satisfies (CH) [9, 4.6.7]. It suffices now to take such a level set A contained in the ball $B(r/2, 0)$, and to apply Theorem 3.1 with $r := r/2$ and $a_0 := 0$. □

Proof of Theorem 3.1. We begin the proof by defining another multifunction $\tilde{\Gamma}$ (more useful than Γ for being uniformly bounded):

$$\tilde{\Gamma}(x) := \text{cl co}\left\{\frac{v}{1 + \|v\|} : v \in \Gamma(x)\right\}.$$

We set

$$\tilde{T_A}(\alpha) := \inf\{T \geq 0 : x(\cdot) \text{ is a } \tilde{\Gamma}\text{-trajectory on } [0, T], x(0) = \alpha, x(T) \in A\}.$$

Evidently (or by convention) we have $\tilde{T_A} = 0$ on A.

LEMMA 3.3.

(a) $\tilde{\Gamma}$ *is locally Lipschitz and has nonempty convex compact values in $\bar{B}(0, 1)$.*

(b) $\tilde{T_A}(\alpha)$ *is finite* $\forall \alpha \in \mathbb{R}^n$.

(c) $\lim_{d(\alpha,A)\downarrow 0} \tilde{T_A}(\alpha) = 0$.

(d) *There exists a positive number ϵ such that whenever $\alpha \in A+\epsilon B$, and whenever the $\tilde{\Gamma}$-trajectory $x(\cdot)$ has $x(0) = \alpha$ and $x(T) \in A$ for some $T \leq \tilde{T_A}(\alpha) + \epsilon$, then we have $\|x - a_0\|_\infty \leq \operatorname{diam}(A) + 1$. We can suppose $\epsilon < 1, \epsilon < r$, and*

$$\text{(3.1)} \qquad \sup\{\tilde{T_A}(\alpha) : \alpha \in A + \epsilon B\} < \frac{r^2}{4(1 + \operatorname{diam}(A))}.$$

Proof. We omit the routine proof of (a). For (b), let $\alpha \in \mathbb{R}^n$ be given. By assumption, there is a Γ-trajectory x on an interval $[0, T]$ such that $x(0) = \alpha, x(T) \in A$. We set

$$\tilde{T} := \int_0^T (1 + \|\dot{x}(t)\|)dt$$

and we define a function $\tilde{x}$ on $[0, \tilde{T}]$ by

$$\tilde{x}(\tau) := x(t),$$

where $t = t(\tau)$ is determined in $[0, T]$ by

$$\tau = \int_0^t (1 + \|\dot{x}(r)\|)dr.$$

(This change of variables or time scale is known as the Erdmann transform.) Then

$$\frac{d\tilde{x}}{d\tau} = \frac{\dot{x}(t)}{1 + \|\dot{x}(t)\|} \in \tilde{\Gamma}(\tilde{x}(\tau)) \quad \text{a.e.}$$

so that $\tilde{x}$ is a $\tilde{\Gamma}$-trajectory. Hence $\tilde{T_A}(\alpha) \leq \tilde{T} < \infty$.

We turn now to (c). Let α_i be a sequence for which $d(\alpha_i, A)$ decreases to 0. Then $T_A(\alpha_i) \to 0$ by assumption. Let m be such that $\|f(x,u)\| \leq m$ for $(x,u) \in (A + \bar{B}) \times \mathcal{U}$. Then, as soon as $T_A(\alpha_i)$ is strictly less than $1/m$, there is a Γ-trajectory x_i on an interval $[0, T_i]$ such that

$$x_i(0) = \alpha_i, \qquad x_i(T_i) \in A, \qquad T_i < \frac{1}{m}, \qquad T_i < T_A(\alpha_i) + \frac{1}{i}.$$

It follows that $x_i(t) \in A + \bar{B}$ for $t \in [0, T_i]$. Now let $\tilde{x_i}$ be the Erdmann transform of x_i as given above. Then

$$\tilde{T_A}(\alpha_i) \leq \tilde{T_i} = \int_0^{T_i} (1 + \|\dot{x_i}(t)\|)dt \leq (1+m)T_i < (1+m)\left(T_A(\alpha_i) + \frac{1}{i}\right).$$

It follows that $\tilde{T_A}(\alpha_i) \to 0$, as required.

We now examine (d). If the assertion is false, there exists a sequence α_i with $d(\alpha_i, A) \downarrow 0$ and corresponding $\tilde{\Gamma}$-trajectories x_i with $x_i(0) = \alpha_i, x_i(T_i) \in A$ such that

$$T_i \leq \tilde{T_A}(\alpha_i) + \frac{1}{i}, \qquad \|x_i - a_0\|_\infty > \operatorname{diam}(A) + 1.$$

Since $\tilde{T_A}(\alpha_i) \to 0$ by (c), we have $T_i \to 0$. On the other hand, there is a subinterval of $[0, T_i]$ in which $\|x_i - a_0\|$ goes from being $\operatorname{diam}(A) + 1$ to at most $\operatorname{diam}(A)$, and since $\|\dot{x_i}(t)\| \leq 1$ the length of that subinterval (and hence, T_i) is at least 1. This contradiction establishes the first part of (d); the rest follows immediately by shrinking ϵ as required, in light of (c). □

Defining a value function. We proceed now to define a new multifunction $F(x)$ whose effect is to enlarge the set $\tilde{\Gamma}(x)$ for $d(x, A) < \epsilon$. We set

$$F(x) := \begin{cases} \tilde{\Gamma}(x) & \text{for } d(x,A) \geq \epsilon, \\ \tilde{\Gamma}(x) + 2[\frac{\epsilon - d(x,A)}{\epsilon}]\bar{B} & \text{for } d(x,A) \leq \epsilon. \end{cases}$$

Having done this, we define a value function $V(\cdot)$ on $\mathbb{R}^n$ in terms of the trajectories of F as follows:

$$V(\alpha) := \inf\left\{\int_0^T \|x(t) - a_0\|dt : T \geq 0, x(0) = \alpha, \dot{x} \in F(x) \text{ a.e.}, x(T) \in A\right\}.$$

We stress that T is a choice variable here, in this free time problem.

Lemma 3.4.

(a) *F is compact and convex-valued, uniformly bounded, and locally Lipschitz.*

(b) *$V(\cdot)$ is nonnegative, finite-valued, and lower semicontinuous, and the infimum defining $V(\alpha)$ is attained for every α.*

(c) *$V(\alpha) = 0$ iff $\alpha \in A$, and $\lim_{d(\alpha,A)\downarrow 0} V(\alpha) = 0$.*

(d) *The sublevel sets $S(b) := \{\alpha : V(\alpha) \le b\}$ of V are compact.*

Proof. The assertions of (a) are immediate. Since $F(x)$ is uniformly bounded, the attainment and the lower semicontinuity asserted in (b) follow from standard "compactness of trajectories" arguments; see [9, Chapter 4] for details. The first assertion of (c) is clear, and the other one stems from Lemma 3.3 as follows.

Let $\alpha \in A + \epsilon B$, and let the $\tilde{\Gamma}$-trajectory x satisfy $x(0) = \alpha$, $x(T) \in A$, and $T \le \tilde{T_A}(\alpha) + \delta$, for some $\delta \in (0, \epsilon)$. Then $\|x - a_0\|_\infty \le \mathrm{diam}(A) + 1$ (by choice of ϵ), and we deduce

$$V(\alpha) \le \int_0^T \|x(t) - a_0\| dt \le (\tilde{T_A}(\alpha) + \delta)(\mathrm{diam}(A) + 1).$$

Since $\tilde{T_A}(\alpha) \downarrow 0$ as $d(\alpha, A) \downarrow 0$, (c) follows.

Finally we turn to (d). If $\|\alpha - a_0\| > \mathrm{diam}(A) + \epsilon$, then the time required for a trajectory x to go from $x = \alpha$ to the boundary of $A + \epsilon B$ is at least $\|\alpha - a_0\| - \mathrm{diam}(A) - \epsilon$. But then $V(\alpha) \ge \epsilon(\|\alpha - a_0\| - \mathrm{diam}(A) - \epsilon)$. This implies assertion (d). □

The next step invokes Hamiltonian conditions for optimal control and uses the lower Hamiltonian h associated with F:

$$h(x, p) := \min\{\langle p, v\rangle : v \in F(x)\}.$$

Lemma 3.5. *Let $\zeta \in \partial_P V(\alpha)$, where α does not lie in A. Let x be a trajectory solving the problem that defines $V(\alpha)$ with associated time T. Then there exists an absolutely continuous function p on $[0, T]$ such that*

$$\left(-\dot{p} - \frac{x - a_0}{\|x - a_0\|}, \dot{x}\right) \in \partial_C h(x, p) \quad \text{a.e.} \quad t \in [0, T], \tag{3.2}$$

$$p(0) = \zeta, \tag{3.3}$$

$$h(x(t), p(t)) + \|x(t) - a_0\| = 0 \qquad \forall t \in [0, T]. \tag{3.4}$$

Proof. By definition of $\partial_P V(\alpha)$, we have for some $\sigma \ge 0$ and for all α' near α

$$V(\alpha') + \sigma\|\alpha' - \alpha\|^2 - \langle\zeta, \alpha'\rangle \ge -\langle\zeta, \alpha\rangle.$$

Let x' be a trajectory near x (in the L^∞ norm), put $\alpha' = x'(0)$ and $\alpha = x(0)$, and rearrange to derive that $x'(\cdot) = x(\cdot)$ solves locally the problem of minimizing

$$\int_0^{T'} \|x'(t) - a_0\| dt - \langle\zeta, x'(0)\rangle + \sigma\|x'(0) - x(0)\|^2$$

over the trajectories x' for F satisfying $x'(T') \in A$. (Here T' and $x'(0)$ are free.) We apply the corollary of Theorem 3.6.1 of [5] (with time reversed) to deduce the existence of an absolutely continuous function q on $[0, T]$ satisfying

$$(-\dot{q}, \dot{x}) \in \partial_C[H(x, q) - \|x - a_0\|](x, q) \quad \text{a.e.} \quad t \in [0, T], \tag{3.5}$$

$$q(0) = -\zeta, \tag{3.6}$$

$$H(x(t), q(t)) = \|x(t) - a_0\|, \quad t \in [0, T], \tag{3.7}$$

where $H(x,p)$ is the function $-h(x,-p)$ and ∂_C denotes the generalized gradient. The Hamiltonian inclusion above implies

$$\left(\dot{q} - \frac{x-a_0}{\|x-a_0\|}, \dot{x}\right) \in \partial_C h(x,-q) \quad \text{a.e.,} \qquad t \in [0,T].$$

Now putting $p := -q$ gives to these conclusions the form asserted in the statement of the lemma. □

LEMMA 3.6. *For any constant $c > 0$, there is a constant M_c with the following property. If $\alpha \in S(c)$ and if the trajectory x on $[0,T]$ attains the infimum defining $V(\alpha)$, then $\|x-a_0\|_\infty \le M_c, T \le M_c$.*

Proof. If $\|x-a_0\|_\infty > c+\mathrm{diam}(A)+1$, then the time required for $\|x-a_0\|$ to attain the value $\mathrm{diam}(A)+1$ exceeds c (since $\|\dot{x}\| \le 1$). But then $V(\alpha) \ge (1+\mathrm{diam}(A))c > c$. This shows that $\|x-a_0\|_\infty$ is bounded by $c+\mathrm{diam}(A)+1$. By Lemma 3.3(c), $\lim_{d(\alpha,A)\downarrow 0} \tilde{T_A}(\alpha) = 0$. So there exists $\rho > 0$ such that

$$\tilde{T_A}(x) \le 1 \quad \forall x \in A + \rho\bar{B}.$$

Now take α outside $A+\rho\bar{B}$, and let τ_ρ denote the first time t that $x(t)$ attains $A+\rho\bar{B}$. Then $V(\alpha) \ge \rho\tau_\rho$, whence $\tau_\rho \le c/\rho$ for $\alpha \in S(c)$.

We deduce that

$$\begin{aligned} \tilde{T_A}(\alpha) &\le \tau_\rho + 1 \quad \text{(by choice of } \rho) \\ &\le \frac{c}{\rho} + 1. \end{aligned}$$

If $\alpha \in A+\rho\bar{B}$, the same bound evidently holds. It suffices now to set

$$M_c := \max\left\{\frac{c}{\rho}+1, c+\mathrm{diam}(A)+1\right\}. \qquad \square$$

LEMMA 3.7. *V is locally Lipschitz on $\mathbb{R}^n$.*

Proof. We prove first that V is locally Lipschitz on the open set $\{V > 0\} = \mathrm{comp}(A)$. Let α_0 belong to this set; take any $\delta > 0$ such that $\delta < d(\alpha_0, A)$, and any element $\zeta \in \partial_P V(\alpha)$, where

$$\|\alpha-\alpha_0\| < \delta, \qquad V(\alpha) \le V(\alpha_0)+\delta =: c. \tag{3.8}$$

The conclusions of Lemma 3.6 are available for any trajectory solving the $V(\alpha)$ problem. If K is a Lipschitz constant for F on the ball $B(0, M_c+1+\|a_0\|)$ (where M_c comes from Lemma 3.6), then the Hamiltonian inclusion (3.2) implies

$$\|\dot{p}\| \le K\|p\| + 1. \tag{3.9}$$

The condition (3.4) at $t = T$ gives $\|p(T)\| \le \mathrm{diam}(A)$ since $x(T) \in A$ and since $F(x(T)) = \tilde{\Gamma}(x(T)) + 2\bar{B} \supset \bar{B}$. This, together with (3.9) and Gronwall's Lemma, leads to

$$\begin{aligned} \|\zeta\| = \|p(0)\| &\le e^{KT}\|p(T)\| + \int_0^T e^{K(T-s)}ds \\ &\le \mathrm{diam}(A)e^{KT} + \frac{e^{KT}-1}{K} \\ &\le \mathrm{diam}(A)e^{KM_c} + \frac{e^{KM_c}-1}{K}, \end{aligned}$$

since $T \leq M_c$ by Lemma 3.6. This establishes a uniform bound on elements of $\partial_P V(\alpha)$ whenever α satisfies (3.8), which proves that V is Lipschitz on a neighborhood of α_0 [9, 1.11.11]. Thus V is locally Lipschitz on the set where it is strictly positive.

There is a neighborhood N of A on which V is bounded above, in view of Lemma 3.4(c). The argument above therefore yields a bound L on elements of $\partial_P V(\alpha)$ for all $\alpha \in N \setminus A$, so that V is uniformly Lipschitz of rank L on $\alpha \in N \setminus A$ by [9, Theorem 1.7.3]. Of course, $V = 0$ on A and is continuous at each point of A in view again of Lemma 3.4(c). That V is Lipschitz on N, and hence locally Lipschitz on $\mathbb{R}^n$, now follows. □

Lemma 3.8.

$$\sup\{V(\alpha) : \alpha \in A + \epsilon B\} < \inf\{V(\alpha) : d(\alpha, A) \geq r\}$$

Proof. Let $d(\alpha, A) < \epsilon$, fix $\delta \in (0, \epsilon)$, and let the trajectory x on $[0, T]$ satisfy $x(0) = \alpha, x(T) \in A, T < \tilde{T_A}(\alpha) + \delta$. Then by Lemma 3.3 we have $\|x - a_0\|_\infty \leq \operatorname{diam}(A) + 1$ and so

$$V(\alpha) \leq \int_0^T \|x(t) - a_0\| dt \leq (\tilde{T_A}(\alpha) + \delta)(\operatorname{diam}(A) + 1).$$

We derive $V(\alpha) \leq (\operatorname{diam}(A) + 1)\tilde{T_A}(\alpha)$, and (from (3.1))

$$V(\alpha) \leq \sup\{\tilde{T_A}(\alpha) : d(\alpha, A) \leq \epsilon\}(\operatorname{diam}(A) + 1) < \frac{r^2}{4}.$$

Now let $d(\alpha, A) \geq r$, and let x solve the problem defining $V(\alpha)$. There is an interval of length at least $r/2$ during which $\|x(t) - a_0\| \geq r/2$ (since $\|\dot{x}\| \leq 1$), whence

$$V(\alpha) > \frac{r^2}{4}.$$

The result follows. □

Lemma 3.9. *There exist positive numbers a, b, η with $a < b$ such that*

$$S(a) + \eta B \subset A + rB \subset \bar{B}(a_0, R) \subset S(b)$$

and

$$S(a, b) + \eta B \subset \{\alpha : d(\alpha, A) > \epsilon\}.$$

Proof. Pick a number $a > 0$ lying between the two quantities in the statement of Lemma 3.8. Then evidently the compact set $S(a)$ satisfies

$$S(a) \subset \operatorname{comp}\{d(\alpha, A) \geq r\} = A + rB,$$

whence $S(a) + \eta B \subset A + rB$ for $\eta > 0$ suitably small. It also follows that (for any $b > a$) the compact set $S(a, b)$ is contained in the open set $\{\alpha : d(\alpha, A) > \epsilon\}$. Any b suitably large will satisfy $\bar{B}(a_0, R) \subset S(b)$, since V is bounded on bounded sets. Finally, by shrinking η further if necessary, we will have the final conclusion of the lemma as well. □

Lemma 3.10. *The infinitesimal decrease condition* (H5) *of section* 1 *holds on $S(a, b) + \eta B$, with $\omega := \epsilon$.*

Proof. As pointed out in Remark 2.3, it suffices to show that for any $\alpha \in S(a,b) + \eta B$, for any $\zeta \in \partial_P V(\alpha)$, one has

$$\inf\{\langle \zeta, f(\alpha,u)\rangle : u \in \mathcal{U}\} \leq -\epsilon. \tag{3.10}$$

Let x be a trajectory solving the problem defining $V(\alpha)$. Then, by Lemma 3.5, we have (at $t = 0$)

$$h(\alpha,\zeta) + \|\alpha - a_0\| = 0.$$

Since $d(\alpha, A) > \epsilon$ by Lemma 3.9, we have $F(\alpha) = \tilde{\Gamma}(\alpha)$, so that the preceding equality yields, for any $\delta > 0$, the existence of some element $v \in \Gamma(\alpha)$ such that

$$\left\langle \zeta, \frac{v}{1 + \|v\|} \right\rangle \leq -\|\alpha - a_0\| + \delta < -\epsilon + \delta.$$

For δ small enough, the right side is negative, whence

$$\langle \zeta, v\rangle < -\epsilon + \delta.$$

Given that $\Gamma(\alpha) := \text{cl co} f(\alpha,\mathcal{U})$, this yields the existence of $u \in \mathcal{U}$ for which

$$\langle \zeta, f(\alpha,u)\rangle < -\epsilon + 2\delta.$$

Since δ is arbitrarily small, (3.10) ensues. □

Since $S(b)$ is compact, f is Lipschitz in x and bounded on $S(b) + \eta B$, in accord with hypotheses (H3) and (H4) of section 1. When the level sets are compact and V is continuous, (H2) always holds. The verification of this fact is the last property to confirm.

LEMMA 3.11. *Hypothesis* (H2) *holds.*

Proof. If (H2) fails, then there exist sequences $\alpha_i \in \mathbb{R}^n$, $\epsilon_i \downarrow 0$, and $u_i \in B(0,1)$ such that

$$V(\alpha_i) \leq a + \epsilon_i \text{ and } V(\alpha_i + \epsilon_i u_i) > b.$$

Since $S(b)$ is compact, we can suppose by passing to a subsequence that $\alpha_i \to \alpha_0$. Then, since V is continuous, we have $V(\alpha_0) \geq b > a \geq V(\alpha_0)$, a contradiction. □

The setting of Theorem 2.1 is established, and Theorem 3.1 is proved.

4. Robustness.

We prove in this section that the feedback constructed in section 1 is robust with respect to small measurement error and persistent external disturbance, in a precise sense that requires two stipulations. The first is that the measurement error must not exceed in order of magnitude the step size of the underlying discretization, a condition which appears to be rather natural. The second requirement is perhaps more surprising and surfaces from the nature of the feedback construction. It dictates that each step be "big enough" (while continuing to be "small enough") so as to counteract the measurement error by means of the attractive effect inherent in the construction. Thus the partitions used to discretize the effect of the control are taken to be "reasonably uniform."

Our perturbed system is modeled by

$$\dot{x} = f(x, k(x+p)) + q,$$

where the external disturbance $q : [0,\infty) \longrightarrow \mathbb{R}^n$ is a bounded measurable function:

$$\|q(t)\| \leq E_q, \qquad t \geq 0 \quad \text{a.e.}$$

Given a partition $\pi = \{t_i\}_{i\geq 0}$ of $[0,\infty)$ and the initial condition x_0, the resulting π-trajectory of our perturbed system is defined by successively solving the differential equation

$$\dot{x}(t) = f(x(t), k(x(t_i) + p_i)) + q(t), \qquad t \in [t_i, t_{i+1}],$$

with $x(0) = x_0$. The continuous function $x(t)$ is the real state of the system, while the sequence $\{x(t_i) + p_i\}$ corresponds to the inexact measurements used to select control values.

THEOREM 4.1. *The feedback $k : S(b) + \gamma B \to \mathcal{U}$ constructed in Theorem* 2.1 *is robust in the sense that there exist positive numbers δ_0, T, and E_q such that, for every $\delta \in (0, \delta_0)$ there exists $E_p(\delta) > 0$ having the following property: for any partition $\pi = \{t_i\}_{i\geq 0}$ having*

$$\frac{\delta}{2} \leq t_{i+1} - t_i \leq \delta, \qquad i \geq 0,$$

where $0 < \delta < \delta_0$, for any set of measurement errors $\{p_i\}_{i\geq 0}$ having

$$\|p_i\| \leq E_p(\delta), \qquad i \geq 0,$$

for any initial condition x_0 such that $x_0 + p_0 \in S(b) + \gamma B$, for any disturbance q having $\|q\|_\infty \leq E_q$, the resulting π-trajectory x satisfies

$$x(t_i) + p_i \in S(b) + \gamma B \quad \forall i \geq 0,$$

$$x(t) \in S(b) + 2\gamma B \quad \forall t \geq 0,$$

$$x(t) \in S(a) + \gamma B \quad \forall t \geq T.$$

Remarks 4.1.

(a) Note that unlike T and E_q, the maximum admissible measurement error E_p depends on δ. Also note that (in contrast to Theorem 2.1) $x(t)$ may not lie in $S(b) + \gamma B \ \forall\ t$, although for large t it must do so. We prove, however, that the "observed values" of the state, namely the values $x(t_i) + p_i (i \geq 0)$, all fall in $S(b) + \gamma B$, the domain of definition of k.

(b) Certain other kinds of error, for example a disturbance $d(\cdot)$ entering into the dynamics in the form $\dot{x} = f(x, k(x) + d)$, can be reduced to that of an external disturbance by positing suitable continuity of f in the control variable.

(c) The maximum admissible disturbance measure E_q will be seen to be proportional to ω/L_V. This has a natural physical meaning, as can easily be seen in the case of smooth V and a continuous feedback $k(x)$ such that

$$\langle \nabla V(x), f(x, k(x)) \rangle \leq -W(x).$$

Then we see that the perturbed system

$$\dot{x} = f(x, k(x)) + q$$

is stabilized by k if $\|q\|_\infty < W(x)/\|\nabla V(x)\|$ for every x, a bound akin to that involving ω/L_V.

Proof of Theorem 4.1. We adapt the proof of Theorem 2.1, whose first five lemmas hold with no change whatever, as does the definition of $k(\cdot)$. Recall that γ, β, and N were introduced earlier; see (2.1) and (2.2). We now define our upper bound for δ:

$$\delta_0 := \min\left\{\frac{6\gamma L_V}{\omega}, 1, \frac{\gamma\omega}{24L_V(m+\frac{\omega}{6L_V}+1)^2}, \frac{\gamma}{20m}\right\}, \tag{4.1}$$

$$E_q := \frac{\omega}{6L_V}, \tag{4.2}$$

$$T := \left(1+\frac{b-a}{\beta}\right)\left(1+\frac{81\gamma L_V}{\omega}\right) \tag{4.3}$$

(this T differs slightly from the one in Theorem 2.1), and we let $\pi = \{t_i\}_{i\geq 0}$ be a partition as described in the statement of Theorem 4.1, with corresponding measurement errors $\{p_i\}_{i\geq 0}$ having $\|p_i\| \leq E_p$ for some $E_p > 0$ satisfying

$$E_p < \min\left\{\frac{3\gamma}{80}, \delta, \frac{\delta\omega}{432L_V}\right\}. \tag{4.4}$$

We also admit any disturbance $q(\cdot)$ for which $\|q\|_\infty \leq E_q$, and we take x_0 such that $x_0 + p_0 \in S(b) + \gamma B$. We shall show that the corresponding π-trajectory has the required properties. We introduce the notation

$$x_i := x(t_i), y_i := x_i + p_i$$

for the actual and the measured state values at time t_i and proceed to develop modified versions of the four last lemmas figuring in the proof of Theorem 2.1. We set

$$\tilde{\Delta} := \frac{\omega}{108L_V}.$$

LEMMA 4.2. *For some $t_j \in \pi$, suppose that $y_j \in \Omega_i$, $i \in \{0, 1, \ldots, N\}$. Then*

$$x(t) \in S(b) + 2\gamma B \subset S(b) + \eta B, \qquad t_j \leq t \leq t_{j+1},$$

$$y_j \in S(b) + \gamma B, \qquad y_{j+1} \in S(b) + \gamma B,$$

$$d(y_{j+1}, S(b-i\beta)) \leq d(y_j, S(b-i\beta)) - \tilde{\Delta}(t_{j+1}-t_j).$$

Proof. Note that $y_j \in S(b) + \gamma B$ by Lemma 2.5; it will follow from the last conclusion of the current lemma that $y_{j+1} \in S(b) + \gamma B$. Also, $\|x_j - y_j\| = \|p_j\| \leq E_p$, together with $\|x(t) - x_j\| \leq \delta m$, yield

$$x(t) \in S(a, b) + \gamma B + (E_p + \delta m)B \subset S(b) + 2\gamma B,$$

since $E_p + \delta m < \gamma$ in view of (4.1) and (4.4). Since $2\gamma < \eta$ by (2.1), this gives $x(t) \in S(b) + \eta B$. By Lemma 2.6 we have

$$\langle y_j - s, f(s, k(y_j))\rangle \leq -\frac{\omega}{2L_V}\|y_j - s\|, \tag{4.5}$$

where $s \in \operatorname{proj}(y_j, S(b-i\beta))$. Fix $t \in (t_j, t_{j+1})$ and set

$$\Psi := \frac{x(t) - s}{\|x(t) - s\|}.$$

Note that $x(t) \neq s$ since $\|y_j - s\| \geq \gamma/4$ by Lemma 2.5(e), while

$$\|y_j - x(t)\| \leq \|y_j - x_j\| + \|x_j - x(t)\| < E_p + \delta m < \frac{\gamma}{4},$$

as already noted. We observe the relations

$$\begin{aligned} d(x(t), S(b - i\beta)) &\leq \|x(t) - s\| = \langle \Psi, x(t) - s \rangle, \\ d(y_j, S(b - i\beta)) &= \|y_j - s\| \geq \langle \Psi, y_j - s \rangle, \end{aligned}$$

whence

$$\begin{aligned} d(x(t), S(b - i\beta)) - d(y_j, S(b - i\beta)) &\leq \langle \Psi, x(t) - y_j \rangle \\ &= \langle \Psi, x_j + \tau(f_j + q_j) - y_j \rangle \\ &\leq \tau \langle \Psi, f_j + q_j \rangle + \|p_j\|, \end{aligned} \tag{4.6}$$

where we have introduced

$$\begin{aligned} \tau &:= t - t_j, \\ f_j &:= \frac{1}{\tau} \int_{t_j}^{t} f(x(r), k(y_j)) dr, \\ q_j &:= \frac{1}{\tau} \int_{t_j}^{t} q(r) dr. \end{aligned}$$

We also set $\hat{f}_j := f(s, k(y_j))$; note $\|\hat{f}_j\| \leq m$. We have

$$\begin{aligned} \|f_j - \hat{f}_j\| &\leq \frac{1}{\tau} \int_{t_j}^{t} L_f \|x(r) - s\| dr \leq \frac{L_f}{\tau} \int_{t_j}^{t} (\|x(r) - x_j\| + \|x_j - s\|) dr \\ &\leq L_f(\delta m + \|x_j - y_j\| + \|y_j - s\|) \\ &\leq L_f(\delta m + E_p + \gamma) \leq \frac{5}{4} L_f \gamma. \end{aligned} \tag{4.7}$$

We deduce

$$\begin{aligned} \langle x(t) - s, f_j + q_j \rangle &= \langle x_j + \tau(f_j + q_j) - s, f_j + q_j \rangle \\ &= \langle y_j - s - p_j, f_j + q_j \rangle + \tau \|f_j + q_j\|^2 \\ &\leq \langle y_j - s, f_j \rangle + E_p(m + E_q) + \delta(m + E_q)^2 \\ &= \langle y_j - s, \hat{f}_j \rangle + \langle y_j - s, f_j - \hat{f}_j \rangle + (m + E_q)[\delta m + \delta E_q + E_p] \\ &\leq -\frac{\omega}{2L_V} \|y_j - s\| + \frac{5}{4} \|y_j - s\| L_f \gamma + (m + E_q)[\delta m + \delta E_q + E_p] \\ &\qquad \text{(where we have used (4.5) and (4.7))} \\ &\leq d(y_j, S(b - i\beta)) \left[-\frac{\omega}{2L_V} + \frac{\omega}{6L_V} \right] + \delta[m + E_q + 1]^2 \\ &\qquad \text{(by (2.1), and since } E_p < \delta \text{ by (4.4))} \\ &\leq \frac{\gamma}{4} \left[-\frac{\omega}{3L_V} \right] + \frac{\gamma\omega}{24L_V} = -\frac{\gamma\omega}{24L_V} \end{aligned}$$

(since $d(y_j, S(b - i\beta)) \geq \gamma/4$ by Lemma 2.5, and since $\delta < \delta_0$ defined by (4.1)). Note also that

$$\begin{aligned} \|x(t) - s\| &= \|y_j - p_j + \tau(f_j + q_j) - s\| \\ &\leq \gamma + E_p + \delta(m + E_q) \leq \frac{5\gamma}{4} + \delta E_q < \frac{9\gamma}{4} \end{aligned}$$

(note $\delta E_q < \gamma$ because of $\delta < \delta_0$, in view of (4.1) and (4.2)).

It follows that

$$\langle \Psi, f_j + q_j \rangle = \left\langle \frac{x(t) - s}{\|x(t) - s\|}, f_j + q_j \right\rangle \leq -\frac{(\gamma\omega)/(24L_V)}{(9\gamma)/4} = -\frac{\omega}{54L_V}.$$

Substituting into (4.6) leads to

$$d(x(t), S(b - i\beta)) - d(y_j, S(b - i\beta)) \leq -2\tilde{\Delta}(t - t_j) + E_p.$$

We obtain from this

$$\begin{aligned} d(y_{j+1}, S(b - i\beta)) - d(y_j, S(b - i\beta)) &\leq \|y_{j+1} - x(t_{j+1})\| - 2\tilde{\Delta}(t - t_j) + E_p \\ &\leq -\tilde{\Delta}(t_{j+1} - t_j) + [2E_p - \tilde{\Delta}(t_{j+1} - t_j)] \\ &\leq -\tilde{\Delta}(t_{j+1} - t_j), \end{aligned}$$

by (4.4), and since $t_{j+1} - t_j \geq \delta/2$. □

LEMMA 4.3. *If $y_j \in \Omega_i$, where $i \leq N$, then y_{j+1} lies in Ω_k for some $k \geq i$.*

Proof. We know that $y_{j+1} \in \Omega_k$ for some k, since y_{j+1} belongs to $S(b) + \gamma B$ by Lemma 4.2. Suppose that $k < i$. We have $d(y_j, S(b - i\beta)) < \gamma$ by definition of Ω_i, and Lemma 4.2 implies $d(y_{j+1}, S(b - i\beta)) < \gamma$. But $S(b - i\beta) \subset S(b - (k+1)\beta)$, so that $d(y_{j+1}, S(b - (k+1)\beta)) < \gamma$. But then $y_{j+1} \notin \Omega_k$ by definition of Ω_k (note that $k \leq N$). This contradiction proves the lemma. □

LEMMA 4.4. *If $\tau \in \pi$ is such that $y(\tau) \in \Omega_{N+1}$, then*

$$x(t) \in S(a) + \gamma B \quad \forall t \geq \tau.$$

Proof. We first establish

$$d(y(\tau'), S(b - N\beta)) \leq \frac{2\gamma}{5} \qquad \forall \text{ nodes } \tau' \geq \tau. \tag{4.8}$$

We consider first $\tau' = \tau + 1$. We have $d(y(\tau), S(b - N\beta)) \leq \gamma/4$, whence

$$\begin{aligned} d(y(\tau+1), S(b - N\beta)) &\leq d(x(\tau+1), S(b - N\beta)) + E_p \\ &\leq d(x(\tau), S(b - N\beta)) + E_p + m\delta \\ &\leq d(y(\tau), S(b - N\beta)) + 2E_p + m\delta \\ &< \frac{\gamma}{4} + \frac{3\gamma}{20} = \frac{2\gamma}{5} \text{ (by (4.4))}. \end{aligned}$$

If $d(y(\tau+1), S(b - N\beta))$ is in fact $\leq \gamma/4$, then this same argument yields

$$d(y(\tau+2), S(b - N\beta)) \leq \frac{2\gamma}{5}.$$

If however $d(y(\tau+1), S(b - N\beta)) > \gamma/4$, then $y(\tau+1)$ lies in Ω_N by definition, and Lemma 4.2 again yields $d(y(\tau+2), S(b - N\beta)) < 2\gamma/5$. Continuing in this way, we obtain (4.8) for all nodes $\tau' \geq \tau$.

We use (4.8) to argue as follows: let $t \geq \tau$, and let $\tau' \geq \tau$ be a node adjacent to t; then

$$\begin{aligned} d(x(t), S(b - N\beta)) &\leq d(x(\tau'), S(b - N\beta)) + \delta m \\ &\leq d(y(\tau'), S(b - N\beta)) + E_p + \delta m \\ &< \frac{2\gamma}{5} + \frac{\gamma}{10} = \frac{\gamma}{2} \text{ (by (4.4))}. \end{aligned}$$

This gives $x(t) \in S(a) + \gamma B$ by Lemma 2.5(f). □

LEMMA 4.5. *Let*

$$T := \left(1 + \frac{b-a}{\beta}\right)\left(1 + \frac{81\gamma L_V}{\omega}\right).$$

Then $x(t) \in S(a) + \gamma B \quad \forall t \geq T$.

Proof. In view of Lemma 4.4, it suffices to prove that some node $\tau \in \pi$ with $\tau \leq T$ is such that $y(\tau) \in \Omega_{N+1}$. The argument is identical to that used to prove Lemma 2.9, with Δ replaced by $\tilde{\Delta}$ and applied to the y_i rather than the x_i. □

REFERENCES

[1] Z. ARTSTEIN, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[2] A. BACCIOTTI, *Local Stabilizability of Nonlinear Control Systems*, World Scientific, River Edge, NJ, 1992.

[3] A. BACCIOTTI AND L. ROSIER, *Lyapunov and Lagrange stability: Inverse theorems for discontinuous systems*, Math. Control Signals Systems, 11 (1998), pp. 101–128.

[4] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millmann, and H. J. Sussmann, eds., Birkhäuser, Basel, Boston, 1983, pp. 181–191.

[5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York; reprinted as Classics Appl. Math. 5, SIAM, Philadelphia, PA, 1990.

[6] F. H. CLARKE, YU. S. LEDYAEV, E. D. SONTAG, AND A. I. SUBBOTIN, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[7] F. H. CLARKE, YU. S. LEDYAEV, AND R. J. STERN, *Asymptotic stability and Lyapunov functions*, J. Differential Equations, 149 (1998), pp. 69–114.

[8] F. H. CLARKE, YU. S. LEDYAEV, R. J. STERN, AND P. R. WOLENSKI, *Qualitative properties of trajectories of control systems: A survey*, J. Dynam. Control Systems, 1 (1995), pp. 1–48.

[9] F. H. CLARKE, YU. S. LEDYAEV, R. J. STERN, AND P. R. WOLENSKI, *Nonsmooth Analysis and Control Theory*, Graduate Texts in Mathematics 178, Springer-Verlag, New York, 1998.

[10] F. H. CLARKE, YU. S. LEDYAEV, AND A. I. SUBBOTIN, *The synthesis of universal feedback pursuit strategies in differential games*, SIAM J. Control Optim., 35 (1997), pp. 552–561.

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

[12] J.-M. CORON AND L. ROSIER, *A relation between continuous time-varying and discontinuous feedback stabilization*, J. Math. Systems Estim. Control, 4 (1994), pp. 67–84.

[13] J.-M. CORON, L. PRALY, AND A. TEEL, *Feedback stabilization of nonlinear systems: Sufficient conditions and Lyapunov and input-output techniques*, in Trends in Control, A. Isidori, ed., Springer-Verlag, New York, 1995, pp. 293–348.

[14] A. F. FILIPPOV, *Differential Equations with Discontinuous Righthand Sides*, Kluwer Academic Publishers, Norwell, MA, 1988.

[15] R. FREEMAN AND P. V. KOKOTOVIC, *Robust Nonlinear Control Design. State-Space and Lyapunov Techniques*, Birkhäuser, Basel, Boston, 1996.

[16] H. HERMES, *Discontinuous vector fields and feedback control*, in Differential Equations and Dynamic Systems, J. P. LaSalle, ed., Academic Press, New York, 1967, pp. 155–165.

[17] H. HERMES, *Resonance, stabilizing feedback controls, and regularity of viscosity solutions of Hamilton-Jacobi-Bellmann equations*, Math. Control Signals Systems, 9 (1996), pp. 59–72.

[18] A. ISIDORI, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, New York, 1995.

[19] N. N. KRASOVSKII AND A. I. SUBBOTIN, *Game-Theoretical Control Problems*, Springer-Verlag, New York, 1988.

[20] V. LAKSHMIKANTHAM, S. LEELA, AND A. MARTYNYUK, *Practical Stability of Nonlinear Systems*, World Scientific, Singapore, 1990.

[21] YU. S. LEDYAEV AND E. D. SONTAG, *A Lyapunov characterization of robust stabilization*, Nonlinear Anal., 37 (1999), pp. 813–840.

[22] Yu. S. Ledyaev and E. D. Sontag, *A Remark on Robust Stabilization of General Asymptotically Controllable Systems*, in Proceedings of the Conference on Information Sciences and Systems (CISS 97), Johns Hopkins University, Baltimore, MD, 1997, pp. 246–251.

[23] G. Leitmann, *One approach to the control of uncertain dynamical systems*, Appl. Math. Comput., 70 (1995), pp. 261–272.

[24] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[25] E. P. Ryan, *On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback*, SIAM J. Control Optim., 32 (1994), pp. 1597–1604.

[26] E. D. Sontag, *A Lyapunov-like characterization of asymptotic controllability*, SIAM. J. Control Optim., 21 (1983), pp. 462–471.

[27] E. D. Sontag, *Mathematical Control Theory*, Texts in Appl. Math. 6, Springer-Verlag, New York, 1990.

[28] E. D. Sontag, *Control of systems without drift via generic loops*, IEEE Trans. Automat. Control, 40 (1995), pp. 1210–1219.

[29] E. D. Sontag and H. J. Sussmann, *Nonsmooth control-Lyapunov functions*, in Proceedings of the IEEE Conference on Decision and Control, New Orleans, LA, Birkhäuser, Basel, Boston, 1990, pp. 61–81.

[30] E. D. Sontag and H. J. Sussmann, *Remarks on continuous feedback*, in Proceedings of the IEEE Conference on Decision and Control, Albuquerque, NM, 1980, IEEE Publications, Piscataway, NJ, 1980, pp. 916–921.

[31] A. I. Subbotin, *Generalized Solutions of First-Order PDEs*, Birkhäuser, Basel, Boston, 1995.

[32] J. Tsinias, *A Lyapunov description of stability in control systems*, Nonlinear Anal., 13 (1989), pp. 3–74.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 49–72

ASYMPTOTIC ALMOST SURE EFFICIENCY OF AVERAGED STOCHASTIC ALGORITHMS[*]

MARIANE PELLETIER[†]

Abstract. First, we define the notion of almost sure efficiency for a decreasing stepsize stochastic algorithm, and then we show that the averaging method, which gives asymptotically efficient algorithms, also gives asymptotically almost surely efficient algorithms. Moreover, we prove that the averaged algorithm also satisfies a law of the iterated logarithm, as well as an almost sure central limit theorem.

Key words. stochastic algorithms, central limit theorem, almost sure invariance principles

AMS subject classification. 62L20, 62F12, 60F05, 60F15

PII. S0363012998308169

1. Introduction.

Many vectorial algorithms are written in the form

$$Z_{n+1} = Z_n + \gamma_n \left[F\left(Z_n, \eta_{n+1} \right) \right],$$

where the gain $(\gamma_n)_{n\geq 0}$ is a nonrandom sequence decreasing to 0 with $\sum \gamma_n = \infty$ and the observed quantity at time $n+1$ is $F(Z_n, \eta_{n+1})$, η_{n+1} being a stochastic disturbance. Such an algorithm is often studied when rewritten as an algorithm used for the search of zeros of a function h,

$$Z_{n+1} = Z_n + \gamma_n \left[h\left(Z_n \right) + e_{n+1} \right], \tag{1}$$

where (e_n) is a "small disturbance" and $h(Z_n)$ corresponds to a mean effect of $F(Z_n, \eta_{n+1})$, given the past. The classical Robbins–Monro algorithm is obtained when (η_n) is a sequence of independent identically distributed random variables and $h(z) = E[F(z, \eta)]$. Extensions to a Markovian disturbance (η_n) are developed in [1].

Throughout this paper, the algorithm (1) is considered in the following framework: (e_n) is a sequence of d-dimensional random vectors defined on a probability space $(\Omega, \mathcal{A}, P)$, adapted to a filtration $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$, and Z_0 is $\mathcal{F}_0$-measurable. The function h is defined on $\mathbb{R}^d$ and $\mathbb{R}^d$-valued, and z^* is a zero of h such that, on a neighborhood of z^*,

$$h(z) = H\left(z - z^* \right) + O\left(\|z - z^*\|^a \right),$$

where $a > 1$ and H is a stable matrix (i.e., all the real parts of the eigenvalues of H are strictly negative).

Many criteria ensure the almost sure convergence or the convergence with a strictly positive probability of (Z_n) towards z^* (see among many others [1], [9], [12], and [18]). In order to ensure their applications to various cases, our results are conditional with respect to the event $\Gamma(z^*) = \{Z_n \to z^*\}$.

[*] Received by the editors October 13, 1998; accepted for publication (in revised form) January 10, 2000; published electronically August 17, 2000.
http://www.siam.org/journals/sicon/39-1/30816.html

[†] Laboratoire de Mathematiques, Bâtiment Fermat, Université de Versailles Saint-Quentin, 45 Avenue des Etats-Unis, 78035 Versailles Cedex, France (pelletier@math.uvsq.fr).

It was proved, under some local assumptions stated in section 2, that if $P\left[\Gamma\left(z^*\right)\right]>0$, then the sequence (Z_n) satisfies a conditional central limit theorem (CLT):

$$\text{given } \Gamma\left(z^*\right), \ \Psi_n=\sqrt{\frac{1}{\gamma_n}}\left(Z_n-z^*\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0, \Sigma), \tag{2}$$

with $\overset{\mathcal{D}}{\to}$ denoting the convergence in distribution, $\mathcal{N}$ denoting the Gaussian distribution, and Σ being a positive definite matrix (see, for instance, [1], [12], [18], [19] for the case $P\left[\Gamma\left(z^*\right)\right]=1$ and [9] or [20] for the case $P\left[\Gamma\left(z^*\right)\right]>0$). (Equation (2) means that the asymptotic conditional law of Ψ_n with respect to $\Gamma\left(z^*\right)$ is $\mathcal{N}(0, \Sigma)$.)

Moreover, the sequence (Z_n) is known to fulfill the three following almost sure properties ([21], [22]), where Σ is the limit covariance matrix of (2), δ_x denotes the point mass at x, and $\Longrightarrow$ is the weak convergence. (Throughout the paper, we say that a property $\mathcal{P}$ holds almost surely (a.s.) on $\Gamma\left(z^*\right)$ if there exists a subset $N \subset \Gamma\left(z^*\right)$ such that $P(N)=0$ and $\mathcal{P}$ holds $\forall \omega \in \Gamma\left(z^*\right) \setminus N$.)

- A quadratic strong law of large numbers:

$$\text{a.s. on } \Gamma\left(z^*\right), \ \lim_{n \to \infty} \frac{1}{\sum_{k=1}^n \gamma_k} \sum_{k=1}^n \left(Z_k-z^*\right)\left(Z_k-z^*\right)^T=\Sigma. \tag{3}$$

- A law of the iterated logarithm: for any eigenvector of H^T, $w \in \mathbb{R}^d$,

$$\text{a.s. on } \Gamma\left(z^*\right), \ \limsup_{n \to \infty} \frac{1}{2\gamma_n \ln\left(\sum_{k=1}^n \gamma_k\right)} w^T\left(Z_k-z^*\right)\left(Z_k-z^*\right)^T w=w^T \Sigma w. \tag{4}$$

- An almost sure central limit theorem (a.s.CLT):

$$\text{a.s. on } \Gamma\left(z^*\right), \ \frac{1}{\sum_{k=1}^n \gamma_k} \sum_{k=1}^n \gamma_k \delta_{\sqrt{\frac{1}{\gamma_k}}\left(Z_k-z^*\right)} \Longrightarrow \mathcal{N}(0, \Sigma), \tag{5}$$

i.e., there exists a P-null set $N \subset \Gamma\left(z^*\right)$ such that $\forall \omega \in \Gamma\left(z^*\right) \setminus N$,

$$\frac{1}{\sum_{k=1}^n \gamma_k} \sum_{k=1}^n \gamma_k \delta_{\sqrt{\frac{1}{\gamma_k}}\left(Z_k(\omega)-z^*\right)} \Longrightarrow \mathcal{N}(0, \Sigma).$$

The optimal weak convergence rate of (1) given $\Gamma\left(z^*\right)$ is reached when $\gamma_n=\gamma_0/n$ with $2L\gamma_0>1$ ($-L$ denoting the greatest real part of the eigenvalues of H), since (2) is then equivalent to

$$\text{Given } \Gamma\left(z^*\right), \ \sqrt{n}\left(Z_n-z^*\right) \overset{\mathcal{D}}{\to} \mathcal{N}\left(0, \gamma_0 \Sigma\right).$$

The question arises as to what the optimal covariance matrix is. For that, let us consider the following class of algorithms:

$$Z_{n+1}=Z_n+\frac{A}{n}\left[h\left(Z_n\right)+e_{n+1}\right], \tag{6}$$

where A is an invertible $d \times d$ matrix such that $AH+I/2$ is stable. For such algorithms (see [9]), it follows from (2) that

$$\text{Given } \Gamma\left(z^*\right), \ \sqrt{n}\left(Z_n-z^*\right) \overset{\mathcal{D}}{\to} \mathcal{N}\left(0, \Sigma(A)\right),$$

where $\Sigma(A)$ is the solution of the Lyapunov equation

$$\left[AH + \frac{I}{2}\right]\Sigma(A) + \Sigma(A)\left[H^T A^T + \frac{I}{2}\right] = -ACA^T,$$

with $C = \lim_{n\to+\infty} E\left(e_{n+1}e_{n+1}^T|\mathcal{F}_n\right)$ a.s. on $\Gamma(z^*)$. The optimal choice of the matrix A in (6) is $A = -H^{-1}$, which leads to $\Sigma(A) = H^{-1}C(H^{-1})^T$, since it minimizes the covariance $\Sigma(A)$ (with respect to the order of the symmetric matrices).

When A is replaced by H^{-1}, (6) is Newton's algorithm, which thus has an asymptotically optimal behavior in distribution. Unfortunately, it is often impossible to use this algorithm, the matrix H being generally unknown.

These considerations lead us to set the following definition.

DEFINITION 1. *If $(Y_n)_{n\geq 0}$ is given by a stochastic algorithm used for the search of zeros of a function h observable only together with a disturbance (e_n), h and (e_n) satisfying the assumptions previously given and y^* being a zero of h, we say that the algorithm is* asymptotically efficient given $\{Y_n \to y^*\}$ *if*

$$\text{Given } \{Y_n \to y^*\},\ \sqrt{n}\,(Y_n - y^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, H^{-1}C(H^{-1})^T).$$

The averaging method, simultaneously introduced by Polyak [23] and Ruppert [25], is known to give asymptotically efficient algorithms (see among others Delyon–Juditsky [5], [6], Dippon–Renz [7], Kushner–Yang [13], Polyak–Juditsky [24], and Yin [31]).

The averaged algorithm is built up in the following way; each iteration requires two steps.

Step 1. Z_{n+1} is found from the algorithm (1) where the gain γ_n is "slow;" typically

$$\gamma_n = \frac{\gamma_0}{n^\alpha},\ \frac{1}{2} < \alpha < 1.$$

Step 2. We compute the empirical mean of all the previous observations,

$$\overline{Z}_{n+1} = \frac{1}{n+1}\sum_{k=1}^{n+1} Z_k.$$

Note that $\overline{Z}_{n+1}$ can be recursively written as

$$\overline{Z}_{n+1} = \overline{Z}_n + \frac{1}{n+1}\left(Z_{n+1} - \overline{Z}_n\right).$$

The aim of this paper is to prove that, conditionally on the set of consistency $\Gamma(z^*)$, the averaged algorithm is not only asymptotically efficient, but that it also satisfies the almost sure properties (3), (4), and (5) with the optimal rate and the optimal covariance matrix Σ. Moreover, the law of the iterated logarithm (4) is obtained for any vector w (and not only for eigenvectors of H^T).

Before stating our main results, let us first introduce the almost sure version of the notion of asymptotic efficiency. When $\gamma_n = \gamma_0/n$ with $2L\gamma_0 > 1$, (3) is equivalent to

$$\text{a.s. on } \Gamma(z^*),\ \lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^{n}(Z_k - z^*)(Z_k - z^*)^T = \gamma_0\Sigma, \tag{7}$$

and, for slower gains $\gamma_n = \dfrac{\gamma_0}{n^\alpha}$, $\dfrac{1}{2} < \alpha < 1$, to

$$\text{a.s. on } \Gamma\left(z^*\right), \ \lim_{n\to\infty} \frac{1-\alpha}{n^{1-\alpha}} \sum_{k=1}^{n} \left(Z_k - z^*\right)\left(Z_k - z^*\right)^T = \gamma_0 \Sigma.$$

Thus, the sum of the "squared" differences between Z_k and the estimated parameter z^* is optimal when (7) is fulfilled with $\gamma_0\Sigma = H^{-1}C(H^{-1})^T$. By analogy with the definition of the asymptotic efficiency, and taking up the terminology introduced by Touati [28] for statistical problems, we introduce the following definition.

DEFINITION 2. *If (Y_n) is given by a stochastic algorithm used for the search of zeros of a function h observable only together with a disturbance (e_n), h and (e_n) satisfying the assumptions previously given and y^* being a zero of h, we say that the algorithm is* asymptotically almost surely efficient on $\{Y_n \to y^*\}$ *if*

$$\text{a.s. on } \left\{Y_n \to y^*\right\}, \ \lim_{n\to\infty} \frac{1}{\ln n} \sum_{k=1}^{n} \left(Y_k - y^*\right)\left(Y_k - y^*\right)^T = H^{-1}C(H^{-1})^T.$$

We shall prove that the averaged algorithm is a.s. efficient on $\Gamma\left(z^*\right)$, i.e., that

$$\text{a.s. on } \Gamma(z^*), \ \lim_{n\to\infty} \frac{1}{\ln n} \sum_{k=1}^{n} (\overline{Z}_k - z^*)(\overline{Z}_k - z^*)^T = H^{-1}C(H^{-1})^T.$$

We shall also show that the averaged algorithm fulfills the following law of the iterated logarithm,

$$\text{a.s. on } \Gamma\left(z^*\right), \ \forall w \in \mathbb{R}^d,$$

$$\limsup_{n\to\infty} \frac{n}{2\ln(\ln n)} w^T \big(\overline{Z}_n - z^*\big)\big(\overline{Z}_n - z^*\big)^T w = w^T H^{-1} C \big(H^{-1}\big)^T w,$$

and the following a.s.CLT:

$$\text{a.s. on } \Gamma\left(z^*\right), \ \frac{1}{\ln n} \sum_{k=1}^{n} \frac{1}{k} \delta_{\sqrt{k}(\overline{Z}_k - z^*)} \Longrightarrow \mathcal{N}(0, H^{-1}C(H^{-1})^T).$$

In fact, we shall extend our framework and study the more general algorithm (including the Kiefer–Wolfowitz algorithm [9])

$$Z_{n+1} = Z_n + \gamma_n \left[h\left(Z_n\right) + r_{n+1}\right] + \sigma_n \varepsilon_{n+1}, \tag{8}$$

where (ε_n) is a noise (i.e., a sequence of martingale increments), (r_n) a residual term, and (σ_n) a nonrandom sequence decreasing to 0 such that $\gamma_n = O\left(\sigma_n\right)$; the algorithm (1) corresponds then to the particular case $\gamma_n = \sigma_n$ and $e_n = r_n + \varepsilon_n$.

Our results are precisely stated in section 2. In section 3, we give an application to efficient recursive estimators. Finally, section 4 is devoted to the proofs.

2. Assumptions and main results. We first precise the required assumptions on (8).

Assumption (A1) *about the function h*: There exist $a > 1$, a stable matrix H, and a neighborhood $\mathcal{U}$ of z^*, such that for any z in $\mathcal{U}$,

$$h(z) = H\left(z - z^*\right) + O\left(\|z - z^*\|^a\right).$$

Assumption (A2) *about the noise* (ε_n):

(i) There exist $M > 0$ and $b > 2$ such that a.s.

$$E\left(\varepsilon_{n+1}|\mathcal{F}_n\right) 1_{\{\|Z_n - z^*\| \le M\}} = 0 \quad \text{and} \quad \sup_{n \ge 0} E(\|\varepsilon_{n+1}\|^b|\mathcal{F}_n) 1_{\{\|Z_n - z^*\| \le M\}} < \infty.$$

(ii) There exists a nonrandom symmetric positive definite matrix C such that $\lim_{n\to\infty} C_n = C$ a.s. on $\Gamma(z^*)$, where $C_n = E\left(\varepsilon_{n+1}\varepsilon_{n+1}^T|\mathcal{F}_n\right)$.

Assumption (A3) *about the gains:* There exist two decreasing positive functions γ and σ, defined over $[0, +\infty[$ such that $\gamma_n = \gamma(n)$ and $\sigma_n = \sigma(n)$ $\forall$ integer n. We define the function v by $v(t) = \gamma(t)/\sigma^2(t)$ and assume there exist two positive real numbers α and β such that the following conditions are fulfilled.

(i) v is a differentiable increasing function, $v(\infty) = \infty$, and its differential v' varies regularly with exponent $(\beta - 1)$ (i.e., $\lim_{t\to\infty} v'(tx)/v'(t) = x^{\beta-1}$; see [11] or [26]).

(ii) γ is differentiable, varies regularly with exponent $(-\alpha)$, and $\theta = (1/\gamma)'$ is decreasing and varies regularly with exponent $(\alpha - 1)$.

(iii) One of the two following conditions (A3.a) or (A3.b) holds.
(A3.a) $\min\{\frac{1}{2}, \frac{2}{b}\} < \alpha < 1$ and $\frac{1-\alpha}{a-1} < \beta \le 1$,
(A3.b) $\frac{1}{2} < \alpha < 1$ and $\frac{1-\alpha}{a-1}(1 + \frac{2a}{b}) < \beta \le 1$.

Assumption (A4) *about the residual term* (r_n): We set

$$J(t) = \int_0^t \frac{1}{\gamma(s)v(s)} ds, \tag{9}$$

and assume $r_{n+1} = r_{n+1}^{(1)} + r_{n+1}^{(2)}$ with

(i) $r_{n+1}^{(2)} = O\left(\|Z_n - z^*\|^a\right)$ a.s.,

(ii) the weak assumption (A4.w) or the stronger one (A4.s) is fulfilled:
(A4.w) There exists $M > 0$ such that

$$\|r_{n+1}^{(1)}\| 1_{\{\|Z_n - z^*\| \le M\}} = O([\sqrt{J(n)}\gamma_n v(n)]^{-1})$$

a.s.;

(A4.s) There exist $M > 0$ and $\rho > \frac{1}{2}(1+\beta-\alpha)$ such that $\|r_{n+1}^{(1)}\| 1_{\{\|Z_n - z^*\| \le M\}} = O\left(n^{-\rho}\right)$ a.s.

Comments on the assumptions.

(a) Our assumptions are *local*. Thus, the results stated below can be applied as soon as $P\left[\Gamma(z^*)\right] > 0$, whatever the behavior of (Z_n) outside of $\Gamma(z^*)$ may be. In particular, they apply to algorithms obtained by projection or truncation in the framework of [3] or [12].

(b) Since the function $s \mapsto \sqrt{J(s)}\gamma(s)v(s)$ varies regularly with exponent $(1 + \beta - \alpha)/2$, assumption (A4.s) implies (A4.w).

(c) Assumptions (A2) about the noise and (A4) about the residual term can be applied to Markovian disturbances in the framework of [1], whose application to the averaged method is precised in [6].

(d) When the conditional moment of order 4 of the noise (ε_n) is bounded (i.e., when $b \ge 4$), assumption (A3)(iii) reduces to

$$\frac{1}{2} < \alpha < 1 \quad \text{and} \quad \frac{1-\alpha}{a-1} < \beta \le 1;$$

thus, the condition (A3.b) is useful only in the case $2 < b < 4$.

(e) In most cases, the function h is regular enough so that assumption (A1) holds with $a = 2$. In this case, assumption (A3) is fulfilled, for instance, by the gains

$$\begin{cases} \gamma_n = \dfrac{\gamma_0}{n^\alpha},\ \sigma_n = \dfrac{\sigma_0}{\sqrt{n^{\alpha+\beta}}}\ (\gamma_0 > 0,\ \sigma_0 > 0), \\ \text{with } \dfrac{2}{3} \leq \alpha < 1 \text{ and } 3(1-\alpha) \leq \beta \leq 1. \end{cases}$$

For these gains, we have

$$v(n) = \frac{\gamma_0 n^\beta}{\sigma_0^2},\quad J(n) = \frac{\sigma_0^2 n^{1+\alpha-\beta}}{\gamma_0^2(1+\alpha-\beta)},$$

and assumption (A4.w) can be rewritten as:

$$\text{there exists } M > 0 \text{ such that } \|r_{n+1}^{(1)}\| 1_{\{\|Z_n - z^*\| \leq M\}} = O(\sqrt{n^{1-\alpha+\beta}}) \text{ a.s.}$$

The Robbins–Monro algorithm is given by (8) with $\gamma_n = \sigma_n$ and $r_n = 0$; we have then $\beta = \alpha$, $J(n) = n$, and, if (A1) holds with $a = 2$, assumptions (A3) and (A4) are fulfilled, for instance, by the gains

$$\gamma_n = \sigma_n = \frac{\gamma_0}{n^\alpha}\quad (\gamma_0 > 0)\quad \text{with } \frac{3}{4} \leq \alpha < 1.$$

The Kiefer–Wolfowitz algorithm corresponds to the case $h = -\nabla V$, where the function V is observable only together with a noise. This algorithm can be written as (8) with $\gamma_n = \dfrac{\gamma_0}{n^\alpha}$ ($\gamma_0 > 0$), $1/2 < \alpha \leq 1$, and $\sigma_n = n^\tau \gamma_n$, $0 < \tau < \alpha/2$. We have then $\beta = \alpha - 2\tau$ and $J(n) = n^{1+2\tau}/(1+2\tau)$. Since $(n^{2\tau} r_{n+1})$ is known to converge a.s. on $\Gamma(z^*)$ toward a deterministic, usually nonzero constant, assumption (A4.w) (respectively, (A4.s)) requires $1/6 \leq \tau < \alpha/2$ (respectively, $1/6 < \tau < \alpha/2$). Consequently, if (A1) holds with $a = 2$, our assumptions (A3) and (A4) are fulfilled by the gains

$$\begin{cases} \gamma_n = \dfrac{\gamma_0}{n^\alpha},\ \sigma_n = n^\tau \gamma_n, \\ \text{with } \gamma_0 > 0,\ \dfrac{3}{4} + \dfrac{\tau}{2} \leq \alpha < 1, \text{ and} \begin{cases} \dfrac{1}{6} \leq \tau < \dfrac{\alpha}{2} & \text{for (A4.w)}, \\ \dfrac{1}{6} < \tau < \dfrac{\alpha}{2} & \text{for (A4.s)}. \end{cases} \end{cases} \tag{10}$$

Our first main result is the following quadratic strong law of large numbers.

THEOREM 3 (quadratic strong law of large numbers). *Assume* (A1), (A2), (A3), *and* (A4.s) *hold. Then, a.s. on* $\Gamma(z^*)$,

$$\lim_{n\to\infty} \frac{1}{\ln n} \sum_{k=1}^{n} k[J(k)]^{-1} \left(\overline{Z}_k - z^*\right) \left(\overline{Z}_k - z^*\right)^T = H^{-1} C \left(H^{-1}\right)^T.$$

COROLLARY 4 (almost sure efficiency). *Assume that* (A1), (A2), (A3), *and* (A4.s) *hold and that* $\gamma_n = \sigma_n$. *Then, a.s. on* $\Gamma(z^*)$,

$$\lim_{n\to\infty} \frac{1}{\ln n} \sum_{k=1}^{n} \left(\overline{Z}_k - z^*\right) \left(\overline{Z}_k - z^*\right)^T = H^{-1} C \left(H^{-1}\right)^T$$

and the almost sure asymptotic efficiency is obtained.

Remarks and examples.

(a) Under the assumptions of Theorem 3, Duflo [9] proved the following conditional CLT:

$$\text{given } \Gamma(z^*), \; Y_n = n[J(n)]^{-1/2}\left(\overline{Z}_n - z^*\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, H^{-1}C(H^{-1})^T). \tag{11}$$

Moreover, the quadratic strong law of large numbers can clearly be rewritten as:

$$\text{a.s. on } \Gamma(z^*),$$

$$\lim_{n\to\infty} \frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\left[k[J(k)]^{-1/2}(\overline{Z}_k - z^*)\right]\left[k[J(k)]^{-1/2}(\overline{Z}_k - z^*)\right]^T = H^{-1}C\big(H^{-1}\big)^T.$$

Thus, the quadratic strong law of large numbers ensures that the logarithmic average of the $Y_k Y_k^T$ converges a.s. toward the covariance matrix of the asymptotic distribution of (11).

(b) The averaged Robbins–Monro algorithm is asymptotically a.s. efficient on $\Gamma(z^*)$.

(c) In the case of the Kiefer–Wolfowitz algorithm, assumption (A4.s) requires that the parameter τ in (10) satisfies $\tau > 1/6$, and we then have

$$\text{a.s. on } \Gamma(z^*), \; \lim_{n\to\infty}\frac{1+2\tau}{\ln n}\sum_{k=1}^{n}\frac{1}{k^{2\tau}}\left(\overline{Z}_k - z^*\right)\left(\overline{Z}_k - z^*\right)^T = H^{-1}C\big(H^{-1}\big)^T.$$

However, we failed in proving a quadratic strong law of large numbers when $\tau = 1/6$. In view of remark (a), it is not surprising since, in this case, $n[J(n)]^{-1/2}\left(\overline{Z}_n - z^*\right)$ is known to converge weakly to a $\mathcal{N}(m, H^{-1}C(H^{-1})^T)$ distribution, where m is a deterministic, usually nonzero constant.

The following corollary gives an estimator of the asymptotic covariance matrix $H^{-1}C\big(H^{-1}\big)^T$, $\overline{Z}_n$ standing for z^* in Theorem 3.

COROLLARY 5 (strongly consistent estimator of the asymptotic covariance). *Set*

$$\widehat{\Sigma}_n = \frac{1}{\ln n}\sum_{k=1}^{n} k[J(k)]^{-1}\left(\overline{Z}_k - \overline{Z}_n\right)\left(\overline{Z}_k - \overline{Z}_n\right)^T.$$

Under assumptions (A1), (A2), (A3), *and* (A4.s), $\widehat{\Sigma}_n$ *is a strongly consistent estimator of* $H^{-1}C\big(H^{-1}\big)^T$ *on* $\Gamma(z^*)$.

Remark. The combination of (11) and Corollary 5 implies the following conditional CLT:

$$\text{given } \Gamma(z^*), \; n[J(n)]^{-1/2}\widehat{\Sigma}_n^{-1/2}\left(\overline{Z}_n - z^*\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I),$$

which permits the construction of confidence regions for z^*.

Our second main result is the following law of the iterated logarithm.

THEOREM 6 (law of the iterated logarithm). *Assume* (A1), (A2), (A3), *and* (A4.w) *hold. Then, for any vector* u *of* $\mathbb{R}^d$, *we have, a.s. on* $\Gamma(z^*)$,

$$\begin{aligned}\limsup_{n\to\infty}\frac{n}{\sqrt{2J(n)\ln(\ln n)}}u^T\left(\overline{Z}_n - z^*\right) &= -\liminf_{n\to\infty}\frac{n}{\sqrt{2J(n)\ln(\ln n)}}u^T\left(\overline{Z}_n - z^*\right)\\ &= \sqrt{u^T H^{-1}C(H^{-1})^T u}.\end{aligned}$$

Moreover, we have, a.s. on $\Gamma(z^*)$,

(12)
$$\forall u \in \mathbb{R}^d, \limsup_{n\to\infty} \frac{n^2}{2J(n)\ln(\ln n)} u^T\left(\overline{Z}_n - z^*\right)\left(\overline{Z}_n - z^*\right)^T u = u^T H^{-1}C\left(H^{-1}\right)^T u.$$

Remarks.

(a) Referring to the order of the symmetric matrices, property (12) can be written as:

$$\text{a.s. on } \Gamma(z^*),\ \limsup_{n\to\infty} \frac{n^2}{2J(n)\ln(\ln n)}\left(\overline{Z}_n - z^*\right)\left(\overline{Z}_n - z^*\right)^T = H^{-1}C\left(H^{-1}\right)^T.$$

(b) When $\gamma_n = \sigma_n$ (in particular, for the Robbins–Monro algorithm), we obtain

$$\text{a.s. on } \Gamma(z^*),\ \limsup_{n\to\infty} \frac{n}{2\ln(\ln n)}\left(\overline{Z}_n - z^*\right)\left(\overline{Z}_n - z^*\right)^T = H^{-1}C\left(H^{-1}\right)^T.$$

We see again that the asymptotic almost sure convergence rate of the averaged algorithm is optimal since the rate $(2\ln(\ln n))/n$ is known to be optimal and the limit covariance matrix $H^{-1}C\left(H^{-1}\right)^T$ is the smallest one (with respect to the order of the symmetric matrices).

(c) In the case of the Kiefer–Wolfowitz algorithm, we have

$$\text{a.s. on } \Gamma(z^*),\ \limsup_{n\to\infty} \frac{(1+2\tau)n^{1-2\tau}}{2\ln(\ln n)}\left(\overline{Z}_n - z^*\right)\left(\overline{Z}_n - z^*\right)^T = H^{-1}C\left(H^{-1}\right)^T$$

for any τ satisfying (10), and here we can choose $\tau = 1/6$, which ensures the optimal convergence rate of the averaged Kiefer–Wolfowitz algorithm.

(d) Theorems 3 and 6 extend previous results of Le Breton [15] and Le Breton and Novikov [16], [17]. They obtained Theorem 3 under the restriction that h is linear; under the same restriction, they obtained Theorem 6 in the unidimensional case ($d = 1$), whereas they obtained only an upper bound of $\left(\overline{Z}_n - z^*\right)$ when $d > 1$.

Our last main result is the following a.s.CLT.

THEOREM 7 (a.s.CLT). *Assume* (A1), (A2), (A3), *and* (A4.s) *hold. Then, a.s. on* $\Gamma(z^*)$,

$$\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\delta_{k[J(k)]^{-1/2}\left(\overline{Z}_k - z^*\right)} \Longrightarrow \mathcal{N}(0, H^{-1}C(H^{-1})^T).$$

The following corollary is a straightforward consequence of Theorems 3 and 7.

COROLLARY 8 (logarithmic strong law of large numbers). *Assume* (A1), (A2), (A3), *and* (A4.s) *hold. Let* $\phi : \mathbb{R}^d \to \mathbb{R}$ *be an almost everywhere continuous function such that, for a positive constant* K, $|\phi(x)| \le K(1 + \|x\|^2)$. *Then, a.s. on* $\Gamma(z^*)$,

$$\lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\phi\left[k[J(k)]^{-1/2}\left(Z_k - z^*\right)\right] = \int \phi(x)dF(x),$$

where F *is the* $\mathcal{N}(0, H^{-1}C(H^{-1})^T)$ *distribution.*

3. Application to efficient recursive estimation. Let (Y_k) be a sequence of independent identically distributed random vectors absolutely continuous with respect to some positive σ-finite measure λ. Let us denote by $f(\theta, .)$ the probability density of Y_k, where $\theta \in \Theta$, Θ is an open subset of $\mathbb{R}^d$, and assume this statistical model to be regular [4].

According to the classical asymptotic theory, the maximum likelihood estimator θ_n^*, which maximizes the likelihood of the sample $(Y_1, \dots, Y_n)$, is strongly consistent and asymptotically efficient, i.e.,

$$\theta_n^* \to \theta \text{ a.s. and } \sqrt{n}\,(\theta_n^* - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, [I(\theta)]^{-1}),$$

where $I(\theta) = E_\theta([\nabla \ln f(\theta, Y_k)]\,[\nabla \ln f(\theta, Y_k)]^T)$ is the Fisher information of the model. Touati [28] proved that θ_n^* is also a.s. asymptotically efficient, i.e.,

$$\frac{1}{\ln n} \sum_{k=1}^{n} (\theta_k^* - \theta)\,(\theta_k^* - \theta)^T \overset{\text{a.s.}}{\to} [I(\theta)]^{-1}.$$

However, the explicit computation of the maximum likelihood estimator is often impossible or very complicated and some approximation procedure is then necessary. For instance, Newton's recursive estimator is given by

$$\widehat{\theta}_{n+1}^* = \widehat{\theta}_n^* + \frac{\left[I(\widehat{\theta}_n^*)\right]^{-1}}{n} \nabla(\ln f(\widehat{\theta}_n^*, Y_{n+1})), \tag{13}$$

i.e.,

$$\widehat{\theta}_{n+1}^* = \widehat{\theta}_n^* + \frac{1}{n}\left[h(\widehat{\theta}_n^*) + \varepsilon_{n+1}\right],$$

where $h(t) = [I(t)]^{-1} \int \nabla\,(\ln f\,(t, x))\, dF_\theta(x)$.

Let us assume that there exists a constant $b > 2$ such that the function

$$t \mapsto \int \|\nabla(\ln f(t, x))\|^b dF_\theta(x)$$

exists and is bounded in the neighborhood of θ for each $\theta \in \Theta$.

It follows from a straightforward application of (2) and (3) (see [9, p. 168]) that

$$\text{given } \{\widehat{\theta}_n^* \to \theta\},\ \sqrt{n}(\widehat{\theta}_n^* - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, [I(\theta)]^{-1}),$$

and

$$\text{a.s. on } \{\widehat{\theta}_n^* \to \theta\},\ \frac{1}{\ln n} \sum_{k=1}^{n} (\widehat{\theta}_k^* - \theta)(\widehat{\theta}_k^* - \theta)^T \to [I(\theta)]^{-1}.$$

Thus, as soon as Newton's estimater is strongly consistent, it is also asymptotically efficient and a.s. asymptotically efficient.

However, the algorithm (13) requires at each step the computation of the inverse of the Fisher information matrix $[I(.)]^{-1}$. The use of an averaged algorithm does not require such a computation; for that, we proceed as follows. We determine $(\widehat{\theta}_n^*)$ from the gradient algorithm

$$\widehat{\theta}_{n+1}^* = \widehat{\theta}_n^* + \frac{1}{n^\alpha} \nabla(\ln f(\widehat{\theta}_n^*, Y_{n+1})), \quad \frac{3}{4} \leq \alpha < 1,$$

and we compute the empirical mean $\overline{\theta}_n^* = \frac{1}{n}\sum_{k=1}^n \widehat{\theta}_k^*$. It follows from a straightforward application of (11) and Theorem 3 that $\overline{\theta}_n^*$ has the same asymptotic properties as the $\widehat{\theta}_n^*$ given by (13); more precisely, given the event $\{\widehat{\theta}_n^* \to \theta\}$, $\overline{\theta}_n^*$ is a strongly consistent estimator, asymptotically efficient and a.s. asymptotically efficient. Moreover, applying Theorems 6 and 7, we also have

$$\text{a.s. on } \{\widehat{\theta}_n^* \to \theta\},\ \limsup_{n\to\infty} \frac{n}{2\ln(\ln n)} (\overline{\theta}_n^* - \theta)(\overline{\theta}_n^* - \theta)^T = [I(\theta)]^{-1}$$

and

$$\text{a.s. on } \{\widehat{\theta}_n^* \to \theta\},\ \frac{1}{\ln n}\sum_{k=1}^n \frac{1}{k}\delta_{\sqrt{k}(\overline{\theta}_k^* - \theta)} \Longrightarrow \mathcal{N}(0, [I(\theta)]^{-1}).$$

4. Proofs. In view of assumptions (A1) and (A4), the algorithm (8) can be rewritten as

$$Z_{n+1} = Z_n + \gamma_n H(Z_n - z^*) + \gamma_n r_{n+1} + \sigma_n \varepsilon_{n+1},$$

and we have

$$H(Z_n - z^*) = \frac{1}{\gamma_n}[(Z_{n+1} - z^*) - (Z_n - z^*)] - r_{n+1} - \frac{\sigma_n}{\gamma_n}\varepsilon_{n+1}.$$

Let us define the sequences (T_n), $(\overline{T}_n)$, (M_n), and (K_n) by

$$T_n = Z_n - z^*,\ \overline{T}_n = \frac{1}{n}\sum_{k=1}^n T_k,\ M_{n+1} = \sum_{k=1}^n \frac{\sigma_k}{\gamma_k}\varepsilon_{k+1},$$

$$K_n = -\frac{T_1}{\gamma_1} + \frac{T_{n+1}}{\gamma_n} - \sum_{k=2}^n T_k\left[\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right].$$

Then we have

$$nH\overline{T}_n = K_n - M_{n+1} - \sum_{k=1}^n r_{k+1}. \tag{14}$$

The proofs of the results stated in section 2 are constructed in the following way. First we establish the almost sure asymptotic properties (a quadratic strong law of large numbers, a law of the iterated logarithm, and an a.s.CLT) of the sequence (M_n). Then we show that (K_n) and $(\sum_{k=1}^n r_{k+1})$ are small enough on $\Gamma(z^*)$ so that the properties obtained for (M_n) are also satisfied by the sequence $\big(nH\overline{T}_n\big)$.

Let us first show how we can strengthen assumptions (A2) and (A4.w).

Note that in order to establish an almost sure property on $\Gamma(z^*)$, it is sufficient to prove it a.s. on $\Gamma_N = \Gamma(z^*) \cap \{\sup_{n\geq N}\|Z_n - z^*\| \leq M\}$ for any N such that $P(\Gamma_N) \neq 0$.

Let $\Gamma_{N,K}$ be the set of the trajectories of Γ_N such that, for a positive integer K,

$$\sup_{n\geq N} E(\|\varepsilon_{n+1}\|^b|\mathcal{F}_n) \leq K \text{ and } \sup_{n\geq N}(\sqrt{J(n)}\gamma_n v(n)\|r_{n+1}^{(1)}\|) \leq K.$$

Since Γ_N equals $\cup_K \Gamma_{N,K}$ up to a negligible set, it is sufficient to establish a property a.s. on $\Gamma_{N,K}$ for each K such that $P\left(\Gamma_{N,K}\right) \neq 0$, in order to prove it on Γ_N.

According to a technique often used by Lai and Wei (see [14], for instance), we modify the algorithm (8), without changing it on $\Gamma_{N,K}$, in order to have, a.s. on *the whole set* Ω,

(15)

$$E\left(\varepsilon_{n+1}|\mathcal{F}_n\right) = 0,\ \sup_{n\geq N} E(\|\varepsilon_{n+1}\|^b|\mathcal{F}_n) \leq K,\ \text{and}\ \sup_{n\geq N}(\sqrt{J(n)}\gamma_n v(n)\|r^{(1)}_{n+1}\|) \leq K.$$

To this end, we replace $r^{(1)}_{n+1}$ by $\widetilde{r}^{(1)}_{n+1} = r^{(1)}_{n+1}.1_{\sqrt{J(n)}\gamma_n v(n)\left\|r^{(1)}_{n+1}\right\|\leq K}$ and ε_{n+1} by $\widetilde{\varepsilon}_{n+1} = \varepsilon_{n+1}1_{B_n}$ with

$$B_n = \{E(\varepsilon_{n+1}|\mathcal{F}_n) = 0 \text{ and } E(\|\varepsilon_{n+1}\|^b|\mathcal{F}_n) \leq K\}.$$

From now on, we shall assume that these modifications have been made. Moreover, substituting (Z_n) for $(Z'_n) = (Z_{n+N})$, we shall assume that (15) is fulfilled with $N = 0$, i.e., that the following condition holds: there exists $K > 0$ such that, a.s. on Ω,

$$E\left(\varepsilon_{n+1}|\mathcal{F}_n\right) = 0,\ \sup_{n\geq 0} E(\|\varepsilon_{n+1}\|^b|\mathcal{F}_n) \leq K,\ \text{and}\ \sup_{n\geq 0}(\sqrt{J(n)}\gamma_n v(n)\|r^{(1)}_{n+1}\|) \leq K.$$

In the same way, we strengthen assumption (A4.s) and assume that there exists $\rho > (1-\alpha+\beta)/2$ such that $\sup_{n\geq 0}(n^\rho\|r^{(1)}_{n+1}\|) \leq K$ a.s. on Ω.

We now state the three lemmas, which give the almost sure properties of the square-integrable martingale (M_n); Lemmas 9 and 10 give, respectively, a law of the iterated logarithm and a quadratic strong law of large numbers for (M_n), whereas Lemma 11 establishes an a.s.CLT for the unidimensional sequence $\left(u^T M_n\right)$ for any vector u of $\mathbb{R}^d$.

Lemma 9 (law of the iterated logarithm for (M_n)). *Assume* (A2) *and* (A3). *Then, for any vector* $u \in \mathbb{R}^d$,

$$\limsup_{n\to\infty}\left[2J(n)\ln(\ln n)\right]^{-1/2} u^T M_n = -\liminf_{n\to\infty}\left[2J(n)\ln(\ln n)\right]^{-1/2} u^T M_n = \sqrt{u^T C u}\ a.s.$$

In particular, $\|M_n\|^2 = O\left(J(n)\ln(\ln n)\right)$ *a.s.*

Lemma 10 (quadratic strong law of large numbers for (M_n)). *Assume* (A2) *and* (A3). *Then,*

$$\lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^{n}[kJ(k)]^{-1}M_k M_k^T = C\ a.s.$$

Lemma 11 (a.s.CLT for $(u^T M_n)$). *Assume* (A2) *and* (A3). *Then, for any vector* $u \in \mathbb{R}^d$,

$$\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\delta_{[J(k)]^{-1/2}u^T M_k} \Longrightarrow \mathcal{N}\left(0, u^T C u\right)\ a.s.$$

Our proofs are now organized as follows. First, we show in section 4.1 how Theorems 3, 6, and 7 can be deduced from Lemmas 9, 10, and 11. Then, Corollary 5 is established in section 4.2. Finally, section 4.3 is devoted to the proofs of Lemmas 9, 10, and 11.

Throughout the proofs, $\mathcal{L}$ denotes a generic, increasing, and slowly varying function.

4.1. Proof of Theorems 3, 6, and 7. In the case assumptions (A2) and (A3.b) hold, we have $\frac{1}{b} < \min\{\frac{1}{2}\ ;\ \frac{1}{2a}[\frac{(a-1)\beta}{1-\alpha} - 1]\}$; throughout this subsection, we then set δ such that

$$\frac{1}{b} < \delta < \min\left\{\frac{1}{2}\ ;\ \frac{1}{2a}\left[\frac{(a-1)\beta}{1-\alpha} - 1\right]\right\}. \tag{16}$$

4.1.1. Preliminaries. In this subsection, we establish two lemmas we shall use in the proofs of Theorems 3, 6, and 7.

LEMMA 12. *Assume* (A1), (A2), *and* (A4.w) *hold. Then, we have, a.s. on* $\Gamma(z^*)$,

(i) *under* (A3.a), $\|K_n\| = O[(1 + n^{\alpha - \frac{\beta}{2}})\mathcal{L}(n)]$ *and*

$$\sum_{k=1}^{n} \|r_{k+1}^{(2)}\| = O\big[(1 + n^{1-\frac{a\beta}{2}})\mathcal{L}(n)\big],$$

(ii) *under* (A3.b), $\|K_n\| = O[(1 + n^{\delta(1-\alpha)-\frac{\beta}{2}+\alpha})\mathcal{L}(n)]$ and $\sum_{k=1}^{n} \|r_{k+1}^{(2)}\| = O[(1 + n^{1+\frac{a}{2}[2\delta(1-\alpha)-\beta]})\mathcal{L}(n)]$, *where* δ *is given by* (16).

Proof of Lemma 12. In view of assumption (A3)(ii), we have

$$\|K_n\| = O\left[1 + \frac{\|T_{n+1}\|}{\gamma_n} + \sum_{k=2}^{n} \|T_k\|\,\theta(k)\right]$$

with $\theta = (1/\gamma)'$. Let us apply the following result proved in [21].

RESULT 1 (almost sure upper bounds of $(Z_n - z^*)$ on $\Gamma(z^*)$). *Assume* (A1), (A2), *and* (A4.w) *hold. Then, we have*

(i) *under* (A3.a), $\|Z_n - z^*\| = O([v(n)]^{-1/2}[\ln(\sum_{k=1}^{n}\gamma_k)]^{1/2})$ *a.s. on* $\Gamma(z^*)$,

(ii) *under* (A3.b), *for any* ζ *such that* $\zeta > \frac{1}{b}$, $\|Z_n - z^*\| = O([v(n)]^{-1/2}(\sum_{k=1}^{n}\gamma_k)^{\zeta})$ *a.s. on* $\Gamma(z^*)$.

It follows that, under assumption (A3.a), we have

$$\begin{aligned}\|K_n\| &= O\left[1 + \frac{[\ln(\sum_{k=1}^{n+1}\gamma_k)]^{1/2}}{\gamma_n\sqrt{v(n+1)}} + \sum_{k=2}^{n}\frac{[\ln(\sum_{k=1}^{n}\gamma_k)]^{1/2}\theta(k)}{\sqrt{v(k)}}\right] \\ &= O\big[(1 + n^{\alpha-\frac{\beta}{2}})\mathcal{L}(n)\big] \text{ a.s. on } \Gamma(z^*)\end{aligned}$$

and, under assumption (A3.b),

$$\begin{aligned}\|K_n\| &= O\left[1 + \frac{(\sum_{k=1}^{n+1}\gamma_k)^{\delta}}{\gamma_n\sqrt{v(n+1)}} + \sum_{k=2}^{n}\frac{(\sum_{j=1}^{k}\gamma_j)^{\delta}\theta(k)}{\sqrt{v(k)}}\right] \\ &= O\big[(1 + n^{\delta(1-\alpha)+\alpha-\frac{\beta}{2}})\mathcal{L}(n)\big] \text{ a.s. on } \Gamma(z^*).\end{aligned}$$

On the other hand, since $\|r_{k+1}^{(2)}\| = O(\|Z_k - z^*\|^{a})$, we deduce from Result 1 that, under assumption (A3.a),

$$\begin{aligned}\sum_{k=1}^{n}\|r_{k+1}^{(2)}\| &= O\left[\sum_{k=1}^{n}([\ln(\textstyle\sum_{j=1}^{k}\gamma_j)]^{a/2}[v(k)]^{-a/2})\right] \\ &= O[(1 + n^{1-\frac{a\beta}{2}})\mathcal{L}(n)] \text{ a.s. on } \Gamma(z^*)\end{aligned}$$

and, under assumption (A3.b),

$$\begin{aligned}\sum_{k=1}^{n} \|r_{k+1}^{(2)}\| &= O\left[\sum_{k=1}^{n} ([\textstyle\sum_{j=1}^{k} \gamma_j]^{a\delta} [v(k)]^{-a/2})\right] \\ &= O[(1 + n^{1+\frac{a}{2}[2\delta(1-\alpha)-\beta]})\mathcal{L}(n)] \text{ a.s. on } \Gamma(z^*),\end{aligned}$$

which concludes the proof of Lemma 12.

LEMMA 13. *Assume* (A1), (A2), *and* (A3) *hold. Then,*

(i) *under* (A4.w), *we have* $[J(n)]^{-1/2} (\|K_n\| + \sum_{k=1}^{n} \|r_{k+1}\|) = O(1)$ *a.s. on* $\Gamma(z^*)$,

(ii) *under* (A4.s), *there exists* $c > 0$ *such that* $[J(n)]^{-1/2} (\|K_n\| + \sum_{k=1}^{n} \|r_{k+1}\|) = O(n^{-c})$ *a.s. on* $\Gamma(z^*)$.

Proof of Lemma 13. The application of Lemma 12 leads to the following almost sure upper bounds on $\Gamma(z^*)$:

Sequence	Under assumption (A3.a)
$[J(n)]^{-1/2} \|K_n\|$ $[J(n)]^{-1/2} \sum_{k=1}^{n} \|r_{k+1}^{(2)}\|$	$O[(n^{-\frac{1}{2}(1-\beta+\alpha)} + n^{-\frac{1}{2}(1-\alpha)})\mathcal{L}(n)]$ $O[(n^{-\frac{1}{2}(1-\beta+\alpha)} + n^{\frac{1}{2}[(1-\alpha)-(a-1)\beta]})\mathcal{L}(n)]$
Sequence	Under assumption (A3.b)
$[J(n)]^{-1/2} \|K_n\|$ $[J(n)]^{-1/2} \sum_{k=1}^{n} \|r_{k+1}^{(2)}\|$	$O[(n^{-\frac{1}{2}(1-\beta+\alpha)} + n^{(\delta-\frac{1}{2})(1-\alpha)})\mathcal{L}(n)]$ $O[(n^{-\frac{1}{2}(1-\beta+\alpha)} + n^{\frac{1}{2}[(1+2a\delta)(1-\alpha)-(a-1)\beta]})\mathcal{L}(n)]$

.

Since all the exponents are strictly negative, there exists $c_1 > 0$ such that

$$[J(n)]^{-1/2} \left(\|K_n\| + \sum_{k=1}^{n} \|r_{k+1}^{(2)}\|\right) = O(n^{-c_1}) \text{ a.s. on } \Gamma(z^*). \tag{17}$$

Now, under assumption (A4.w), we have

$$[J(n)]^{-1/2} \sum_{k=1}^{n} \|r_{k+1}^{(1)}\| = O\left([J(n)]^{-1/2} \sum_{k=1}^{n} \left[\sqrt{J(k)}\gamma_k v(k)\right]^{-1}\right) \quad \text{a.s. on } \Gamma(z^*).$$

Since the function $s \mapsto [\sqrt{J(s)}\gamma(s)v(s)]^{-1}$ varies regularly with exponent $-(1-\alpha+\beta)/2 > -1$, we have [11, p. 281]

$$\lim_{t\to\infty} \frac{t\left[\sqrt{J(t)}\gamma(t)v(t)\right]^{-1}}{\int_0^t \left[\sqrt{J(s)}\gamma(s)v(s)\right]^{-1}} = \frac{1}{2}[(1-\beta)+\alpha], \quad \frac{1}{2}[(1-\beta)+\alpha] \neq 0.$$

It follows that

$$\sum_{k=1}^{n} \left[\sqrt{J(k)}\gamma_k v(k)\right]^{-1} = O\left(n\left[\sqrt{J(n)}\gamma_n v(n)\right]^{-1}\right)$$

$$\text{and } [J(n)]^{-1/2} \sum_{k=1}^{n} \|r_{k+1}^{(1)}\| = O(n[J(n)\gamma_n v(n)]^{-1}) \quad \text{a.s. on } \Gamma(z^*).$$

However, the function $s \mapsto \gamma(s)v(s)$ varies regularly with exponent $\alpha - \beta > -1$; thus

$$\frac{t[\gamma(t)v(t)]^{-1}}{J(t)} \to (1-\beta)+\alpha \neq 0,$$

which implies that

$$[J(n)]^{-1/2}\sum_{k=1}^{n}\|r_{k+1}^{(1)}\| = O(1) \text{ a.s. on } \Gamma(z^*). \tag{18}$$

The first part of Lemma 13 then follows from the combination of (17) and (18).

Now, under assumption (A4.s), we have

$$\begin{aligned}[J(n)]^{-1/2}\sum_{k=1}^{n}\|r_{k+1}^{(1)}\| &= O\left([J(n)]^{-1/2}\sum_{k=1}^{n}k^{-\rho}\right)\\ &= O(n^{-\frac{1}{2}(1+\alpha-\beta)}\mathcal{L}(n)[n^{1-\rho}+\ln n])\\ &= O([n^{\frac{1}{2}(1-\alpha+\beta)-\rho}+n^{-\frac{1}{2}(1+\alpha-\beta)}]\mathcal{L}(n))\\ &= O(n^{-c_2}) \text{ with } c_2>0, \qquad (19)\end{aligned}$$

and the second part of Lemma 13 follows from the combination of (17) and (19).

4.1.2. Proof of Theorem 6. In view of (14) and Lemma 13 (i), we have, for any $u \in \mathbb{R}^d$,

$$\begin{aligned}\frac{u^T nH\overline{T}_n}{\sqrt{2J(n)\ln(\ln n)}} &= \frac{-u^T M_{n+1}}{\sqrt{2J(n)\ln(\ln n)}} + \frac{u^T\left(K_n - \sum_{k=1}^{n} r_{k+1}\right)}{\sqrt{2J(n)\ln(\ln n)}}\\ &= \frac{-u^T M_{n+1}}{\sqrt{2J(n)\ln(\ln n)}} + o(1) \text{ a.s.}\end{aligned}$$

It then follows from Lemma 9 that, a.s. on $\Gamma(z^*)$,

$$\limsup_{n\to\infty}\frac{nu^T H\left(\overline{Z}_n - z^*\right)}{\sqrt{2J(n)\ln(\ln n)}} = -\liminf_{n\to\infty}\frac{nu^T H\left(\overline{Z}_n - z^*\right)}{\sqrt{2J(n)\ln(\ln n)}} = \sqrt{u^T C u}$$

and, replacing u by $[(H^T)^{-1}u]$ (H^T is nonsingular), we deduce that, a.s. on $\Gamma(z^*)$,

$$\limsup_{n\to\infty}\frac{nu^T\left(\overline{Z}_n - z^*\right)}{\sqrt{2J(n)\ln(\ln n)}} = -\liminf_{n\to\infty}\frac{nu^T\left(\overline{Z}_n - z^*\right)}{\sqrt{2J(n)\ln(\ln n)}} = \sqrt{u^T H^{-1}C(H^T)^{-1}u}, \tag{20}$$

which concludes the proof of the first assertion of Theorem 6.

Now, (20) implies that for any $u \in \mathbb{R}^d$, a.s. on $\Gamma(z^*)$,

$$\limsup_{n\to\infty}\frac{n^2}{2J(n)\ln(\ln n)}u^T\left(\overline{Z}_n - z^*\right)\left(\overline{Z}_n - z^*\right)^T u = u^T H^{-1}C\left(H^T\right)^{-1}u$$

and, $\mathbb{Q}$ being a countable set, there exists a P-null set N such that $\forall\omega \in \Gamma(z^*)\setminus N$, $\forall u \in \mathbb{Q}^d$,

$$\limsup_{n\to\infty}\frac{n^2}{2J(n)\ln(\ln n)}u^T\left(\overline{Z}_n(\omega) - z^*\right)\left(\overline{Z}_n(\omega) - z^*\right)^T u = u^T H^{-1}C\left(H^T\right)^{-1}u. \tag{21}$$

To conclude the proof of Theorem 6, we have to show that for any $\omega_0 \in \Gamma(z^*)\setminus N$ and any $v \in \mathbb{R}^d$

$$\limsup_{n\to\infty} v^T\Sigma_n v = v^T\Sigma v, \tag{22}$$

$$\text{where } \Sigma_n = \frac{n^2\left(\overline{Z}_n(\omega_0)-z^*\right)\left(\overline{Z}_n(\omega_0)-z^*\right)^T}{2J(n)\ln(\ln n)} \text{ and } \Sigma = H^{-1}C(H^T)^{-1}.$$

Set $\varepsilon > 0$ and $u \in \mathbb{Q}^d$ such that $\|v-u\| \le \varepsilon$; we have

$$\begin{aligned} v^T\left(\Sigma_n-\Sigma\right)v &\le u^T\left(\Sigma_n-\Sigma\right)u + \left\|(v-u)^T\left(\Sigma_n-\Sigma\right)v + u^T\left(\Sigma_n-\Sigma\right)(v-u)\right\|, \\ \limsup_{n\to\infty} v^T\left(\Sigma_n-\Sigma\right)v &\le \limsup_{n\to\infty} u^T\left(\Sigma_n-\Sigma\right)u \\ &\quad + \limsup_{n\to\infty}\left[\|v-u\|\left(\|\Sigma_n\|+\|\Sigma\|\right)\left(\|u\|+\|v\|\right)\right], \\ &\le \limsup_{n\to\infty}\left[\varepsilon\left(\|\Sigma_n\|+\|\Sigma\|\right)\left(\varepsilon+2\|v\|\right)\right] \text{ in view of (21).} \end{aligned}$$

However, (21) implies that $\|\overline{Z}_n(\omega_0)-z^*\| = O(1)$; thus

$$\limsup_{n\to\infty} v^T\left(\Sigma_n-\Sigma\right)v \le \varepsilon C(\varepsilon+2\|v\|), \text{ where } C>0.$$

It follows that $\limsup_{n\to\infty} v^T\Sigma_n v \le v^T\Sigma v$.

On the other hand, (21) implies that there exists a sequence of integers $(t(n))$ and $n_0 \in \mathbb{N}$ such that $\lim_{n\to\infty} t(n) = \infty$ and $\forall n \ge n_0$, $\|u^T(\Sigma_{t(n)}-\Sigma)u\| \le \varepsilon$. We then have, $\forall n \ge n_0$,

$$\left\|v^T\left(\Sigma_{t(n)}-\Sigma\right)v\right\| \le \varepsilon C(\varepsilon+2\|v\|).$$

Thus $\lim_{n\to\infty} v^T\Sigma_{t(n)}v = v^T\Sigma v$ and (22) is proved.

4.1.3. Proof of Theorem 3. In view of (14), we have

$$\begin{aligned} n^2 H\overline{T}_n\overline{T}_n^T H^T &= \left(K_n - M_{n+1} - \sum_{k=1}^n r_{k+1}\right)\left(K_n - M_{n+1} - \sum_{k=1}^n r_{k+1}\right)^T \\ &= M_n M_n^T + \widetilde{R}_n, \end{aligned}$$

where $\widetilde{R}_n = n^2 H\overline{T}_n\overline{T}_n^T H^T - M_{n+1}M_{n+1}^T$; thus

$$\frac{1}{\ln n}\sum_{k=1}^n k[J(k)]^{-1}H\overline{T}_k\overline{T}_k^T H^T = \frac{1}{\ln n}\sum_{k=1}^n [kJ(k)]^{-1}M_k M_k^T + \frac{1}{\ln n}\sum_{k=1}^n [kJ(k)]^{-1}\widetilde{R}_k.$$

Lemmas 9 and 12 and assumption (A4.s) lead to the following almost sure upper bounds on $\Gamma(z^*)$:

Sequence	Under assumption (A3.a)
$[J(k)]^{-1}\Vert K_k\Vert^2$	$O[(k^{-(1-\beta+\alpha)}+k^{-(1-\alpha)})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert\sum_{j=1}^k r_{j+1}^{(1)}\Vert^2$	$O[(k^{-(1-\beta+\alpha)}+k^{(1+\beta-\alpha)-2\rho})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert\sum_{j=1}^k r_{j+1}^{(2)}\Vert^2$	$O[(k^{-(1-\beta+\alpha)}+k^{(1-\alpha)-(a-1)\beta})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert K_k\Vert\Vert M_k\Vert$	$O[(k^{-\frac{1}{2}(1+\alpha-\beta)}+k^{-\frac{1}{2}(1-\alpha)})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert\sum_{j=1}^k r_{j+1}^{(1)}\Vert\Vert M_k\Vert$	$O[(k^{-\frac{1}{2}(1+\alpha-\beta)}+k^{\frac{1}{2}(1+\beta-\alpha)-\rho})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert\sum_{j=1}^k r_{j+1}^{(2)}\Vert\Vert M_k\Vert$	$O[(k^{-\frac{1}{2}(1+\alpha-\beta)}+k^{\frac{1}{2}[(1-\alpha)-(a-1)\beta]})\mathcal{L}(k)]$

Sequence	Under assumption (A3.b)
$[J(k)]^{-1}\Vert K_k\Vert^2$	$O[(k^{-(1-\beta+\alpha)}+k^{(2\delta-1)(1-\alpha)})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert\sum_{j=1}^k r_{j+1}^{(1)}\Vert^2$	$O[(k^{-(1-\beta+\alpha)}+k^{(1+\beta-\alpha)-2\rho})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert\sum_{j=1}^k r_{j+1}^{(2)}\Vert^2$	$O[(k^{-(1-\beta+\alpha)}+k^{(1+2a\delta)(1-\alpha)-(a-1)\beta})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert K_k\Vert\Vert M_k\Vert$	$O[(k^{-\frac{1}{2}(1+\alpha-\beta)}+k^{(\delta-\frac{1}{2})(1-\alpha)})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert\sum_{j=1}^k r_{j+1}^{(1)}\Vert\Vert M_k\Vert$	$O[(k^{-\frac{1}{2}(1+\alpha-\beta)}+k^{\frac{1}{2}(1+\beta-\alpha)-\rho})\mathcal{L}(k)]$
$[J(k)]^{-1}\Vert\sum_{j=1}^k r_{j+1}^{(2)}\Vert\Vert M_k\Vert$	$O[(k^{-\frac{1}{2}(1+\alpha-\beta)}+k^{\frac{1}{2}[(1+2a\delta)(1-\alpha)-(a-1)\beta]})\mathcal{L}(k)]$

Since all the exponents are strictly negative, we deduce that $\sum[kJ(k)]^{-1}\|\widetilde{R}_k\|<\infty$ a.s. on $\Gamma(z^*)$, and thus

$$\frac{1}{\ln n}\sum_{k=1}^{n}k[J(k)]^{-1}H\overline{T}_k\overline{T}_k^T H^T=\frac{1}{\ln n}\sum_{k=1}^{n}[kJ(k)]^{-1}M_kM_k^T+o(1)\text{ a.s. on }\Gamma(z^*).$$

Lemma 10 then implies

$$\lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^{n}k[J(k)]^{-1}H\overline{T}_k\overline{T}_k^T H^T=C\text{ a.s. on }\Gamma(z^*)$$

and thus

$$\lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^{n}k[J(k)]^{-1}\left(\overline{Z}_k-z^*\right)\left(\overline{Z}_k-z^*\right)^T=H^{-1}C\big(H^{-1}\big)^T\text{ a.s. on }\Gamma(z^*).$$

4.1.4. Proof of Theorem 7. We have to prove that there exists a P-null set N such that $\forall\omega\in\Gamma(z^*)\setminus N$

$$\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\delta_{k[J(k)]^{-1/2}\left(\overline{Z}_k(\omega)-z^*\right)}\Longrightarrow\mathcal{N}(0,H^{-1}C(H^{-1})^T).$$

We first study the behavior of the characteristic functions of the random measures $\frac{1}{\ln n}\sum_{k=1}^n\frac{1}{k}\delta_{k[J(k)]^{-1/2}\left(\overline{Z}_k-z^*\right)}$.

Let u be any vector of $\mathbb{R}^d$. In view of (14), we have

$$\begin{aligned}&\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\exp[i[J(k)]^{-1/2}u^T(kH\overline{T}_k)]\\&=\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\exp\left[i[J(k)]^{-1/2}u^T\bigg(-M_{k+1}+K_k-\sum_{j=1}^{k}r_{j+1}\bigg)\right].\end{aligned}$$

Applying Lemma 13 (ii), it follows that, a.s. on $\Gamma(z^*)$,

$$\begin{aligned}\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}&\exp[i[J(k)]^{-1/2}u^T(kH\overline{T}_k)]\\&=\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\exp[-i[J(k)]^{-1/2}u^TM_{k+1}+O(k^{-c})]\\&=\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\exp[-i[J(k)]^{-1/2}u^TM_{k+1}]+o(1).\end{aligned}$$

Thus, in view of Lemma 11,

$$\lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\exp[i(H^Tu)^T(k[J(k)]^{-1/2}\overline{T}_k)]=\exp\left[\frac{-u^TCu}{2}\right].$$

It follows that, for any vector u in $\mathbb{R}^d$, we have, a.s. on $\Gamma(z^*)$,

$$\lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\exp[iu^T(k[J(k)]^{-1/2}(\overline{Z}_k-z^*))]=\exp\left[\frac{-u^TH^{-1}C\big(H^{-1}\big)^Tu}{2}\right].$$

Since $\mathbb{Q}$ is a countable set, there exists a P-null set $N\subset\Gamma(z^*)$ such that $\forall\omega\in\Gamma(z^*)\setminus N$, $\forall u\in\mathbb{Q}^d$,

(23)
$$\lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\exp[iu^T(k[J(k)]^{-1/2}(\overline{Z}_k(\omega)-z^*))]=\exp\left[\frac{-u^TH^{-1}C\big(H^{-1}\big)^Tu}{2}\right].$$

Let us now set $\omega_0\in\Gamma(z^*)\setminus N$ and prove that the sequence of the deterministic measures $(\mu_n(\omega_0))$ defined by

$$\mu_n(\omega_0)=\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\delta_{k[J(k)]^{-1/2}(\overline{Z}_k(\omega_0)-z^*)}$$

converges weakly to the $\mathcal{N}(0,H^{-1}C(H^{-1})^T)$ distribution.

Let μ_0 be a closure point of $(\mu_n(\omega_0))$ and $\mu_{p(n)}(\omega_0)$ a subsequence such that $\mu_{p(n)}(\omega_0)\Longrightarrow\mu_0$. Since $(\mu_{p(n)}(\omega_0))$ is a bounded sequence of measures, μ_0 is a bounded measure; let ϕ_0 (respectively, $\phi_{p(n)}$) be the characteristic function of μ_0 (respectively, $\mu_{p(n)}(\omega_0)$). We then have $\lim_{n\to\infty}\phi_{p(n)}(u)=\phi_0(u)$ for any $u\in\mathbb{R}^d$, and, in view of (23), $\phi_0(u)=\exp[-u^TH^{-1}C(H^{-1})^Tu/2]$ for any $u\in\mathbb{Q}^d$. However, the function ϕ_0 is continuous, thus $\phi_0(u)=\exp[-u^TH^{-1}C(H^{-1})^Tu/2]$ for any $u\in\mathbb{R}^d$. We finally deduce that μ_0 is the $\mathcal{N}(0,H^{-1}C(H^{-1})^T)$ distribution and $\mu_n(\omega_0)\Longrightarrow\mathcal{N}(0,H^{-1}C(H^{-1})^T)$, which concludes the proof of Theorem 7.

4.2. Proof of Corollary 5. The estimator $\widehat{\Sigma}_n$ can be written as

$$\begin{aligned}\widehat{\Sigma}_n=&\frac{1}{\ln n}\sum_{k=1}^{n}k[J(k)]^{-1}\left(\overline{Z}_k-z^*\right)\left(\overline{Z}_k-z^*\right)^T+\frac{1}{\ln n}\sum_{k=1}^{n}k[J(k)]^{-1}\left(\overline{Z}_n-z^*\right)\left(\overline{Z}_n-z^*\right)^T\\&-\frac{1}{\ln n}\sum_{k=1}^{n}k[J(k)]^{-1}\left(\overline{Z}_k-z^*\right)\left(\overline{Z}_n-z^*\right)^T-\frac{1}{\ln n}\sum_{k=1}^{n}k[J(k)]^{-1}\left(\overline{Z}_n-z^*\right)\left(\overline{Z}_k-z^*\right)^T.\end{aligned}$$

Applying Theorem 6, we have, a.s. on $\Gamma\left(z^{*}\right)$,

$$\left(\sum_{k=1}^{n} k[J(k)]^{-1}\right)\left\|\overline{Z}_{n}-z^{*}\right\|^{2}=O\left[\left(\int_{1}^{n} s[J(s)]^{-1} d s\right) \frac{J(n) \ln (\ln n)}{n^{2}}\right].$$

Since the function $s \mapsto s[J(s)]^{-1}$ varies regularly with exponent $\beta-\alpha>-1$, we have

$$\lim_{n \rightarrow \infty} \frac{n^{2}[J(n)]^{-1}}{\int_{1}^{n} s[J(s)]^{-1} d s}=1+\beta-\alpha, \quad 1+\beta-\alpha \neq 0;$$

thus, a.s. on $\Gamma\left(z^{*}\right)$,

$$\left(\sum_{k=1}^{n} k[J(k)]^{-1}\right)\left\|\overline{Z}_{n}-z^{*}\right\|^{2}=O[\ln (\ln n)]$$

and

$$\lim_{n \rightarrow \infty} \frac{1}{\ln n} \sum_{k=1}^{n} k[J(k)]^{-1}\left(\overline{Z}_{n}-z^{*}\right)\left(\overline{Z}_{n}-z^{*}\right)^{T}=0.$$

On the other hand, applying Theorem 6 again, we obtain, a.s. on $\Gamma\left(z^{*}\right)$,

$$\begin{aligned}
&\left(\sum_{k=1}^{n} k[J(k)]^{-1}\left\|\overline{Z}_{k}-z^{*}\right\|\right)\left\|\overline{Z}_{n}-z^{*}\right\| \\
&\quad=O\left[\left(\sum_{k=1}^{n}[J(k)]^{-1/2} \sqrt{\ln (\ln k)}\right) \frac{\sqrt{J(n) \ln (\ln n)}}{n}\right] \\
&\quad=O\left[\left(\int_{1}^{n}[J(s)]^{-1/2} \sqrt{\ln (\ln s)} d s\right) \frac{\sqrt{J(n) \ln (\ln n)}}{n}\right].
\end{aligned}$$

Since the function $s \mapsto [J(s)]^{-1/2} \sqrt{\ln (\ln s)}$ varies regularly with exponent $-\frac{1}{2}(1+\alpha-\beta)<1$, we have

$$\lim_{n \rightarrow \infty} \frac{n[J(n)]^{-1/2} \sqrt{\ln \ln n}}{\int_{1}^{n}[J(s)]^{-1/2} \sqrt{\ln (\ln s)} d s}=\frac{1}{2}(1-\alpha+\beta), \quad \frac{1}{2}(1-\alpha+\beta) \neq 0.$$

Thus, a.s. on $\Gamma\left(z^{*}\right)$,

$$\left(\sum_{k=1}^{n} k[J(k)]^{-1}\left\|\overline{Z}_{k}-z^{*}\right\|\right)\left\|\overline{Z}_{n}-z^{*}\right\|=O(\ln \ln n)$$

and

$$\begin{aligned}
\lim_{n \rightarrow \infty} & \left[\frac{1}{\ln n} \sum_{k=1}^{n} k[J(k)]^{-1}\left(\overline{Z}_{k}-z^{*}\right)\left(\overline{Z}_{n}-z^{*}\right)^{T}\right. \\
&\left.+\frac{1}{\ln n} \sum_{k=1}^{n} k[J(k)]^{-1}\left(\overline{Z}_{n}-z^{*}\right)\left(\overline{Z}_{k}-z^{*}\right)^{T}\right]=0.
\end{aligned}$$

Applying Theorem 3, we finally deduce that $\lim_{n \rightarrow \infty} \widehat{\Sigma}_{n}=H^{-1} C\left(H^{-1}\right)^{T}$ a.s. on $\Gamma\left(z^{*}\right)$.

4.3. Proof of Lemmas 9, 10, and 11. Throughout this subsection, $\mathrm{Tr}(A)$ denotes the trace of a matrix A, and $\langle\widetilde{M}\rangle_n$ the increasing process of a square-integrable martingale $(\widetilde{M}_n)$. Recall that $\langle\widetilde{M}\rangle_0 = I$ and $E((\widetilde{M}_{n+1} - \widetilde{M}_n)(\widetilde{M}_{n+1} - \widetilde{M}_n)^T|\mathcal{F}_n) = \langle\widetilde{M}\rangle_{n+1} - \langle\widetilde{M}\rangle_n$.

4.3.1. Proof of Lemma 9. The proof of Lemma 9 is based upon the following adaptation of the law of the iterated logarithm of Stout [27] (see [8] or [10]).

Result 2 (law of the iterated logarithm for unidimensional martingales). *Let (η_n) be a sequence of unidimensional random variables adapted to a filtration $\mathcal{F}$ such that*

$$E(\eta_{n+1}|\mathcal{F}_n) = 0\ \forall n \geq 0, \quad \limsup_{n\to\infty} E(|\eta_{n+1}|^2|\mathcal{F}_n) = c^2 < \infty, \quad \textit{and}$$

$$\exists \xi \in]0,1[\ \textit{s.t.}\ \sup_{n\geq 0} E(|\eta_{n+1}|^{2(1+\xi)}|\mathcal{F}_n) < +\infty \quad \textit{a.s.}$$

Let (Φ_n) be a sequence of unidimensional random variables adapted to $\mathcal{F}$ and set $\tau_n = \sum_{k=0}^n \Phi_k^2$; if $\tau_\infty = +\infty$, $\sum \Phi_n^{2+2\xi}\tau_n^{-1-\xi} < +\infty$ and $\Phi_n^2 = o(\tau_n(\ln\ln\tau_n)^{-1/\xi})$ a.s.; then

$$\limsup_{n\to\infty} [2\tau_n \ln(\ln\tau_n)]^{-1/2} \sum_{k=0}^n \Phi_k\eta_{k+1} = -\liminf_{n\to\infty} [2\tau_n \ln(\ln\tau_n)]^{-1/2} \sum_{k=0}^n \Phi_k\eta_{k+1} = c \ \textit{a.s.}$$

Set $u \in \mathbb{R}^d$; the application of Result 2 with $\Phi_n = \sigma_n\gamma_n^{-1}$ and $\eta_{n+1} = u^T\varepsilon_{n+1}$ leads to

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{\sum_{k=1}^n \sigma_k\gamma_k^{-1}u^T\varepsilon_{k+1}}{\left[2\left(\sum_{k=1}^n \sigma_k^2\gamma_k^{-2}\right)\ln\ln\left(\sum_{k=1}^n \sigma_k^2\gamma_k^{-2}\right)\right]^{1/2}} \\
&\quad = -\liminf_{n\to\infty} \frac{\sum_{k=1}^n \sigma_k\gamma_k^{-1}u^T\varepsilon_{k+1}}{\left[2\left(\sum_{k=1}^n \sigma_k^2\gamma_k^{-2}\right)\ln\ln\left(\sum_{k=1}^n \sigma_k^2\gamma_k^{-2}\right)\right]^{1/2}} \\
&\quad = \sqrt{u^TCu} \quad \text{a.s.}
\end{aligned}$$

However, $\sigma_k^2\gamma_k^{-2} = [\gamma_k v(k)]^{-1}$; thus $\sum_{k=1}^n \sigma_k^2\gamma_k^{-2} \sim J(n)$ and we obtain

$$\limsup_{n\to\infty} [2J(n)\ln(\ln n)]^{-1/2}u^TM_n = -\liminf_{n\to\infty} [2J(n)\ln(\ln n)]^{-1/2}u^TM_n = \sqrt{u^TCu} \ \text{a.s.}$$

4.3.2. Proof of Lemma 10. The proof of Lemma 10 is based upon the following martingale version of a result established by Wei [30] for regressive sequences.

Lemma 14 (a strong law of large numbers for martingales). *Let $(\widetilde{M}_n)$ be a d-dimensional, square-integrable martingale with respect to a filtration $\mathcal{F}$.*

Set

$$H_n = \sum_{k=1}^n \widetilde{M}_k^T[\langle\widetilde{M}\rangle_{k-1}^{-1} - \langle\widetilde{M}\rangle_k^{-1}]\widetilde{M}_k,$$

$$f_n = \mathrm{Tr}(\langle\widetilde{M}\rangle_{n+1}^{-1/2}[\langle\widetilde{M}\rangle_{n+1} - \langle\widetilde{M}\rangle_n]\langle\widetilde{M}\rangle_{n+1}^{-1/2}) = d - \mathrm{Tr}(\langle\widetilde{M}\rangle_n\langle\widetilde{M}\rangle_{n+1}^{-1}),$$

$$F_n = f_0 + \cdots + f_n,$$

and assume there exists a constant $a > 1$ such that

$$\sup_n E(([\widetilde{M}_{n+1} - \widetilde{M}_n]^T \langle \widetilde{M} \rangle_{n+1}^{-1} [\widetilde{M}_{n+1} - \widetilde{M}_n])^a | \mathcal{F}_n) < \infty \ a.s.$$

Then, a.s. on $\{F_n \to +\infty\}$,

$$\lim_{n \to \infty} \frac{\widetilde{M}_n^T \langle \widetilde{M} \rangle_n^{-1} \widetilde{M}_n + H_n}{F_n} = 1.$$

Let us first take up briefly the outlines of the proof of Lemma 14. Setting $V_n = \widetilde{M}_n^T \langle \widetilde{M} \rangle_n^{-1} \widetilde{M}_n$, we have

$$\begin{aligned} V_{n+1} &= [\widetilde{M}_n + (\widetilde{M}_{n+1} - \widetilde{M}_n)]^T \langle \widetilde{M} \rangle_{n+1}^{-1} [\widetilde{M}_n + (\widetilde{M}_{n+1} - \widetilde{M}_n)] \\ &= V_n - A_n + B_{n+1} + D_n + D_n^T \end{aligned}$$

with $A_n = \widetilde{M}_n^T (\langle \widetilde{M} \rangle_n^{-1} - \langle \widetilde{M} \rangle_{n+1}^{-1}) \widetilde{M}_n$, $B_{n+1} = (\widetilde{M}_{n+1} - \widetilde{M}_n)^T \langle \widetilde{M} \rangle_{n+1}^{-1} (\widetilde{M}_{n+1} - \widetilde{M}_n)$, and $D_n = (\widetilde{M}_{n+1} - \widetilde{M}_n)^T \langle \widetilde{M} \rangle_{n+1}^{-1} \widetilde{M}_n$. We deduce that

$$V_{n+1} = V_1 - H_n + \sum_{k=1}^{n} B_{k+1} + \sum_{k=1}^{n} \left(D_k^T + D_k \right).$$

Under the moment assumption, $\lim_{n \to \infty} \left(\sum_{k=1}^n B_{k+1} \right) / F_n = 1$ a.s. on $\{F_n \to +\infty\}$. On the other hand,

$$E(|D_{n+1}|^2 | \mathcal{F}_n) = \widetilde{M}_n^T \langle \widetilde{M} \rangle_{n+1}^{-1} [\langle \widetilde{M} \rangle_{n+1} - \langle \widetilde{M} \rangle_n] \langle \widetilde{M} \rangle_{n+1}^{-1} \widetilde{M}_n = O(A_n).$$

Thus the series $\sum_{k=1}^\infty (D_k^T + D_k)$ converges a.s. on $\{H_n < +\infty\}$ and $\sum_{k=1}^n (D_k^T + D_k) = o(H_n)$ a.s. on $\{H_n \to +\infty\}$. It follows that, a.s. on $\{F_n \to +\infty\}$,

$$\lim_{n \to \infty} \frac{V_{n+1} + H_n + o\left(H_n\right) 1_{\{H_n \to +\infty\}}}{F_n} = 1$$

and thus

$$\lim_{n \to \infty} \frac{V_{n+1} + H_n}{F_n} = 1,$$

which concludes the proof of Lemma 14.

We now prove Lemma 10. Let u be any nonzero vector of $\mathbb{R}^d$, and set $W_n = u^T M_n$; (W_n) is a square-integrable martingale, whose increasing process $\langle W \rangle_{n+1} = \Sigma_{k=1}^n \sigma_k^2 \gamma_k^{-2} u^T C_k u = \Sigma_{k=1}^n [\gamma_k v(k)]^{-1} u^T C_k u$ satisfies

$$\langle W \rangle_{n+1} \sim J(n) u^T C u. \tag{24}$$

We have

$$\begin{aligned} [W_{n+1} - W_n]^T \langle W \rangle_{n+1}^{-1} [W_{n+1} - W_n] &= \left[u^T \left(M_{n+1} - M_n \right) \right]^2 \langle W \rangle_{n+1}^{-1} \\ &= \left(\frac{1}{\gamma_n v(n)} \right) \left[u^T \varepsilon_{n+1} \right]^2 \langle W \rangle_{n+1}^{-1}. \end{aligned}$$

Thus,

$$E(([W_{n+1}-W_n]^T\langle W\rangle_{n+1}^{-1}[W_{n+1}-W_n])^{b/2}|\mathcal{F}_n)$$
$$=\left(\frac{1}{\gamma_n v(n)}\right)^{b/2}\langle W\rangle_{n+1}^{-b/2}E(|u^T\varepsilon_{n+1}|^b|\mathcal{F}_n)$$

and, in view of (24),

$$(25)\qquad E(([W_{n+1}-W_n]^T\langle W\rangle_{n+1}^{-1}[W_{n+1}-W_n])^{b/2}|\mathcal{F}_n)=O([\gamma_n v(n)J(n)]^{-b/2}) \\ =O[n^{-b/2}\mathcal{L}(n)].$$

Since $b>2$, (W_n) fulfills the moment assumption of Lemma 14, and we deduce that, a.s. on $\{\Sigma_{k=1}^n(1-\langle W\rangle_k\langle W\rangle_{k+1}^{-1})\to+\infty\}$,

$$(26)\qquad \lim_{n\to\infty}\frac{W_n^T\langle W\rangle_n^{-1}W_n+\sum_{k=1}^n W_k^T[\langle W\rangle_{k-1}^{-1}-\langle W\rangle_k^{-1}]W_k}{\sum_{k=1}^n(1-\langle W\rangle_k\langle W\rangle_{k+1}^{-1})}=1.$$

Now, in view of (24), $(1-\langle W\rangle_k\langle W\rangle_{k+1}^{-1})\sim 1-J(k-1)[J(k)]^{-1}$. Since J' varies regularly with exponent $\alpha-\beta$, we have $J(k-1)[J(k)]^{-1}=1-(\alpha-\beta+1)k^{-1}+o(k^{-1})$; thus

$$(27)\qquad 1-\langle W\rangle_k\langle W\rangle_{k+1}^{-1}\sim\frac{(\alpha-\beta+1)}{k}$$

and

$$(28)\qquad \sum_{k=1}^n(1-\langle W\rangle_k\langle W\rangle_{k+1}^{-1})\sim(\alpha-\beta+1)\ln n.$$

On the other hand, Lemma 9 gives $\|W_n\|^2=O(J(n)\ln(\ln n))$ a.s. Using (24) again, we obtain $\|W_n^T\langle W\rangle_{n+1}^{-1}W_n\|=O(\ln(\ln n))$ and, in view of (28),

$$(29)\qquad W_n^T\langle W\rangle_{n+1}^{-1}W_n=o\left[\sum_{k=1}^n(1-\langle W\rangle_k\langle W\rangle_{k+1}^{-1})\right].$$

Finally,

$$\sum_{k=1}^n W_k^T[\langle W\rangle_{k-1}^{-1}-\langle W\rangle_k^{-1}]W_k=\sum_{k=1}^n\langle W\rangle_{k-1}^{-1}[1-\langle W\rangle_{k-1}\langle W\rangle_k^{-1}]W_kW_k^T$$

and, in view of (24) and (27),

$$(30)\qquad \sum_{k=1}^n W_k^T[\langle W\rangle_{k-1}^{-1}-\langle W\rangle_k^{-1}]W_k\sim\frac{(\alpha-\beta+1)}{u^TCu}\sum_{k=1}^n[kJ(k)]^{-1}u^TM_kM_k^Tu.$$

It follows from the combination of (26), (28), (29), and (30) that

$$(31)\qquad \lim_{n\to\infty}\frac{1}{\ln n}\sum_{k=1}^n[kJ(k)]^{-1}u^TM_kM_k^Tu=u^TCu\text{ a.s.}$$

Let us define Σ_n by

$$\Sigma_n = \frac{1}{\ln n}\left(\sum_{k=1}^{n}[kJ(k)]^{-1}M_k M_k^T\right) - C$$

and denote by $\Sigma_n^{(i,j)}$ the coefficient of the ith line and jth column of Σ_n. Let $(e_1,\ldots,e_d)$ be the canonical basis of $\mathbb{R}^d$. It is easy to see that, for $i,j\in\{1,\ldots,d\}$,

$$\Sigma_n^{(i,j)} = \frac{1}{2}[(e_i+e_j)^T\Sigma_n(e_i+e_j) - e_i^T\Sigma_n e_i - e_j^T\Sigma_n e_j].$$

Applying (31) to the three terms of the right-hand side of this equation, it follows that $\lim_{n\to\infty}\Sigma_n^{(i,j)} = 0$ a.s. and thus $\lim_{n\to\infty}\Sigma_n = 0$ a.s., which completes the proof of Lemma 10.

4.3.3. Proof of Lemma 11. The proof of Lemma 11 is based upon the following result proved by Chaabane [2].

Result 3 (an a.s.CLT for unidimensional martingales). *Let $(\widetilde{M}_n)$ be an unidimensional square integrable martingale with respect to a filtration $\mathcal{F}$ and assume there exists $a>1$ such that*

$$\sum_{n\geq 1} E(([\widetilde{M}_{n+1}-\widetilde{M}_n]^T\langle\widetilde{M}\rangle_{n+1}^{-1}[\widetilde{M}_{n+1}-\widetilde{M}_n])^a|\mathcal{F}_n) < \infty \text{ a.s.}$$

Then,

$$\frac{1}{\ln\langle\widetilde{M}\rangle_n}\sum_{k=1}^{n}\frac{\langle\widetilde{M}\rangle_k-\langle\widetilde{M}\rangle_{k-1}}{\langle\widetilde{M}\rangle_k}\delta_{\langle\widetilde{M}\rangle_k^{-1/2}\widetilde{M}_k}\Longrightarrow\mathcal{N}(0,1) \text{ a.s.}$$

Set $u\in\mathbb{R}^d$, $u\neq 0$, and $W_n = u^T M_n$. In view of (25), (W_n) satisfies the assumption of Result 3. It follows that

$$\frac{1}{\ln\langle W\rangle_n}\sum_{k=1}^{n}[1-\langle W\rangle_{k-1}\langle W\rangle_k^{-1}]\delta_{\langle W\rangle_k^{-1/2}W_k}\Longrightarrow\mathcal{N}(0,1)\text{ a.s.}$$

Then, in view of (24) and (27),

$$\frac{1}{(\alpha-\beta+1)\ln n}\sum_{k=1}^{n}\frac{\alpha-\beta+1}{k}\delta_{[u^TCuJ(k)]^{-1/2}u^TM_k}\Longrightarrow\mathcal{N}\left(0,1\right)\text{ a.s.}$$

and thus

$$\frac{1}{\ln n}\sum_{k=1}^{n}\frac{1}{k}\delta_{[J(k)]^{-1/2}u^TM_k}\Longrightarrow\mathcal{N}\left(0,u^TCu\right)\text{ a.s.}$$

This last property is also clearly satisfied by $u=0$, and thus the proof of Lemma 11 is completed.

Acknowledgments. I gratefully acknowledge Marie Duflo for her invaluable help. I also deeply thank Abdelkader Mokkadem for his thoughtful comments and an anonymous referee whose suggestions have significantly improved the overall presentation.

REFERENCES

[1] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, New York, 1990.

[2] F. Chaabane, *Version forte du théorème de la limite centrale fonctionnel pour les martingales*, C. R. Acad. Sci. Paris Sec. I Math., 323 (1996), pp. 195–198.

[3] H. F. Chen, L. Guo, and A. J. Gao, *Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds*, Stochastic Process. Appl., 27 (1988), pp. 217–231.

[4] D. D. Castelle and M. Duflo, *Probability and Statistics*, Vol. 1, Springer-Verlag, New York, 1986.

[5] B. Delyon and A. Juditsky, *Stochastic optimization with averaging of trajectories*, Stochastics Stochastic Rep., 39 (1992), pp. 107–118.

[6] B. Delyon and A. Juditsky, *Stochastic Approximation with Averaging*, preprint 952, Institut Recherche en Informatique et Systèmes Aléatoires, Rennes, France, 1995.

[7] J. Dippon and J. Renz, *Weighted means of processes in stochastic approximation*, Math. Methods Statist., 5 (1996), pp. 32–60.

[8] M. Duflo, *Méthodes Récursives Aléatoires*, Masson, Paris, 1990. *Random Iterative Models,* Springer-Verlag, New York, 1997 (translation).

[9] M. Duflo, *Algorithmes Stochastiques,* Math. Appl. 23, Springer-Verlag, Berlin, 1996.

[10] M. Duflo, R. Senoussi, and A. Touati, *Sur la loi des grands nombres pour les martingales vectorielles et l'estimateur des moindres carrés d'un modèle de régression*, Ann. Inst. H. Poincaré Probab. Statist., 26 (1990), pp. 549–566.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 3rd ed., Wiley, New York, 1968.

[12] H. J. Kushner and D. S. Clark, *Stochastic approximation for constrained and unconstrained systems*, Appl. Math. Sci. Ser. 26, Springer-Verlag, New York, 1978.

[13] H. J. Kushner and J. Yang, *Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes*, SIAM J. Control Optim., 31 (1993), pp. 1045–1062.

[14] T. Z. Lai and C. Z. Wei, *A note on martingale difference sequences satisfying the local Marcinkiewicz-Zigmund condition*, Bull. Inst. Math. Acad. Sinica, 11 (1983), pp. 1–13.

[15] A. Le Breton, *About the averaging approach schemes for stochastic approximation*, Math. Methods Statist., 2 (1993), pp. 295–315.

[16] A. Le Breton and A. Novikov, *Averaging for estimating covariances in stochastic approximation*, Math. Methods Statist., 3 (1994), pp. 244–266.

[17] A. Le Breton and A. Novikov, *Some results about averaging in stochastic approximation*, Metrika, 42 (1995), pp. 153–171.

[18] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation and Optimization of Random Systems*, Birkhäuser, Basel, 1992.

[19] M. B. Nevelson and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimations*, Nauka, Moscow, 1972. Transl. Math. Monogr. 47, Amer. Math. Soc., Providence, RI, 1973.

[20] M. Pelletier, *Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing*, Ann. Appl. Probab., 8 (1998), pp. 10–44.

[21] M. Pelletier, *On the almost sure asymptotic behaviour of stochastic algorithms,* Stoch. Process. Appl., 78 (1998), pp. 217–244.

[22] M. Pelletier, *An almost sure central limit theorem for stochastic approximation algorithms*, J. Multivariate Anal., 71 (1999), pp. 76–93.

[23] B. T. Polyak, *A new method of stochastic approximation type*, Automat. i Telemekh, 7 (1990), pp. 98–107 (in Russian); Automat. Remote Control, 51 (1990), pp. 937–946 (in English).

[24] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

[25] D. Ruppert, *Stochastic approximation*, in Handbook in Sequential Analysis, B. K. Ghosh and P. K. Sen, eds., Marcel Dekker, New York, 1991, pp. 503–529.

[26] E. Seneta, *Regularly Varying Function*, Lecture Notes in Math. 508, Springer-Verlag, New

York, 1976.

[27] W. F. Stout, *A martingale analogue of Kolmogorov's law of the iterated logarithm*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 15 (1970), pp. 279–290.

[28] A. Touati, *Sur les Versions Fortes du Théorème de la Limite Centrale*, Preprint 23, Université de Marne-la-Vallée, Marne-la-Vallée, France, 1995.

[29] C. Z. Wei, *Asymptotic properties of least-squares estimates in stochastic regression models*, Ann. Statist., 13 (1985), pp. 1498–1508.

[30] C. Z. Wei, *Adaptive prediction by least squares predictors in stochastic regression models with applications to time series*, Ann. Statist., 15 (1987), pp. 1667–1682.

[31] G. Yin, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics Stochastic Rep., 36 (1991), pp. 245–264.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 73–96

SUPERREPLICATION UNDER GAMMA CONSTRAINTS[*]

H. METE SONER[†] AND NIZAR TOUZI[‡]

Abstract. In a financial market consisting of a nonrisky asset and a risky one, we study the minimal initial capital needed in order to superreplicate a given contingent claim under a gamma constraint. This is a constraint on the unbounded variation part of the hedging portfolio. We first consider the case in which the prices are given as general Markov diffusion processes and prove a verification theorem which characterizes the superreplication cost as the unique solution of a quasi-variational inequality. In the context of the Black–Scholes model (i.e., when volatility is constant), this theorem allows us to derive an explicit solution of the problem. These results are based on a new dynamic programming principle for general "stochastic target" problems.

Key words. stochastic control, viscosity solutions, stochastic analysis, superreplication, gamma constraint

AMS subject classifications. 35K55, 49J20, 60H30, 90A09

PII. S0363012998348991

1. Introduction.

We study the problem of superreplicating a contingent claim under a *gamma* constraint. This is a constraint on the unbounded part of the hedging portfolio.

To explain this constraint and the idea of superreplication, let us first consider the classical Black–Scholes framework with one riskless asset which is normalized to $S^0 = 1$ and one risky asset whose price process evolves according to the stochastic differential equation $dS(t)/S(t) = \mu dt + \sigma dW(t)$. Then given a European contingent claim of the type $g(S(T))$, the unconstrained superreplication cost $v^{BS}(0, S(0))$ is defined as the minimal initial capital which allows us to hedge $g(S(T))$ through some portfolio strategy on the assets S^0 and S. It is known that the solution of this problem coincides with the Black–Scholes arbitrage price of $g(S(T))$ and therefore it is given by $v^{BS}(t, s) = E^Q[g(S(T))|S(t) = s]$. Here $E^Q(.)$ is the expectation operator under the equivalent martingale measure, i.e., Q is the probability measure equivalent to P under which the process S is a martingale. Then the optimal hedging strategy consists of holding $\Delta(t, S(t)) := v_s^{BS}(t, S(t))$ units of the risky asset at each time $t \in [0, T]$.

In practice, traders are faced with shortselling, borrowing, or another type of constraint. These restrictions render this optimal strategy impossible to use in practice, and the notion of superreplication is introduced to replace the no-arbitrage price of Black and Scholes, in the presence of such constraints. We refer to Jouini and Kallal (1995) and Cvitanič and Karatzas (1993) for the superreplication problem with general portfolio constraints. They provide a characterization of the minimal superreplication cost as the value of a stochastic optimal control problem. Broadie, Cvitanič, and Soner (1998) observe that, for a contingent claim of the type $g(S(T))$, this con-

[*] Received by the editors December 4, 1998; accepted for publication (in revised form) January 27, 2000; published electronically August 17, 2000.
http://www.siam.org/journals/sicon/39-1/34899.html

[†] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (soner@princeton.edu). The research of this author was partially supported by National Science Foundation grant DMS-98-17525. Part of this work was completed during the author's visit to the Feza Gursey Institute for Basic Sciences in Istanbul.

[‡] CERMSEM, Université Paris I Panthéon-Sorbonne, 106-112 Boulevard de l'Hopital, 75647 Paris Cedex 13, France (touzi@univ-paris1.fr). This paper was written while the author was affiliated with CEREMADE (Université Paris IX Dauphine) and CREST.

trol problem can be explicitly solved by proving that the minimal superreplication cost is the unconstrained Black–Scholes price of a modified claim. For the stochastic volatility model, a similar explicit solution is provided in Cvitanič, Pham, and Touzi (1999).

Another problem which in practice faces traders is the variation of the optimal hedging strategy. The gamma associated to the optimal hedging strategy is defined by $\gamma(t, S(t)) := v_{ss}^{BS}(t, S(t))$ and describes the variation of the holdings in S, in the optimal hedging strategy, with respect to an infinitesimal change of the process S. Since traders act only in discrete-time, a large γ induces an important risk exposure between two transaction dates. This problem was raised by Broadie, Cvitanić, and Soner (1998) who provided an upper bound for the superreplication cost under gamma constraint, as well as the associated hedging strategy. However, they did not formulate a precise statement of the problem.

The chief goals of this paper are first to define the superreplication problem under a gamma constraint and then to obtain an explicit solution.

Formulation of the problem is obtained by observing that the gamma constraint is equivalent to a bound on the variation of the hedging portfolio. We then provide a simple solution to this problem. To describe this solution, let $\hat{g}$ be the smallest function greater than g which satisfies the gamma constraint. Then the minimal superreplicating cost with a gamma constraint solves a variational inequality with terminal condition $\hat{g}$. When the volatility is a given constant, the solution of the problem is given by $E^Q[\hat{g}(S(t))]$, i.e., the Black and Scholes no-arbitrage price of the contingent claim $\hat{g}(S(T))$. We explicitly calculate the $\hat{g}$ function for several standard options such as European calls, puts, and digital options.

Previously, the convex duality argument was used to characterize the minimal superreplicating cost. In this approach, the dual formulation of the problem is obtained by suitable changes of measure. However, in the case of gamma constraints, it seems that the diffusion coefficients need to be modified in order to follow a similar technique. Since this cannot be accomplished by equivalent changes of measure, we were not able to use the convex duality arguments. Instead, we introduce a dynamic programming argument to identify the superreplication cost as the viscosity solution of a differential inequality. To our knowledge, this is the first use of dynamic programming in this context. We believe that this is a powerful tool in analyzing "stochastic target" problems and establishing the connection between the backward-forward stochastic differential equations and viscosity solutions as developed in an accompanying paper by the authors Soner and Touzi (2000).

A technical contribution of this paper is a result on the behavior of double stochastic integrals with respect to Brownian motion. This is needed because our formulation of the problem involves a nonclassical constraint on the unbounded variation part of the portfolio process, which is itself the integrand of the martingale part of the state process.

This paper is organized as follows. Section 2 describes the general problem. We introduce the modified terminal data in section 3 and state the assumptions in section 4. After stating the dynamic programming in section 5, we state and prove the main result in section 6. Section 7 focuses on the constant coefficient (i.e., the Black–Scholes) case, and several examples are discussed in section 8. The remainder of the paper is devoted to technical results: section 9 proves the viscosity property, a comparison result is proved in section 10, and finally a property of stochastic double integrals is proved in section 11.

2. Problem. We consider a financial market which consists of one bank account, with constant price process $S^0(t) = 1$ for all $t \in [0, T]$, and one risky asset with price process evolving according to the following stochastic differential equation:

$$S_{t,s}(t) = s \quad \text{and} \quad \frac{dS_{t,s}(u)}{S_{t,s}(u)} = \mu(u, S_{t,s}(u))dt + \sigma(u, S_{t,s}(u))dW^0(u), \qquad t \le u \le T.$$

Here W^0 is a standard Brownian motion in $\mathbb{R}$ defined on a complete probability space $(\Omega, \mathcal{F}, P^0)$. We shall denote by $\mathbb{F} = \{\mathcal{F}(t),\, 0 \le t \le T\}$ the P^0-augmentation of the filtration generated by W^0. The drift and the volatility functions $s\mu(t, s)$ and $s\sigma(t, s)$ satisfy the usual Lipschitz and linear growth conditions in order for the process $S_{t,s}$ to be well defined; we also assume that $\sigma(t, s) > 0$ for all $(t, s) \in [0, T] \times (0, \infty)$ and

$$E^{P^0}\left[\mathcal{E}\left(-\int_0^T \frac{\mu(t, S_{0,s}(t))}{\sigma(t, S_{0,s}(t))} dW^0(t)\right)\right] = 1,$$

where $E^{P^0}(.)$ is the expectation operator under the probability measure P^0 and $\mathcal{E}(.)$ is the Doléans–Dade exponential martingale, i.e.,

$$\mathcal{E}\left(\int_0^T b(t)dW^0(t)\right) = \exp\left(\int_0^T b(t)dW^0(t) - \frac{1}{2}\int_0^T b^2(t)dt\right).$$

As usual, the assumption that the interest rate of the bank account is zero can be easily dispensed with by appropriate discounting.

Consider now an economic agent, endowed with an initial capital x at time t, who invests at each time $u \in [t, T]$ an amount $Y(u)S(u)$ of his wealth in the risky asset and the remaining wealth in the bank account. The process $Y = \{Y(u),\, t \le u \le T\}$ represents the number of shares of risky asset S held by the agent during the time interval $[t, T]$. Then, by the self-financing condition, the wealth process evolves according to the stochastic differential equation

$$X(t) = x \qquad \text{and} \qquad dX(u) = Y(u)dS(u), \quad t \le u \le T.$$

The purpose of this paper is to introduce constraints on the variations of the hedging portfolio Y. We consider portfolios which are continuous semimartingales with respect to the filtration $\mathbb{F}$. Since $\mathbb{F}$ is the Brownian filtration, we define the controlled portfolio strategy $Y^{\alpha,\gamma}_{t,s,y}$ by

$$\begin{aligned} Y^{\alpha,\gamma}_{t,s,y}(t) &= y, \\ dY^{\alpha,\gamma}_{t,s,y}(u) &= \alpha(u)du + \gamma(u)\frac{dS_{t,s}(u)}{S_{t,s}(u)}, \quad t \le u \le T, \end{aligned} \tag{2.1}$$

where $y \in \mathbb{R}$ is the initial portfolio and the control pair (α, γ) takes values in

$$\mathcal{D}_t := (L^\infty([t, T] \times \Omega;\ \text{Lebesgue} \otimes P^0))^2.$$

Hence a *trading strategy* is defined by the triple (y, α, γ) with $y \in \mathbb{R}$ and $(\alpha, \gamma) \in \mathcal{D}_t$. Then the associated wealth process, denoted by $X^{\alpha,\gamma}_{t,x,s,y}$, satisfies

$$X^{\alpha,\gamma}_{t,x,s,y}(u) = x + \int_t^u Y^{\alpha,\gamma}_{t,s,y}(r)dS_{t,s}(r), \quad t \le u \le T. \tag{2.2}$$

We shall formulate the gamma constraint by requiring that the process γ be bounded from above. Before making this definition precise, we give a formal discussion. Formally, we expect the hedging portfolio to satisfy

$$Y(u) = v_s(u, S_{t,s}(u)),$$

where v is the minimal superreplication cost. Indeed, this is true in the classical Black–Scholes theory as well as in the case of portfolio constraints; see Broadie, Cvitanič, and Soner (1998). Assuming enough regularity, we apply the Itô formula. The result is

$$dY(u) = A(u)du + \sigma(u, S_{t,s}(u))S_{t,s}(u)v_{ss}(u, S_{t,s}(u))dW^0(u),$$

where $A(u)$ is given in terms of derivatives of v. Compare this equation with (2.1) to conclude that

$$\gamma(u) = S_{t,s}(u)\ v_{ss}(u, S_{t,s}(u)).$$

Therefore a bound on the process γ translates to a bound on sv_{ss}. Notice that, by changing the definition of the process γ in (2.1), we may bound v_{ss} instead of sv_{ss}. However, we choose to study sv_{ss} because it is a dimensionless quantity, i.e., if all the parameters in the problem are increased by the same factor, sv_{ss} remains unchanged.

We now formulate the gamma constraint in the following way. Let Γ be a constant fixed throughout the paper. Given some initial capital $x > 0$, a trading strategy (y, α, γ) is said to be x-admissible if it satisfies the gamma constraint $\gamma(u) \leq \Gamma$ for all $t \leq u \leq T$ almost surely (a.s.) and the associated wealth process $X^{\alpha,\gamma}_{t,x,s,y}$ is nonnegative. We shall denote by

$$\mathcal{A}_{t,s}(x) := \left\{(y, \alpha, \gamma) \in \mathbb{R} \times \mathcal{D}_t : \gamma(.) \leq \Gamma \text{ and } X^{\alpha,\gamma}_{t,x,s,y}(.) \geq 0\right\}$$

the set of all admissible trading strategies.

We consider a European-type contingent claim $g(S_{t,s}(T))$ defined by the terminal payoff function g. Given such a contingent claim, we then consider the infimum $v(t, s)$ of initial capitals x which induce a wealth process $X^{\alpha,\gamma}_{t,x,s,y}$ through some admissible trading strategy (y, α, γ) such that $X^{\alpha,\gamma}_{t,x,s,y}$ hedges $g(S_{t,s}(T))$, i.e.,

$$(2.3) \qquad v(t, s) = \inf\left\{x : \exists\ (y, \alpha, \gamma) \in \mathcal{A}_{t,s}(x),\ \ X^{\alpha,\gamma}_{t,x,s,y}(T) \geq g(S_{t,s}(T)) \text{ a.s.}\right\}.$$

Note that if g is convex so is v in the s-variable; hence, in this case, gamma is bounded from below as well.

Our goal is to prove that function $v(t, s)$ solves a variational inequality and that its terminal value is given by some function $\hat{g}$ dominating g. When we focus on the constant volatility case, these observations allow us to derive an explicit solution of the hedging problem (2.3): $v(t, s)$ is the (unconstrained) Black and Scholes price of the modified contingent claim $\hat{g}(S_{t,s}(T))$. This function $\hat{g}$ can be easily computed and several examples are provided in section 7.

Throughout this paper, we shall introduce a probability measure $P \sim P^0$ defined by

$$P(A) = E^{P^0}\left[1_A \mathcal{E}\left(-\int_0^T \frac{\mu(t, S_{0,s}(t))}{\sigma(t, S_{0,s}(t))} dW^0(t)\right)\right] \qquad \text{for all } A \in \mathcal{F}.$$

We shall denote by $E(.)$ the expectation operator under the probability measure P. By Girsanov's theorem, the process W defined by

$$W(u) := W^0(u) + \int_t^u \frac{\mu(r, S_{t,s}(r))}{\sigma(r, S_{t,s}(r))} dr, \qquad t \le u \le T,$$

is a Brownian motion under P. In terms of the Brownian motion W, the risky asset price process is defined by

$$\text{(2.4)} \qquad S_{t,s}(t) = s \qquad \text{and} \qquad \frac{dS_{t,s}(u)}{S_{t,s}(u)} = \sigma(u, S_{t,s}(u))dW(u), \qquad t \le u \le T.$$

3. Modified terminal data. Due to the constraint, the limit of the value function $v(t,s)$ of (2.3), as t tends to the terminal time T, may not be equal to the contingent claim g. Indeed the determination of this limit is an important step toward the solution of the problem.

We will show in the following sections that the following function $\hat{g}$ is equal to the limit

$$\hat{g}(s) := h^{\mathrm{conc}}(s) + \Gamma s \ln(s), \quad s > 0,$$

where $h(s) = g(s) - \Gamma s \ln(s)$ and h^{conc} is the concave envelope of h, i.e., the smallest concave function greater than h. In other words, function $\hat{g}(s) - \Gamma s \ln(s)$ is the concave envelope of function $g(s) - \Gamma s \ln(s)$. The chief property of $\hat{g}$ that we will use is the following.

LEMMA 3.1. *$\hat{g}$ is the smallest function satisfying the following two conditions:* (i) *$\hat{g} \ge g$ and* (ii) *$\hat{g}(s) - \Gamma s \ln(s)$ is concave.*

Proof. Clearly, $\hat{g}$ satisfies these conditions. Let u be another function satisfying both of them. Set $w(s) := u(s) - \Gamma s \ln(s)$. Clearly $w \ge h$. Since w is concave and $w \ge h$, by the definition of the concave envelope of h, $w \ge h^{\mathrm{conc}}$. Therefore,

$$u(s) = w(s) + \Gamma s \ln(s) \ge h^{\mathrm{conc}}(s) + \Gamma s \ln(s) = \hat{g}(s). \qquad \square$$

In section 6 below, we will show that the terminal data of the minimal superreplicating cost are equal to $\hat{g}$. The formal reason for this is that if v is sufficiently smooth, we formally expect $v(t,s)$ to satisfy the gamma constraint $sv_{ss}(t,s) \le \Gamma$. This is equivalent to the statement that $v(t,s) + \Gamma s \ln(s)$ is concave. Therefore we formally expect the terminal data $\lim_{t \uparrow T} v(T,s)$ to be the smallest function satisfying the two conditions of the previous lemma.

4. Assumptions. We always assume that

$$\text{(4.1)} \qquad \text{function } g \text{ is nonnegative and lower semicontinuous.}$$

We start with several assumptions on the payoff function g and $\hat{g}$. In sections 7 and 8 below, we will verify that all these assumptions are satisfied by standard claims in the Black–Scholes model, i.e., in the case of a constant volatility function $\sigma(t,s)$; see section 7 below.

Assumption 4.1. We assume that $\hat{g}$ is finite and there exists a nonnegative, strictly concave function $\phi \in C^2$ with $\lim_{s \to \infty} \phi(s) = \infty$ such that

$$\text{(4.2)} \qquad \limsup_{s \to \infty} \frac{\hat{g}^{\mathrm{conc}}(s) - E\left[\hat{g}^{\mathrm{conc}}(S_{t,s}(T))\right]}{\phi(s)} < \infty.$$

Remark 4.2. Any function g which is growing at most linearly at infinity satisfies $\hat{g}^{\rm conc}(.) < \infty$. Indeed in this case, $h(s) \le H(s) := K - \Gamma s \ln(s)/2$ for some constant K. Since H is concave, $h^{\rm conc} \le H$, and therefore $\hat{g}$ is finite.

The main use of Assumption 4.1 is to prove a comparison result. The statement and the proof of this result are given in section 10.

Our final assumption is the existence of a smooth solution to the variational inequality

$$\min\left\{-\mathcal{L}u;\ \Gamma - su_{ss}\right\}(t,s) = 0 \quad \text{on } [0,T)\times(0,\infty) \tag{4.3}$$

together with the terminal condition

$$u(T,s) = \hat{g}(s) \qquad \text{for all } s > 0, \tag{4.4}$$

where $\mathcal{L}$ is the parabolic operator related to the infinitesimal generator of the stock price process,

$$\mathcal{L} := \frac{\partial}{\partial t} + \frac{1}{2}\sigma^2(t,s)s^2\frac{\partial^2}{\partial s^2}.$$

We will prove in section 6 that this solution is equal to the minimal superreplication cost.

Assumption 4.3. The variational inequality (4.3)–(4.4) has a $C^{1,2}\left([0,T),(0,\infty)\right)$ solution $\hat{v}$ satisfying

(i) $\hat{v}(t,0) = \hat{g}(0)$ for all $t \in [0,T]$,

(ii) $\hat{v}$ is polynomially growing in its s variable at infinity,

(iii) $s\hat{v}_{ss}$, $\mathcal{L}\hat{v}$ are bounded,

(iv) v_s is a $W^{1,2}$ function with generalized derivatives satisfying $\mathcal{L}\hat{v}_s$ bounded.

In section 7, for the constant volatility model, we verify this assumption by providing an explicit solution.

Remark 4.4. By a classical comparison theorem for the equation $\mathcal{L}v = 0$ (see, for instance, Friedman (1964)), we see that $\hat{v}(t,s) \ge E[\hat{g}(S_{t,s}(T))]$. Since g is nonnegative, so is $\hat{g}$; therefore we have

$$\hat{v}(t,s) \ge E\left[\hat{g}(S_{t,s}(T))\right] \ge 0 \qquad \text{for all } (t,s) \in [0,T]\times(0,\infty).$$

5. Dynamic programming. The following is the analogue of the principle of dynamic programming which is standard in the theory of stochastic optimal control theory first proved by R. Bellman.

LEMMA 5.1. *Let $(t,s) \in [0,T)\times[0,\infty)$ and consider an arbitrary stopping time θ valued in $[t,T]$. Suppose that $X^{\alpha,\gamma}_{t,x,s,y}(T) \ge g(S_{t,s}(T))$ P-a.s. for some $(\alpha,\gamma) \in \mathcal{A}_{t,s}(x)$, $y \in \mathbb{R}$, and initial wealth $x \in \mathbb{R}$. Then, for the value function v of* (2.3)*, we have*

$$X^{\alpha,\gamma}_{t,x,s,y}(\theta) \ge v\left(\theta, S_{t,s}(\theta)\right), \quad P\text{-a.s.}$$

Proof. Let x, y, θ, and (α,γ) be as in the above statement. Set $\hat{x} = X^{\alpha,\gamma}_{t,x,s,y}(\theta)$, $\hat{s} = S_{t,s}(\theta)$, $\hat{y} = Y^{\alpha,\gamma}_{t,s,y}(\theta)$. Clearly $Y_{t,s,y} = Y_{\theta,\hat{s},\hat{y}}$. By definition of the wealth process (2.2), this provides

$$X^{\alpha,\gamma}_{t,x,s,y}(T) = X^{\alpha,\gamma}_{\theta,\hat{x},\hat{s},\hat{y}}(T).$$

Also, by uniqueness of the solution for the stochastic differential equation defining the stock price S, we have $S_{t,s} = S_{\theta,\hat{s}}$. Since $X^{\alpha,\gamma}_{\theta,\hat{x},\hat{s},\hat{y}}(T) = X^{\alpha,\gamma}_{t,x,s,y}(T) \ge g\left(S_{t,s}(T)\right)$

$= g\left(S_{\theta,\hat{s}}(T)\right)$, it follows that $\hat{x} \geq v(\theta, \hat{s})$ by definition of the control problem $v(\theta, \hat{s})$. □

Remark 5.2. As in optimal control theory, the second part of the dynamic programming is also available. A systematic study of dynamic programming is given in an accompanying paper by the authors. Since we do not need the second part of the dynamic programming in this paper, we refer the reader to Soner and Touzi (2000) for a discussion of the full dynamic programming.

6. Main result. Let $\hat{v}$ be the solution of the variational inequality (4.3)–(4.4) introduced in Assumption 4.3.

THEOREM 6.1. *Let Assumptions* 4.1 *and* 4.3 *hold. Then, the value function* v *of the hedging problem* (2.3) *is equal to the unique smooth solution of the variational inequality* (4.3)–(4.4), *i.e.,*

$$v = \hat{v}.$$

Notice that the variational inequality (4.3)–(4.4) was not assumed to have a unique solution satisfying the requirement of Assumption 4.3. Uniqueness is obtained as a consequence of the above theorem.

Let v_* be the lower semicontinuous envelope of v:

$$v_*(t,s) := \liminf_{(t',s')\to(t,s)} v(t,s). \tag{6.1}$$

We prove the theorem after assuming two properties of the value function v.

P1. Function $s \longmapsto v_*(t,s) - \Gamma s \ln(s)$ is concave for all $t \in [0,T)$.

P2. v_* is a viscosity supersolution of the equation $-\mathcal{L}u = 0$ on $[0,T) \times (0,\infty)$.

These properties will be verified in section 9 below.

Proof. We start with the inequality $v \leq \hat{v}$. For $t \leq u \leq T$, set

$$y = \hat{v}_s(t,s), \qquad \alpha(u) = \mathcal{L}\hat{v}_s(u, S(u)), \qquad \gamma(u) = S_{t,s}(u)\hat{v}_{ss}(u, S(u)).$$

Since $\mathcal{L}\hat{v} \leq 0$,

$$\begin{aligned} g\left(S_{t,s}(T)\right) &\leq \hat{g}\left(S_{t,s}(T)\right) = \hat{v}\left(T, S_{t,s}(T)\right) \\ &= \hat{v}(t,s) + \int_t^T \mathcal{L}\hat{v}(u, S_{t,s}(u))du + \hat{v}_s(u, S_{t,s}(u))dS_{t,s}(u) \\ &\leq \hat{v}(t,s) + \int_t^T Y_{t,y}^{\alpha,\gamma}(u)dS_{t,s}(u); \end{aligned}$$

in the last step we applied the generalized Itô's formula to $v_s \in W^{1,2}$. (See Krylov (1980, Theorem 1, p. 122) for Itô's formula with generalized derivatives.) By Assumption 4.3, $(\alpha,\gamma) \in \mathcal{D}_t$. Furthermore, since $\hat{v}$ solves the variational inequality (4.3), $\gamma(u) \leq \Gamma$ for all $u \in [t,T]$. By Remark 4.4, $\hat{v}(u, S_{t,s}(u)) = X_{t,x,s,y}^{\alpha,\gamma}(u) \geq 0$ with $x = \hat{v}(t,s)$. Hence $(y,\alpha,\gamma) \in \mathcal{A}_{t,s}(\hat{v}(t,s))$, and by the definition of the minimal replicating price, we conclude that $v \leq \hat{v}$.

We now prove the reverse inequality. Fix $(t,s) \in (0,T) \times (0,\infty)$, and $\delta > 0$. By the definition of v, there exist an initial wealth $x \in [v(t,s), v(t,s)+\delta)$ and a trading strategy $(y^t, \alpha^t, \gamma^t) \in \mathcal{A}_{t,s}(x)$ satisfying

$$X_{t,x,s,y^t}^{\alpha^t,\gamma^t}(T) \geq g\left(S_{t,s}(T)\right) \quad P\text{-a.s.}$$

Therefore,

$$\delta + v(t,s) + \int_t^T Y_{t,s,y^t}^{\alpha^t,\gamma^t}(u)dS_{t,s}(u) \geq g\left(S_{t,s}(T)\right) \qquad P\text{-a.s.}$$

By the definition of $\mathcal{A}_{t,s}(x)$, the local martingale $\{\int_t^u Y_{t,s,y^t}^{\alpha^t,\gamma^t}(r)dS_{t,s}(r),\ u \geq t\}$ is bounded from below and is therefore a supermartingale. We take the expected value in the last inequality and then use this fact. The result is

$$\delta + v(t,s) \geq E\left[g(S_{t,s}(T))\right].$$

Since $\delta > 0$ is arbitrary and g is lower semicontinuous, Fatou's lemma yields

$$v_*(T,s) = \liminf_{(t,s')\to(T,s)} v(t,s') \geq g(s) \qquad \text{for all } s > 0.$$

In view of the property P1, $v_*(T,\cdot)$ satisfies both conditions stated in Lemma 3.1, and therefore $v_*(T,s) \geq \hat{g}(s)$.

By dynamic programming, for any $(y,\alpha,\gamma) \in \mathcal{A}_{t,s}(x)$ satisfying $X_{t,x,s,y}^{\alpha,\gamma}(T) \geq g(S_{t,s}(T))$,

$$X_{t,x,s,y}^{\alpha,\gamma}(u) \geq v(u, S_{t,s}(u)) \qquad \text{for all } u \in [t,T].$$

Since we have shown that $v_*(T,s) \geq \hat{g}(s)$, by taking the limit as u tends to T, we conclude that

$$X_{t,x,s,y}^{\alpha,\gamma}(T) \geq \hat{g}(S_{t,s}(T)).$$

Therefore, any strategy that dominates g also dominates $\hat{g}$. Since $\hat{g} \geq g$, this provides

$$v(t,s) = \inf\left\{x : \exists\ (y,\alpha,\gamma) \in \mathcal{A}_{t,s}(x),\ \ X_{t,x,s,y}^{\alpha,\gamma}(T) \geq \hat{g}(S_{t,s}(T))\ \text{ a.s.}\right\}, \tag{6.2}$$

i.e., v is the minimal superreplication cost for the claim $\hat{g}$. By definition, the Black–Scholes price (i.e., unconstrained superreplication cost) is always smaller than the superreplication cost with gamma constraint,

$$v(t,s) \geq E\left[\hat{g}(S_{t,s}(T))\right] \qquad \text{for all } (t,s) \in [0,T) \times (0,\infty). \tag{6.3}$$

Moreover, by (6.2), $v(t,0+) = \hat{g}(0)$ for all $t \in [0,T)$. Therefore, $v_*(t,0) \leq v(t,0) = \hat{g}(0)$. Also (6.3) together with Fatou's lemma yield $v_*(t,0) \geq \hat{g}(0)$. Hence $v_*(t,0) = \hat{g}(0)$.

In view of Lemma 9.2 below, v_* is a lower semicontinuous viscosity supersolution of (4.3)–(4.4). By Theorem 10.1, $v_* \geq \hat{v}$. This completes the proof of the theorem since $v \geq v_*$. □

Remark 6.2. In the first part of the above proof, the optimal hedging strategy (y,α,γ) is expressed explicitly in terms of the derivatives of the minimal superreplication cost function $\hat{v}$.

Remark 6.3. In the proof above, it is shown (without appealing to Theorem 10.1) that the (unconstrained) Black and Scholes price of $\hat{g}(S_{t,s}(T))$ is a trivial lower bound for v

$$v(t,s) \geq E\left[\hat{g}(S_{t,s}(T))\right] \qquad \text{for all } (t,s) \in [0,T) \times (0,\infty).$$

We shall use this lower bound in the proof of the comparison Theorem 10.1.

7. The Black and Scholes model.

In this section, we focus on a discussion of the Black and Scholes model in which the volatility function $\sigma(t,s)$ is constant, i.e., $\sigma(t,s) = \sigma$ for all $(t,s) \in [0,T] \times (0,\infty)$.

We shall provide an explicit solution to the hedging problem (2.3) under the following condition.

Assumption 7.1. Function $s \longmapsto h^{\text{conc}}(s) - Cs\ln(s)$ is convex for some constant C.

Remark 7.2. Suppose that function g is such that $s \longmapsto g(s) + As\ln(s)$ is convex for some constant A. Then, since $h(s) = g(s) + As\ln(s) - (\Gamma + A)s\ln(s)$, it follows from the construction of the concave envelope that Assumption 7.1 is satisfied by $C = \Gamma + A$.

THEOREM 7.3. *Let Assumptions* 4.1 *and* 7.1 *hold. Then, Assumption* 4.3 *holds and the value function v of the hedging problem* (2.3) *is simply the unconstrained Black and Scholes price $\hat{v}$ of the contingent claim $\hat{g}(S_{t,s}(T))$, i.e.,*

$$v(t,s) = \hat{v}(t,s) = E\left[\hat{g}(S_{t,s}(T))\right] \quad \textit{for all } (t,s) \in [0,T] \times (0,\infty).$$

Proof. Denote $\tilde{v}(t,s) := E\left[\hat{g}(S_{t,s}(T))\right]$. Then $\tilde{v}$ is a classical solution to the equation

$$-\mathcal{L}u = 0 \quad \text{on } [0,T) \times (0,\infty) \qquad \text{and} \qquad u(T,s) = \hat{g}(s), \quad s > 0.$$

Furthermore, by the definition of $\hat{g}$,

$$\tilde{v}(t,s) - \Gamma s\ln(s) = E\left[h^{\text{conc}}(S_{t,s}(T))\right] + \frac{1}{2}\sigma^2(T-t)\Gamma s. \tag{7.1}$$

Since h^{conc} is concave and $S_{t,s}(T)$ is linear in s, this proves that for all $t \in [0,T]$, function $s \to \tilde{v}(t,s) - \Gamma s\ln(s)$ is concave, and therefore $s\tilde{v}_{ss}(t,s) \le \Gamma$ for all $(t,s) \in [0,T) \times (0,\infty)$. A similar argument using Assumption 7.1 shows that $s\tilde{v}_{ss}(t,s) \ge C$.

Consequently $\tilde{v} = \hat{v}$ is a classical solution of the variational inequality (4.3)–(4.4). By Friedman (1964, Theorem 10, p. 72), function $\hat{v}_s$ is $C^{1,2}$, which provides all the regularity required in Assumption 4.3, except the property (iii). To verify Assumption 4.3 (iii), we differentiate the equation $\mathcal{L}\hat{v} = \hat{v}_t(t,s) + \sigma^2 s^2 \hat{v}_{ss}(t,s) = 0$ to obtain $\mathcal{L}\hat{v}_s(t,s) = \sigma^2 s\hat{v}_{ss}(t,s)$. Since we have already proved that $s\hat{v}_{ss}$ is bounded, so is $\mathcal{L}\hat{v}$. □

Remark 7.4. Observe that Assumption 4.1 is only used in the proof of the comparison Theorem 10.1 which is needed to show that $\hat{v} \le v$. Since in the Black and Scholes case $\hat{v}(t,s) = E[\hat{g}(S_{t,s}(T))]$, the variational inequality (4.3) reduces to the linear equation $-\mathcal{L}v = 0$. Then we can appeal to the standard comparison theorem for this equation, and Assumption 4.1 can be relaxed by requiring only that $\hat{g}(.) < \infty$.

8. Examples.

European call option. Let $g(s) = (s-K)^+$, $s > 0$. Since g is convex, Assumption 7.1 is satisfied; see Remark 7.2. Next, it is easily checked that the concave envelope of function $h(s) = (s-K)^+ - \Gamma s\ln(s)$ is given by

$$h^{\text{conc}}(s) = \begin{cases} h(s), & s \in (0,\infty) \setminus [s_1,s_2], \\ h(s_1) + h'(s_1)(s-s_1), & s \in [s_1,s_2], \end{cases}$$

i.e., h^{conc} coincides with h outside the interval $[s_1,s_2]$ and is defined by a straight line in $[s_1,s_2]$. The values s_1 and s_2 are characterized by

$$s_1 < K < s_2\, h'(s_1) = h'(s_2) \qquad \text{and} \qquad h(s_2) = h(s_1) + h'(s_1)(s_2 - s_1).$$

A direct calculation yields

$$s_1 = \frac{K}{\Gamma(e^{1/\Gamma} - 1)} \qquad \text{and} \qquad s_2 = \frac{Ke^{1/\Gamma}}{\Gamma(e^{1/\Gamma} - 1)}.$$

Therefore,

$$\hat{g}(s) = \begin{cases} (s-K)^+, & s \in (0,\infty) \setminus [s_1, s_2], \\ \Gamma\left(s \ln \frac{s}{s_1} + s_1 - s\right), & s \in [s_1, s_2]. \end{cases}$$

Since $\hat{g}^{\text{conc}}(s) = s$ for all $s > 0$, Assumption 4.1 is clearly satisfied and Theorem 7.3 applies.

European put option. We now consider the case $g(s) = (K-s)^+$, $s > 0$. As in the previous example, g is convex, and therefore Assumption 7.1 is satisfied. The concave envelope of function $h(s) = (K-s)^+ - \Gamma s \ln(s)$ is given by

$$h^{\text{conc}}(s) = \begin{cases} h(s), & s \in (0,\infty) \setminus [s_1, s_2], \\ h(s_1) + h'(s_1)(s - s_1), & s \in [s_1, s_2], \end{cases}$$

i.e., h^{conc} coincides with h outside the interval $[s_1, s_2]$ and is defined by a straight line in $[s_1, s_2]$. The values s_1 and s_2 are characterized by

$$s_1 < K < s_2 \, h'(s_1) = h'(s_2) \qquad \text{and} \qquad h(s_2) = h(s_1) + h'(s_1)(s_2 - s_1).$$

We directly calculate that

$$s_1 = \frac{K}{\Gamma(e^{1/\Gamma} - 1)} \qquad \text{and} \qquad s_2 = \frac{Ke^{1/\Gamma}}{\Gamma(e^{1/\Gamma} - 1)}$$

(the same values as in the first example) and

$$\hat{g}(s) = \begin{cases} (K-s)^+, & s \in (0,\infty) \setminus [s_1, s_2], \\ K - s + \Gamma\left(s \ln \frac{s}{s_1} + s_1 - s\right), & s \in [s_1, s_2]. \end{cases}$$

Since $\hat{g}$ is bounded, Assumption 4.1 holds and therefore Theorem 7.3 applies.

Straddle option. We now study the contingent claim defined by $g(s) = (s-K)^+ + (K-s)^+$, $s > 0$. The same argument as in the previous examples yields

$$\hat{g}(s) = \begin{cases} (s-K)^+ + (K-s)^+, & s \in (0,\infty) \setminus [s_1, s_2], \\ K - s + \Gamma\left(s \ln \frac{s}{s_1} + s_1 - s\right), & s \in [s_1, s_2], \end{cases}$$

where $s_1 = \frac{2K}{\Gamma(e^{2/\Gamma}-1)}$ and $s_2 = s_1 e^{2/\Gamma}$.

Digital option. Our last example is the contingent claim defined by $g(s) = 1_{\{s>K\}}$, $s > 0$. Then, it is easily seen that the concave envelope of function $h(s) = 1_{s>K} - \Gamma s \ln(s)$ is given by

$$h^{\text{conc}}(s) = \begin{cases} h(s), & s \in (0,\infty) \setminus [s^*, K], \\ h(s^*) + h'(s^*)(s - s^*), & s^* \leq s \leq K, \end{cases}$$

where s^* is the unique solution of

$$0 < s^* < \Gamma \qquad \text{and} \qquad s^* - K\ln(s^*) = K - K\ln(K) + \frac{1}{\Gamma}.$$

Clearly, the above function satisfies Assumption 7.1. This provides the candidate for the hedging problem under the gamma constraint:

$$\hat{g}(s) = \begin{cases} 0, & s \ \leq s^*, \\ \Gamma s \ln(s) + h(s^*) + h'(s^*)(s - s^*), & s^* \leq s \leq K, \\ 1, & s \ \geq K. \end{cases}$$

Since $\hat{g} \leq 1$, we have $\hat{g}^{\text{conc}} \leq 1$ and Assumption 4.1 holds. Then, Theorem 7.3 again applies.

9. Viscosity property. In this section, we prove properties P1 and P2 of section 6.

THEOREM 9.1. *v_* is a viscosity supersolution of the variational inequality*

$$\min\{-\mathcal{L}u(t,s),\ \Gamma - su_{ss}(t,s)\} = 0 \tag{9.1}$$

on $(0,T) \times (0,\infty)$.

Proof. For $\varepsilon \in (0,1]$, set

$$\mathcal{A}^{\varepsilon}_{t,s}(x) := \left\{(y,\alpha,\gamma) \in \mathcal{A}_{t,s}(x) : |\alpha(.)| + |\gamma(.)| \leq \varepsilon^{-1}\right\},$$

and

$$v^{\varepsilon}(t,s) = \inf\left\{x : \exists\ (y,\alpha,\gamma) \in \mathcal{A}^{\varepsilon}_{t,s}(x),\ \ X^{\alpha,\gamma}_{t,x,s,y}(T) \geq g(S_{t,s}(T))\ \text{ a.s.}\right\}.$$

Let v^{ε}_* be the lower semicontinuous envelope of v^{ε}; cf. (6.1). It is clear that v^{ε} also satisfies the dynamic programming equation of Lemma 5.1.

First we will show that v^{ε}_* is a viscosity supersolution of (9.1). Let $\varphi \in C^{\infty}(\mathbb{R}^2)$ and $(t_0,s_0) \in (0,T) \times (0,\infty)$ satisfy

$$(v^{\varepsilon}_* - \varphi)(t_0,s_0) = \min_{(t,s)\in(0,T)\times(0,\infty)} (v^{\varepsilon}_* - \varphi)(t,s).$$

We need to show that

$$-\mathcal{L}\varphi(t_0,s_0) \geq 0 \qquad \text{and} \qquad s_0\varphi_{ss}(t_0,s_0) \leq \Gamma. \tag{9.2}$$

We may assume that $(v^{\varepsilon}_* - \varphi)(t_0,s_0) = 0$ so that $v^{\varepsilon}_* \geq \varphi$.

Choose $(t_n,s_n) \to (t_0,s_0)$ so that $v^{\varepsilon}(t_n,s_n)$ converges to $v^{\varepsilon}_*(t_0,s_0)$. For each n, by the definition of v^{ε} and the dynamic programming, there are $x_n \in [v^{\varepsilon}(t_n,s_n), v^{\varepsilon}(t_n,s_n)+1/n]$ hedging strategies $(y_n,\alpha_n,\gamma_n) \in \mathcal{A}^{\varepsilon}_{t_n,s_n}(x_n)$ satisfying

$$X^{\alpha_n,\gamma_n}_{t_n,x_n,s_n,y_n}(t_n+t) - v^{\varepsilon}(t_n+t, S_{t_n,s_n}(t_n+t)) \geq 0$$

for every $t > 0$. Since $v^{\varepsilon} \geq v^{\varepsilon}_* \geq \varphi$,

$$x_n + \int_{t_n}^{t_n+t} Y^{\alpha_n,\gamma_n}_{t_n,s_n,y_n}(u)dS_{t_n,s_n}(u) - \varphi(t_n+t, S_{t_n,s_n}(t_n+t)) \geq 0.$$

Set

$$\beta_n := x_n - \varphi(t_n,s_n)$$

and observe that $\beta_n \to 0$ as $n \to \infty$, since $\varphi(t_n,s_n) \to \varphi(t_0,s_0) = v^{\varepsilon}_*(t_0,s_0)$, $|x_n - v^{\varepsilon}(t_n,s_n)| \leq 1/n$, and $v^{\varepsilon}(t_n,s_n) \longrightarrow v^{\varepsilon}_*(t_0,s_0)$.

By Itô's lemma,

$$M_n(t) \le D_n(t) + \beta_n \tag{9.3}$$

for every $t \ge 0$, where

$$\begin{aligned} M_n(t) &= \int_0^t \left[\varphi_s(t_n+u, S_{t_n,s_n}(t_n+u)) - Y^{\alpha_n,\gamma_n}_{t_n,s_n,y_n}(t_n+u)\right] dS_{t_n,s_n}(t_n+u),\\ D_n(t) &= -\int_0^t \mathcal{L}\varphi(t_n+u, S_{t_n,s_n}(t_n+u))du. \end{aligned}$$

For some sufficiently large positive constant λ, define the stopping time $t_n+\theta_n$ by

$$\theta_n := \inf\left\{u > 0 : \left|\ln\left(S_{t_n,s_n}(t_n+u)/s_n\right)\right| \ge \lambda\right\}$$

and observe that the sequence of stopping times (θ_n) satisfies

$$\liminf_{n\to\infty} t \wedge \theta_n \ge \frac{1}{2} t \wedge \theta_0 \qquad P\text{-a.s.}$$

for all $t > 0$; see Remark 11.2. By the smoothness of $\mathcal{L}\varphi$, the integrand in the definition of M_n is bounded up to the stopping time θ_n and therefore, taking the expectation in (9.3) provides

$$-E\left[\int_0^{t\wedge\theta_n} \mathcal{L}\varphi(t_n+u, S_{t_n,s_n}(t_n+u))du\right] \ge -\beta_n.$$

By sending n to infinity, we obtain

$$-E\left[\int_0^{t\wedge\theta_0} \mathcal{L}\varphi(t_0+u, S_{t_0,s_0}(t_0+u))du\right] \ge 0$$

by dominated convergence and continuity of $\mathcal{L}\varphi$. Then, dividing by t and taking the limit as $t \searrow 0$, we get by dominated convergence

$$-\mathcal{L}\varphi(t_0, s_0) \ge 0,$$

which is the first part of (9.2). It remains to prove the second inequality.

By another application of Itô's lemma, it follows that

$$M_n(t) = \int_0^t \left(z_n + \int_0^u a_n(r)dr + \int_0^u b_n(r)dS_{t_n,s_n}(t_n+r)\right) dS_{t_n,s_n}(t_n+u),$$

where

$$\begin{aligned} z_n &= \varphi_s(t_n, s_n) - y_n,\\ a_n(r) &= \mathcal{L}\varphi_s(t_n+r, S_{t_n,s_n}(t_n+r)) - \alpha_n(t_n+r),\\ b_n(r) &= \varphi_{ss}(t_n+r, S_{t_n,s_n}(t_n+r)) - \frac{\gamma_n(t_n+r)}{S_{t_n,s_n}(t_n+r)}. \end{aligned}$$

Observe that the processes $a_n(.\wedge\theta_n)$ and $b_n(.\wedge\theta_n)$ are bounded uniformly in n since $\mathcal{L}\varphi_s$ and φ_{ss} are smooth functions. By (9.3),

$$M_n(t\wedge\theta_n) \le D_n(t\wedge\theta_n) + \beta_n \le Ct\wedge\theta_n + \beta_n$$

for some positive constant C. We now apply the results of Propositions 11.5 and 11.6 to the martingales M_n. The result is

$$\lim_{n\to\infty} y_n = \varphi_s(t_0, y_0) \qquad \text{and} \qquad \liminf_{n\to\infty,\, t\searrow 0} b(t) \le 0,$$

where b is the L^2 weak limit of the sequence (b_n). The remaining inequality in (9.2) is obtained after recalling that $\gamma_n(t) \le \Gamma$.

Hence v_*^ε is a viscosity supersolution of (9.1). Since

$$v_*(t,s) = \liminf_* v^\varepsilon(t,s) = \liminf_{\varepsilon\to 0,(t',s')\to(t,s)} v_*^\varepsilon(t',s'),$$

the Barles–Perthame technique implies that v_* is a viscosity supersolution of (9.1) as well. □

The following result completes the proof of the properties P1 and P2 of section 6.

LEMMA 9.2. *Let f be a lower semicontinuous function defined on $(0,\infty)$. Then, f is a viscosity supersolution of $\Gamma - sf_{ss}(s) \ge 0$ if and only if $f(s) - \Gamma s \ln(s)$ is concave.*

Proof. Suppose that $h(s) := f(s) - \Gamma s \ln(s)$ is a concave function and a smooth test function φ and $s_0 > 0$ satisfy

$$0 = (f-\varphi)(s_0) = \min \{\ (f-\varphi)(s)\ :\ s \ge 0\ \}.$$

Set $\psi(s) := \varphi(s) - \Gamma s \ln(s)$, so that for any $\delta > 0$,

$$\psi(s_0+\delta) + \psi(s_0-\delta) - 2\psi(s_0) \le h(s_0+\delta) + h(s_0-\delta) - 2h(s_0) \le 0.$$

We divide by δ^2 and let δ go to zero. The result is $\varphi_{ss}(s_0) \le \Gamma/s_0$. Hence, f is a viscosity supersolution of $-sf_{ss}(s) + \Gamma \ge 0$.

Now suppose that f is a viscosity supersolution of $-sf_{ss}(s) + \Gamma \ge 0$. We need to show that

$$h(s+\delta) + h(s-\delta) - 2h(s) \le 0$$

for any $\delta > 0$. Suppose that there exist s_0 and $\delta > 0$ such that

$$\alpha := h(s_0+\delta) + h(s_0-\delta) - 2h(s_0) > 0.$$

Set

$$\psi(s) := h(s_0) + \frac{h(s_0+\delta) - h(s_0-\delta)}{2\delta}\,(s-s_0) + \frac{\alpha}{4\delta^2}(s-s_0)^2.$$

Then, $(h-\psi)(s_0) = 0$ and

$$(h-\psi)(s_0 \pm \delta) = \frac{1}{2}\left[h(s_0+\delta) + h(s_0-\delta) - 2h(s_0)\right] - \frac{\alpha}{4} = \frac{\alpha}{4}.$$

Hence, $(h-\psi)$ attains a local minimum in $(s_0-\delta, s_0+\delta)$. Set $\varphi(s) := \psi(s) + \Gamma s \ln(s)$ so that $(f-\varphi)$ attains a local minimum in the same interval, say at s^*. We calculate that

$$\Gamma - s^*\varphi_{ss}(s^*) = -s^*\frac{\alpha}{2\delta^2} < 0.$$

This contradicts the supersolution property of f. □

10. The comparison result. This section is devoted to the proof of a comparison theorem which was used in the proof of our main result. We refer to Crandall, Ishii, and Lions (1992) and Fleming and Soner (1993) for the definition and the properties of viscosity solutions.

THEOREM 10.1. *Let Assumption* 4.1 *hold. Suppose that the variational inequality* (4.3)–(4.4) *has a solution* $\hat{v} \in C^{1,2}([0,T] \times (0,\infty))$ *which is polynomially growing and has bounded* $\mathcal{L}\hat{v}$. *Let* u *be a lower semicontinuous viscosity supersolution of* (4.3) *satisfying* $u(T,\cdot) \geq \hat{g}$, $u(\cdot,0) \geq \hat{v}(\cdot,0)$, *and* $u(t,s) \geq E\left[\hat{g}(S_{t,s}(T)\right]$. *Then,*

$$u \geq \hat{v} \qquad \text{on } [0,T] \times (0,\infty).$$

We start with deriving an upper bound for the solution $\hat{v}$ of (4.3)–(4.4).

LEMMA 10.2. *For all* $(t,s) \in [0,T] \times (0,\infty)$, $\hat{v}(t,s) \leq \hat{g}^{\mathrm{conc}}(s)$.

Proof. To prove this result, we first show that $\hat{v}$ is related to some stochastic control problem. Let $\mathcal{N}$ be the set of all bounded nonnegative progressively measurable processes. For all $\nu \in \mathcal{N}$, consider the controlled process $S^{\nu}_{t,s}$ defined by

$$\frac{dS^{\nu}_{t,s}(u)}{S^{\nu}_{t,s}(u)} = \left[\frac{\nu(u)}{1+S^{\nu}_{t,s}(u)} + \sigma^2\left(t, S^{\nu}_{t,s}(u)\right)\right]^{1/2} dW(u).$$

Notice that the random function $s \longmapsto s\left[\nu(1+s)^{-1} + \sigma^2(t,s)\right]^{1/2}$ is Lipschitz uniformly in t and therefore the process S^{ν} is well defined. Next, for some small parameter $\eta > 0$, define the stochastic control problem

$$u(t,s) := \sup_{\nu \in \mathcal{N}} E\left[\hat{g}(S^{\nu}_{t,s}(T)) - \frac{1}{2}(\Gamma - \eta)\int_t^T \nu(u)\frac{S^{\nu}_{t,s}(u)}{1+S^{\nu}_{t,s}(u)}du\right]$$

and consider the approximating problems

$$u^n(t,s) := \sup_{\nu \in \mathcal{N}^n} E\left[\hat{g}(S^{\nu}_{t,s}(T)) - \frac{1}{2}(\Gamma - \eta)\int_t^T \nu(u)\frac{S^{\nu}_{t,s}(u)}{1+S^{\nu}_{t,s}(u)}du\right]$$

with $\mathcal{N}^n$ consisting of elements in $\mathcal{N}$ which are bounded by n. Clearly, for every n we have $u^n(t,s) \leq u(t,s)$ for all $(t,s) \in [0,T] \times (0,\infty)$. By classical arguments, it is easily checked that u^n is a viscosity solution of the Hamilton–Jacobi–Bellman (HJB) equation

$$-\sup_{0 \leq \nu \leq n}\left\{w_t + \frac{1}{2}s^2\left(\sigma^2(t,s) + \frac{\nu}{1+s}\right)w_{ss} - \frac{1}{2}(\Gamma - \eta)\nu\frac{s}{1+s}\right\} = 0$$

which can be written as

$$-\mathcal{L}w - \frac{1}{2}n\frac{s}{1+s}[sw_{ss} - (\Gamma - \eta)]^+ = 0 \quad \text{on } [0,T) \times (0,\infty). \tag{10.1}$$

Now recall that $\hat{v}$ is a classical solution to (4.3).

Case 1. $s\hat{v}_{ss} < \Gamma$, then $\mathcal{L}\hat{v} = 0$ and therefore $-\mathcal{L}\hat{v} - \frac{1}{2}n\frac{s}{1+s}[s\hat{v}_{ss} - (\Gamma - \eta)]^+ \leq 0$.

Case 2. $s\hat{v}_{ss} = \Gamma$, then $\mathcal{L}\hat{v} \geq 0$ and $-\mathcal{L}\hat{v} - \frac{1}{2}n\frac{s}{1+s}[s\hat{v}_{ss} - (\Gamma - \eta)]^+ \leq -\mathcal{L}\hat{v} - \frac{1}{2}n\eta\frac{s}{1+s} \leq 0$ for sufficiently large n; recall that $\mathcal{L}\hat{v}$ is assumed to be bounded uniformly in (t,s).

We have then proved that $\hat{v}$ is a subsolution of the HJB equation (10.1) for sufficiently large n. Since $\hat{v}(T,s) = u^n(T,s) = \hat{g}(s)$, it follows from the comparison theorem (which will be verified at the end of this proof) that $\hat{v} \leq u^n$ and therefore

$$\hat{v}(t,s) \leq u(t,s) \quad \text{ for all } (t,s) \in [0,T] \times (0,\infty).$$

We then have

$$\hat{v}(t,s) \le \sup_{\nu\in\mathcal{N}} E\left[\hat{g}(S^{\nu}_{t,s}(T))\right] \le \sup_{\nu\in\mathcal{N}} E\left[\hat{g}^{\mathrm{conc}}(S^{\nu}_{t,s}(T))\right].$$

By the Jensen inequality and the martingale property of the process $S^{\nu}_{t,s}$, this provides

$$\hat{v}(t,s) \le \sup_{\nu\in\mathcal{N}} \hat{g}^{\mathrm{conc}}\left(E\left[S^{\nu}_{t,s}(T)\right]\right) = \hat{g}^{\mathrm{conc}}(s).$$

It remains to prove the comparison theorem for (10.1). Let m be the growth rate of $\hat{v}$, i.e., $\hat{v}(t,s) \le C(1+s^m)$ for some constant C. Take some $\lambda \ge m(m+1)\sigma^2(t,s)/2$ (recall that $s \longmapsto s\sigma(t,s)$ is Lipschitz uniformly in t and therefore σ is bounded). Choose a minimizer at (t_0,s_0) of

$$\psi(t,s) = e^{\lambda t}u^n(t,s) - e^{\lambda t}\hat{v}(t,s) + \varepsilon s^{m+1},$$

where ε is a small positive parameter. Since $u^n \ge 0$ and $\hat{v}$ is growing at the rate m, ϕ attains its minimum. If $s_0 = 0$ or $t_0 = T$, then $\psi(t_0,s_0) \ge 0$ by the boundary conditions. Now, suppose that $s_0 > 0$ and $t_0 < T$. Since u^n is a viscosity solution of (10.1) and $\hat{v}$ is a classical subsolution of (10.1), it follows that

$$\begin{aligned}
&\lambda e^{\lambda t_0}[u^n(t_0,s_0) - \hat{v}(t_0,s_0)] + \varepsilon\frac{1}{2}\sigma^2(t_0,s_0)m(m+1)s_0^{m(m+1)} \\
&\quad \ge e^{\lambda t_0}\frac{n}{2}\frac{s_0}{1+s_0}\left\{[s_0\hat{v}_{ss}(t_0,s_0) - \Gamma]^+ - [s_0\hat{v}_{ss}(t_0,s_0) - e^{-\lambda t_0}\Gamma - \varepsilon m(m+1)s_0^m]^+\right\} \\
&\quad \ge 0.
\end{aligned}$$

Then $\psi(t_0,s_0) \ge 0$ from the choice of the parameter λ. By sending ε to zero, we obtain the comparison result for (10.1). □

Proof of Theorem 10.1. Fix some positive scalar λ and set $\hat{w}(t,s) = \hat{v}(t,s)e^{-\lambda t}$ and $w(t,s) = u(t,s)e^{-\lambda t}$ for all $(t,s) \in [0,T]\times(0,\infty)$. Then $\hat{w}$ is a $C^{1,2}([0,T)\times(0,\infty))$ solution of the variational inequality

$$\begin{aligned}
\min\left\{\lambda\hat{w} - \mathcal{L}\hat{w};\ \Gamma e^{-\lambda t} - s\hat{w}_{ss}\right\} &= 0 \quad \text{on } [0,T)\times(0,\infty), \\
\hat{w}(T,s) &= \hat{g}(s)e^{-\lambda T}, \quad s > 0,
\end{aligned} \tag{10.2}$$

and w is a lower semicontinuous viscosity supersolution of the above equation. Given $\varepsilon > 0$, define the test function

$$\varphi(t,s) = \hat{w}(t,s) - \varepsilon\phi(s), \quad (t,s) \in [0,T)\times(0,\infty),$$

where ϕ is the function introduced in Assumption 4.1; recall that ϕ is positive, C^2 is strictly concave, and $\lim_{s\to\infty}\phi(s) = +\infty$. By Remark 6.3, we have $v(t,s) \ge E\left[\hat{g}(S_{t,s}(T))\right]$. Moreover, since g is nonnegative, we have $\hat{g} \ge 0$ and by the definition of the concave envelope, it follows that $\hat{g} \ge \hat{g}^{\mathrm{conc}} - C$ for some positive constant C. Then, from Lemma 10.2 together with condition (4.2), we can conclude that

$$\liminf_{s\to\infty}(w-\varphi)(t,s) \ge \liminf_{s\to\infty}\left\{E\left[\hat{g}(S_{t,s}(T))\right] - \hat{g}^{\mathrm{conc}}(s) + \varepsilon\phi(s)\right\} = +\infty$$

for all $t \in [0,T]$. Then there exists $(t_0,s_0) \in [0,T]\times[0,\infty)$ such that

$$(w-\varphi)(t_0,s_0) = \min_{[0,T]\times[0,\infty)}(w-\varphi).$$

In order to prove the required result, we have to show that

$$(w-\varphi)(t_0,s_0)\geq 0, \tag{10.3}$$

which implies that $w(t,s)-\hat{w}(t,s)+\varepsilon\phi(s)\geq 0$ for all $(t,s)\in[0,T]\times(0,\infty)$ and the result of the theorem follows by sending ε to zero.

Inequality (10.3) is trivially satisfied if $s_0=0$ or $t_0=T$. We then concentrate on the case $t_0<T$ and $s_0>0$. Since (t_0,s_0) is an interior minimum, it follows from the viscosity supersolution property of w that

$$\lambda w(t_0,s_0)-\mathcal{L}\varphi(t_0,s_0)\geq 0 \qquad \text{and} \qquad \Gamma e^{-\lambda t_0}-s_0\varphi_{ss}(t_0,s_0)\geq 0. \tag{10.4}$$

Recalling the definition of φ, the second inequality provides

$$\Gamma e^{-\lambda t_0}-s_0\hat{w}_{ss}(t_0,s_0)\geq -\varepsilon\phi_{ss}(s_0)>0.$$

By (10.2), we then see that $\mathcal{L}\hat{w}(t_0,s_0)=\lambda\hat{w}(t_0,s_0)$. Plugging this into the first inequality of (10.4) provides

$$\lambda(w-\varphi)(t_0,s_0)\geq\varepsilon\left[\lambda\phi(s_0)-\frac{1}{2}\sigma^2(t_0,s_0)\phi_{ss}(t_0,s_0)\right]\geq 0,$$

which is the required inequality (10.3). □

11. Appendix: Properties of stochastic integrals. In this section we prove several properties of double stochastic integrals with respect to Brownian motion. The key idea in our analysis was provided by Professor F. Delbaen. Our main result is Proposition 11.6 below.

It is known that if

$$\int_0^{t\wedge\theta}h(u)dW(u)\leq Ct\wedge\theta \qquad \text{for all } t\geq 0, \tag{11.1}$$

for some continuous adapted process $h(\cdot)$, standard Brownian motion $W(\cdot)$, positive stopping time θ, and a constant C, then $h(0)=0$. This result is contained in Soner, Shreve, and Cvitanič (1995).[1]

In the analysis of gamma constraints, in particular in proving the viscosity property of the value function in section 9, we are led to study a similar situation for double stochastic integrals such as

$$\int_0^t\int_0^u b(r)dW(r)\ dW(u)\leq Ct. \tag{11.2}$$

In this section, we analyze several inequalities of the type (11.2) ordered by increasing difficulty.

First, suppose that the process $b(\cdot)$ in (11.2) is equal to a constant b_0. Then,

$$\frac{b_0}{2}[W^2(t)-t]=\int_0^t\int_0^u b(r)dW(r)\ dW(u)\leq Ct.$$

[1] Here is an alternative simple proof of this result. Given an arbitrary $\nu\in\mathbb{R}$, introduce the exponential martingale $Z^\nu=\mathcal{E}(\nu W)$. Then, multiplying both sides of (11.1) by $Z^\nu(t\wedge\theta)$, and taking expectations, it follows from the optional sampling theorem that $\nu E[Z^\nu(t\wedge\theta)\int_0^{t\wedge\theta}h(u)du]\leq CE[Z^\nu(t\wedge\theta)(t\wedge\theta)]$. Dividing by t, sending t to zero, and recalling that the process $h(.)$ is continuous, and the stopping time θ is positive P-a.s., we see that $\nu h(0)\leq C$. By arbitrariness of ν, this proves that $h(0)=0$.

Hence, $b_0 \leq 0$ by the law of iterated logarithm.

Next, suppose that b is a bounded, progressively measurable process and (11.2) holds for all $t \in [0, \eta]$ where η is a positive constant. Delbaen proved the following:

$$P\left[\inf_{0 \leq u \leq t} b(u) \geq c\right] < 1 \qquad \text{for all } c > 0,\ t \leq \eta. \tag{11.3}$$

Suppose to the contrary, i.e., suppose that there are $c > 0, t \leq \eta$ such that $b(u) \geq c$ for all $u \in [0, t]$. Let $Z^\nu(t) := \exp\big(\nu W(t) - (\nu^2 t/2)\big)$. A direct calculation shows that

$$E\left[Z^\nu(t) \int_0^t \int_0^u b(v) dW(v) dW(u)\right] = \nu^2 E\left[\int_0^t \int_0^u b(v) Z^\nu(v) dv du\right] \geq c\nu^2 t^2/2.$$

By (11.2),

$$E\left[Z^\nu(t) \int_0^t \int_0^u b(v) dW(v) dW(u)\right] \leq Ct.$$

Hence, $c\nu^2 t^2/2 \leq Ct$ for all ν, which cannot happen. This proves (11.3).

We continue the analysis when (11.2) holds only up to a stopping time.

Lemma 11.1. *Let θ be some bounded positive stopping time and $\{b(t),\ t \geq 0\}$ be a bounded progressively measurable process satisfying* (11.2) *for all $t \leq \theta$. Then,*

$$\liminf_{t \searrow 0} b(t) \leq 0. \tag{11.4}$$

Proof. Suppose to the contrary. Then, there exist a positive stopping time τ and a constant $c > 0$ such that $b(t \wedge \tau) \geq c$ for all t. Rename the stopping time $\tau \wedge \theta$ to be θ.

Step 1. We employ a time change and then use standard properties of Brownian motion to obtain a contradiction. Set

$$h(t) := \int_0^t [b(u)^2 + 1_{\{u > \theta\}}] du, \quad t \geq 0,$$

so that h is a continuous strictly increasing function on $[0, \theta]$. Let

$$\hat{W}(t) := \int_0^{h^{-1}(t)} b(u) dW(u), \quad t \geq 0,$$

and $\mathbb{G} = \{\mathcal{G}_t,\ t \geq 0\}$ be given by $\mathcal{G}_t := \mathcal{F}_{h^{-1}(t)}$. Then the time-changed process $(\hat{W}, \mathbb{G})$ is a standard Brownian motion. By the time-change formula (see, e.g., Karatzas and Shreve (1991, Proposition 4.8, p. 176)), we rewrite (11.2) as

$$\begin{aligned} Ct \wedge \theta \geq \int_0^{t \wedge \theta} \int_0^u b(r)\ dW(r)\ dW(u) &= \int_0^{t \wedge \theta} \hat{W}(h(u))\ dW(u) \\ &= \int_0^{h(t \wedge \theta)} \phi(u) \hat{W}(u)\ d\hat{W}(u) \\ &= \frac{1}{2} \int_0^{h(t \wedge \theta)} \phi(u) d[\hat{W}(u)^2] - \frac{1}{2} \int_0^{h(t \wedge \theta)} \phi(u) du, \end{aligned}$$

where $\phi(u) := 1/b(h^{-1}(u))$. Since b is bounded away from zero, ϕ is bounded and

$$\int_0^{h(t\wedge\theta)} \phi(u)d[\hat{W}(u)^2] \leq C't \wedge \theta, \qquad t \geq 0, \tag{11.5}$$

for some constant C'.

Step 2. By the law of iterated logarithm, there exists a sequence of bounded positive $\mathbb{F}$-stopping times $(\tau_n)_n$ converging to zero such that

$$\frac{\hat{W}(\tau_n)^2}{\tau_n} \longrightarrow +\infty \qquad P\text{-a.s.}$$

Set

$$\theta_n := \theta \wedge h^{-1}(\tau_n)\ .$$

Since θ is positive, for sufficiently large n, $h(\theta_n) = h(h^{-1}(\tau_n)) = \tau_n$. Hence,

$$\frac{\hat{W}\left(h(\theta_n)\right)^2}{h(\theta_n)} \longrightarrow +\infty \qquad P\text{-a.s.} \tag{11.6}$$

Step 3. Choose M so that $|b| < M$. Let ϕ be as in Step 1. Since $b > 0$ on $[0,\theta]$, we have $\phi > 1/M$ on this interval.

Set $\ell := \liminf_{t\downarrow 0} \frac{2}{t}\int_0^t [\phi(u) - \frac{1}{M}]d[W^2(u)]$, and let $(\zeta_n)_n$ be a sequence of positive stopping times converging to zero P-a.s. such that

$$\int_0^{\zeta_n} \left[\phi(u) - \frac{1}{M}\right] d[W^2(u)] \leq \ell\zeta_n.$$

Direct calculation provides

$$\ell\ E[\zeta_n] \geq E\left[\int_0^{\zeta_n}\left[\phi(u) - \frac{1}{M}\right] d[W^2(u)]\right] = E\left[\int_0^{\zeta_n}\left[\phi(u) - \frac{1}{M}\right] du\right] \geq 0.$$

This proves that $\ell \geq 0$, and consequently

$$\liminf_{t\downarrow 0} \frac{\int_0^t \phi(u)d[\hat{W}(u)^2]}{\hat{W}(t)^2} \geq \frac{1}{M}.$$

Let θ_n be the sequence constructed in Step 2. Since θ_n tends to zero as n approaches to zero,

$$\liminf_{n\to\infty} \frac{\int_0^{h(\theta\wedge\tau_n)} \phi(u)d[\hat{W}(u)^2]}{\hat{W}\left(h(\theta\wedge\tau_n)\right)^2} \geq \frac{1}{M}. \tag{11.7}$$

Step 4. Since $b(\theta \wedge t) \geq c$, the definition of h implies that

$$\lim_{n\to\infty} \frac{h(\theta_n)}{\theta_n} \geq c^2.$$

Combining this inequality with (11.6) and (11.7), we arrive at

$$\limsup_{n\to\infty} \frac{h(\theta_n)}{\theta_n}\frac{\hat{W}\left(h(\theta_n)\right)^2}{h(\theta_n)}\frac{\int_0^{h(\theta_n)} \phi(u)d[\hat{W}(u)^2]}{\hat{W}\left(h(\theta_n)\right)^2} = +\infty.$$

Step 5. By (11.5), we have

$$\frac{h(\theta_n)}{\theta_n}\frac{\hat{W}\left(h(\theta_n)\right)^2}{h(\theta_n)}\frac{\int_0^{h(\theta_n)}\phi(u)d[\hat{W}(u)^2]}{\hat{W}\left(h(\theta_n)\right)^2}\leq C'\frac{\theta_n}{\theta_n}.$$

Clearly this is in contradiction with the previous step. □

Our next generalization is to replace W in (11.2) by the stock price process.

We introduce some notation that will be used throughout this section. Let (t_n, s_n) be a sequence converging to some $(t_0, s_0) \in [0,T) \times (0,\infty)$. To simplify the notation, we set

$$S_n(t) := S_{t_n,s_n}(t) \qquad \text{and} \qquad \bar{\sigma}_n(t) := S_{t_n,s_n}(t)\sigma\left(t, S_{t_n,s_n}(t)\right).$$

Since the processes S_n may take very large values, we need to introduce a sequence of stopping times defined as follows. For a large constant $\lambda > 0$ let

$$\bar{\tau}_n := \inf\left\{t > t_n : \left|\ln\left(S_n(t)/s_n\right)\right| \geq \lambda\right\}. \tag{11.8}$$

In our notation, we do not show the dependence of $\bar{\tau}_n$ on λ.

Remark 11.2. The sequence of stopping times $(\bar{\tau}_n)_n$ satisfies

$$\liminf_{n\to\infty} t\wedge\bar{\tau}_n \geq \frac{1}{2}t\wedge\bar{\tau}_0 \qquad P\text{-a.s.}$$

Indeed, since $(t_n, s_n) \longrightarrow (t_0, s_0)$, it follows from Protter (1990, Theorem 37, p. 246) that for almost everywhere (a.e.) $\omega \in \Omega$, we have

$$S_{t_n,s_n} \longrightarrow S_{t_0,s_0} \qquad \text{uniformly on } [t_0, t_0+t],$$

which implies the announced claim.

Lemma 11.3. *Let b, θ, C be as in Lemma* 11.1. *Suppose that*

$$\int_0^{t\wedge\theta}\int_0^r b(r)\ dS_0(r)\ dS_0(u) \leq Ct\wedge\theta \qquad \textit{for all } t \geq 0.$$

Then, b satisfies 11.4.

Proof. We follow the proof of Lemma 11.1. We replace θ by the stopping time $\bar{\theta} := \theta \wedge \bar{\tau}_0$ and the time-change function h by

$$\bar{h}(t) := \int_0^t [b(u)^2\bar{\sigma}(u)^2 + 1_{\{u>\bar{\theta}\}}]du.$$

We define the time-changed Brownian motion $\hat{W}$ in the obvious way. Then, the time-change formula implies that

$$\int_0^{t\wedge\bar{\theta}}\int_0^u b(r)dS_0(r)\ dS_0(u) = \int_0^{t\wedge\bar{\theta}}\hat{W}(u)\ dS_0(u) = \int_0^{h(t\wedge\bar{\theta})}\bar{\phi}(u)\hat{W}(u)d\hat{W}(u),$$

where $\bar{\phi} = 1/[\bar{b}(\bar{h}^{-1})]$. We then proceed as in Lemma 11.1. □

Remark 11.4. The conclusion of Lemma 11.1 is still valid if t is substituted for $t\wedge\theta$ in the right-hand side of inequality (11.2). This is easily checked by going through the proof. The same observation prevails for Lemma 11.3.

Finally, we provide two results which deal with a slightly general double integral:

$$M_n(t\wedge\theta_n) := \int_{t_n}^{t_n+t\wedge\theta_n}\left(z_n+\int_{t_n}^{u}a_n(r)dr+\int_{t_n}^{u}b_n(r)dS_n(r)\right)dS_n(u)\le\beta_n+Ct. \tag{11.9}$$

We will first show that if β_n tends to zero, then z_n also converges to zero. This is a slight generalization of the result on single stochastic integrals stated in the beginning of this section. The second result provides information on the limit behavior of the sequence $(b_n)_n$.

Proposition 11.5. *Let $(\{a_n(u),\ u\ge 0\})_n$ and $(\{b_n(u),\ u\ge 0\})_n$ be two sequences of real-valued, progressively measurable processes that are uniformly bounded in n. Suppose that* (11.9) *holds with real numbers $(z_n)_n$, $(\beta_n)_n$, and stopping times $(\theta_n)_n$. Assume further that, as n tends to zero,*

$$\beta_n \longrightarrow 0 \qquad \textit{and} \qquad t\wedge\theta_n\longrightarrow t\wedge\theta_0 \quad P\text{-a.s.},$$

where θ_0 is a strictly positive stopping time. Then

$$\lim_{n\to\infty} z_n = 0.$$

Proof. For each $n\ge 0$, define the stopping time

$$\tau_n := 1\wedge\bar\tau_n\wedge\theta_n.$$

By Remark 11.2, $\liminf_n t\wedge\tau_n \ge t\wedge\tau_0/2$ with probability one. Let ν be an arbitrary real parameter and define the local martingales Z_n^ν by

$$Z_n^\nu(t)=\mathcal{E}\left(\int_0^t\frac{\nu dW(u)}{\bar\sigma_n(u)}\right),\quad t\ge 0.$$

By the definition of τ_n in (11.8), the process $\{Z_n^\nu(t\wedge\tau_n),\ t\ge 0\}$ is a P-martingale. We then define the probability measure P_n^ν equivalent to P by its density process $\{Z_n^\nu(t\wedge\tau_n),\ t\ge 0\}$ with respect to P. We shall denote by E_n^ν the expectation operator under P_n^ν. By Girsanov's theorem, the process $W_n^\nu(.\wedge\tau_n)$ defined by

$$W_n^\nu(t)=W(t)-\int_0^t\frac{\nu\, du}{\bar\sigma_n(u)},\quad t\ge 0,$$

is a Brownian motion under P_n^ν. We also define the local martingale Z^ν by

$$Z^\nu(t)=\mathcal{E}\left(\int_0^t\frac{\nu dW(u)}{\bar\sigma_0(u)}\right),\quad t\ge 0.$$

By the same argument as above, the process $\{Z^\nu(t\wedge\tau_0),\ t\ge 0\}$ is a P-martingale and is therefore the density process of some probability measure P^ν equivalent to P. We shall denote by E^ν the expectation operator under P^ν. It is easily checked that $Z_n^\nu(.)\longrightarrow Z^\nu(.)$ P-a.s. Then, since $t\wedge\tau_0/2\le\liminf_n t\wedge\tau_n\le\limsup_n t\wedge\tau_n\le t$, it follows from the continuity of Z_n^ν and Z^ν that

$$Z_\infty^\nu := \liminf_{n\to\infty} Z_n^\nu(t \wedge \tau_n) > 0. \tag{11.10}$$

Rewrite $M_n(t \wedge \tau_n)$ in terms of W_n^ν,

$$\begin{aligned} M_n(t \wedge \tau_n) &= \mathrm{mart}(P_n^\nu) + \nu z_n t \wedge \tau_n + \nu \int_{t_n}^{t_n+t\wedge\tau_n} \int_{t_n}^{u} a_n(r) dr du \\ &\quad + \nu t \wedge \tau_n \int_{t_n}^{t_n+t\wedge\tau_n} b_n(r)\bar{\sigma}_n(r) dW_n^\nu(r) + \nu^2 \int_{t_n}^{t_n+t\wedge\tau_n} \int_{t_n}^{u} b_n(r) dr du, \end{aligned}$$

where $\mathrm{mart}(P_n^\nu)$ is a martingale under P_n^ν starting from zero. Take the expectation under P_n^ν, apply the Cauchy–Schwartz inequality for the third term on the right-hand side, and also utilize the bounds on $(a_n)_n$ and $(b_n)_n$ to obtain

$$\begin{aligned} \nu z_n E_n^\nu[t \wedge \tau_n] &\le \beta_n + C' \left(E_n^\nu[t \wedge \tau_n] + (|\nu| + \nu^2) E_n^\nu[(t \wedge \tau_n)^2]^{3/4} \right) \\ &\le \beta_n + C' \left(t + (|\nu| + \nu^2) t^{3/4} \right). \end{aligned}$$

Let ℓ denote either $\liminf_n z_n$ or $\limsup_n z_n$, and restrict ν to have the same sign as ℓ, so that $\nu\ell \ge 0$. Now, let n go to infinity. Then, it follows from Fatou's lemma together with (11.2) and (11.10) that

$$\frac{1}{2} \nu \ell E[t \wedge \tau_0 Z_\infty^\nu] \le C' \left(t + (|\nu| + \nu^2) t^{3/4} \right).$$

We now divide by t and take the limit as $t \searrow 0$. Since τ_0 and Z_∞^ν are positive P (and then P^ν)-a.s., we get by dominated convergence

$$\nu\ell \le C' \qquad \text{for all } \nu \in \mathbb{R}.$$

Since ν is arbitrary, we conclude that $\liminf_n z_n = \limsup_n z_n = 0$. □

The following result is a stronger version of Lemma 11.1 which was used in section 9. We shall denote by $\mathbb{H}^2$ the Hilbert space of all progressively measurable Lebesgue(0,T)$\otimes P$-square integrable processes.

Let $(b_n)_n$ be as in Lemma 11.5. By assumption, $(b_n)_n$ is bounded in L^∞(Lebesgue$(0, T) \otimes P)$. Then it is bounded in $\mathbb{H}^2$ and, therefore, converges weakly to some b, possibly along a subsequence.

PROPOSITION 11.6. *Assume the hypothesis of Lemma* 11.5. *Let b be as above. Then*

$$\liminf_{u\searrow 0} b(u) \le 0.$$

Proof. Define the stopping times τ_n as in the proof of Lemma 11.5. To simplify the notation, we rename process $b_n(t) 1_{t_n \le t \le t_n + t\wedge\tau_n}$ by $b_n(t)$. By Mazur's lemma, there exists a sequence of coefficients $(\lambda_k^n,\ k \ge n)_n$ with $\lambda_k^n \ge 0$ and $\sum_{k\ge n} \lambda_k^n = 1$ such that

$$\hat{b}_n := \sum_{k\ge n} \lambda_k^n b_k \longrightarrow b \qquad \text{strongly in } \mathbb{H}^2. \tag{11.11}$$

Integrating by parts and using the bound on a_n and $S_n(.\wedge\tau_n)$ provide

$$\begin{aligned}M_n(t\wedge\tau_n) &= z_n[S_n(t_n+t\wedge\tau_n)-s_n]\\ &\quad + S_n(t_n+t\wedge\tau_n)\int_{t_n}^{t_n+t\wedge\tau_n} a_n(r)dr - \int_{t_n}^{t_n+t\wedge\tau_n} S_n(u)a_n(u)du\\ &\quad + \int_{t_n}^{t_n+t\wedge\tau_n}\int_{t_n}^{u} b_n(r)dS_n(r)dS_n(u)\\ &\geq -C't\wedge\tau_n - |z_n|s_n(e^\lambda-1) + \int_{t_n}^{t_n+t\wedge\tau_n}\int_{t_n}^{u} b_n(r)dS_n(r)dS_n(u).\end{aligned}$$

Set $\hat\beta_n := \beta_n + |z_n|s_n(e^\lambda-1)$. Then, from Lemma 11.5, $\hat\beta_n \longrightarrow 0$ as $n\to\infty$ and we get from the inequality satisfied by M_n

$$\int_{t_n}^{t_n+t\wedge\tau_n}\int_{t_n}^{u} b_n(r)dS_n(r)dS_n(u) \leq \hat\beta_n + Kt\wedge\tau_n \tag{11.12}$$

for some positive constant K. Set

$$\varepsilon_n(t) := \int_{t_n}^{t_n+t\wedge\tau_n}\int_{t_n}^{u} b_n(r)dS_n(r)dS_n(u) - \int_{t_0}^{t_0+t\wedge\tau_0}\int_{t_0}^{u} b_n(r)dS_0(r)dS_0(u).$$

We shall later prove that

$$\varepsilon_n(t) \longrightarrow 0 \qquad P\text{-a.s.} \tag{11.13}$$

possibly along a subsequence. Take convex combinations in (11.12) to conclude that

$$\sum_{k\geq n}\lambda_k^n\varepsilon_k(t) + \int_{t_0}^{t_0+t\wedge\tau_0}\int_{t_0}^{u}\hat b_n(r)dS_0(r)dS_0(u) \leq \sum_{k\geq n}\lambda_k^n\left(\hat\beta_k + Kt\wedge\tau_k\right). \tag{11.14}$$

We directly calculate that

$$\begin{aligned}&E\left[\left(\int_{t_0}^{t_0+t\wedge\tau_0}\int_{t_0}^{u}\left(\hat b_n(r)-b(r)\right)dS_0(r)dS_0(u)\right)^2\right]\\ &\quad = E\left[\int_{t_0}^{t_0+t\wedge\tau_0}\left(\int_{t_0}^{u}\left(\hat b_n(r)-b(r)\right)dS_0(r)\right)^2\bar\sigma_0(u)^2du\right]\\ &\quad \leq C_1\, E\left[\int_{t_0}^{t_0+t}\left(\int_{t_0}^{u}\left(\hat b_n(r)-b(r)\right)dS_0(r)\right)^2du\right]\\ &\quad \leq C_2\, E\left[\int_{t_0}^{t_0+t}\int_{t_0}^{u}\left(\hat b_n(r)-b(r)\right)^2drdu\right]\\ &\quad \leq C_3t\|\hat b_n-b\|^2_{\mathbb{H}^2},\end{aligned}$$

where C_i's are constants independent of n. This proves that

$$\int_{t_0}^{t_0+t\wedge\tau_0}\int_{t_0}^{u}\hat b_n(r)dS_0(r)dS_0(u)\longrightarrow\int_{t_0}^{t_0+t\wedge\tau_0}\int_{t_0}^{u}b(r)dS_0(r)dS_0(u)\qquad\text{as } n\to\infty$$

in $L^2(P)$, and therefore P-a.s. along some subsequence. Then, taking a.s. limits in (11.14) and using (11.13), we get

$$\int_{t_0}^{t_0+t\wedge\tau_0} \int_{t_0}^{u} b(r)dS_0(r)dS_0(u) \leq Kt.$$

Since the limit process b inherits the bound on b_n, we apply the result of Lemma 11.3 to complete the proof; see also Remark 11.4.

It remains to prove the convergence result stated in (11.13). Set $\zeta_n = t_n + t\wedge\tau_n$ for $n \geq 0$. By Itô's lemma,

$$\varepsilon_n(t) = A_n + B_n + C_n,$$

where

$$\begin{aligned}
A_n &= [S_n(\zeta_n) - s_n] \int_{t_n}^{\zeta_n} b_n(u)dS_n(u) - [S_0(\zeta_0) - s_0] \int_{t_0}^{\zeta_0} b_n(u)dS_0(u),\\
B_n &= -\int_{t_n}^{\zeta_n} b_n(u)S_n(u)dS_n(u) + \int_{t_0}^{\zeta_0} b_n(u)S_0(u)dS_0(u),\\
C_n &= -\int_{t_n}^{\zeta_n} b_n(u)\bar{\sigma}_n(u)^2du + \int_{t_0}^{\zeta_0} b_n(u)\bar{\sigma}_0(u)^2du.
\end{aligned}$$

It suffices to prove that A_n, B_n, and C_n converge to zero P-a.s. along some subsequence. We prove only the convergence of A_n; the remaining claims are proved similarly.

(i) To simplify the presentation, set $\bar{\sigma}(.) = 0$ outside the stochastic interval $[t_n, \zeta_n]$ and observe that

$$S_n(\zeta_n) - S_0(\zeta_0) = s_n + \int_{t_n}^{\zeta_n} \bar{\sigma}_n(u)dW(u).$$

Since $\bar{\sigma}_n$ is bounded inside the stochastic interval $[t_n, \zeta_n]$, by dominated convergence,

$$\begin{aligned}
&E\left[\left(\int_{t_n}^{\zeta_n} \bar{\sigma}_n(u)dW(u) - \int_{t_0}^{\zeta_0} \bar{\sigma}_0(u)dW(u)\right)^2\right]\\
&\qquad = E\left[\int_{t_0\wedge t_n}^{\zeta_0\vee\zeta_n} (\bar{\sigma}_n(u) - \bar{\sigma}_0(u))^2\, du\right] \longrightarrow 0.
\end{aligned}$$

This proves that

$$S_n(\zeta_n) \longrightarrow S_0(\zeta_0) \qquad P\text{-a.s.}$$

along some subsequence.

(ii) Recall that we have set $b_n(.) = 0$ outside the interval $[t_n, \zeta_n]$. Thus,

$$\begin{aligned}
\int_{t_n}^{\zeta_n} b_n(u)dS_n(u) - \int_{t_0}^{\zeta_0} b_n(u)dS_0(u) = \int_{t_n}^{t_0\vee t_n} b_n(u)dS_0(u) + \int_{\zeta_0\wedge\zeta_n}^{\zeta_n} b_n(u)dS_0(u)\\
+ \int_{t_0\vee t_n}^{\zeta_0\wedge\zeta_n} b_n(u)\left(\bar{\sigma}_n(u) - \bar{\sigma}_0(u)\right)dW(u).
\end{aligned}$$

From the bound on b_n, the first two terms on the right-hand side converge to zero in $L^2(P)$ and therefore P-a.s. along some subsequence. As for the third term,

$$\begin{aligned}
E&\left[\left(\int_{t_0\vee t_n}^{\zeta_0\wedge\zeta_n} b_n(u)\left(\bar\sigma_n(u)-\bar\sigma_0(u)\right)dW(u)\right)^2\right]\\
&= E\left[\int_{t_0\vee t_n}^{\zeta_0\wedge\zeta_n} b_n(u)^2\left(\bar\sigma_n(u)-\bar\sigma_0(u)\right)^2 du\right]\\
&\le C_1 E\left[\int_{t_0\vee t_n}^{\zeta_0\wedge\zeta_n} \left(\bar\sigma_n(u)-\bar\sigma_0(u)\right)^2 du\right]\\
&\le C_2 E\left[\int_{t_0}^{\zeta_0} \left(\bar\sigma_n(u)-\bar\sigma_0(u)\right)^2 du\right],
\end{aligned}$$

where C_i's are constants and we have set $\sigma_n(.) = 0$ outside the stochastic interval $[t_n, \zeta_n]$. Since $\bar\sigma_n$ is bounded, we see by dominated convergence that the third term of interest converges to zero in $L^2(P)$ and therefore P-a.s. along some subsequence. This proves that

$$\int_{t_n}^{t_n+t\wedge\tau_n} b_n(u)dS_n(u) - \int_{t_0}^{t_0+t\wedge\tau_0} b_n(u)dS_0(u) \longrightarrow 0 \qquad P\text{-a.s.}$$

along some subsequence.

By (i) and (ii), $A_n \to 0$ P-a.s. along some subsequence. □

Acknowledgment. We gratefully acknowledge conversations with Professor F. Delbaen; in particular, he has provided us with the key idea in the proof of Lemma 11.5.

REFERENCES

F. Black and M. Scholes (1973), *The pricing of options and corporate liabilities*, J. Political Economy, 81, pp. 637–654.

M. Broadie, J. Cvitanić, and M. Soner (1998), *Optimal replication of contingent claims under portfolio constraints*, Review of Financial Studies, 11, pp. 59–79.

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

J. Cvitanič and I. Karatzas (1993), *Hedging contingent claims with constrained portfolios*, Ann. Appl. Probab., 3, pp. 652–681.

J. Cvitanič, H. Pham, and N. Touzi (1999), *Super-replication in stochastic volatility models under portfolio constraints*, J. Appl. Probab., 36, pp. 523–545.

W.H. Fleming and H.M. Soner (1993), *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York.

A. Friedman (1964), *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ.

E. Jouini and H. Kallal (1995), *Arbitrage in securities markets with transaction costs*, J. Econom. Theory, 5, pp. 197–232.

N.V. Krylov (1980), *Controlled Diffusion Processes*, Springer-Verlag, New York, Heidelberg, Berlin.

I. Karatzas and S.E. Shreve (1991), *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York.

P. Protter (1990), *Stochastic Integration and Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin.

H.M. Soner, S.E. Shreve, and J. Cvitanič (1995), *There is no nontrivial hedging portfolio for option pricing with transaction costs*, Ann. Appl. Probab., 5, pp. 327–355.

H.M. Soner and N. Touzi (2000), *Stochastic Target Problems, Dynamic Programming and Viscosity Solutions*, preprint.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 97–112

HAMILTONIAN NECESSARY CONDITIONS FOR A MULTIOBJECTIVE OPTIMAL CONTROL PROBLEM WITH ENDPOINT CONSTRAINTS*

QIJI J. ZHU†

Abstract. This paper discusses Hamiltonian necessary conditions for a nonsmooth multiobjective optimal control problem with endpoint constraints involving a general preference. The transversality condition in our necessary conditions is stated in terms of normal cones to the level sets of the preference. Examples involving a number of useful preferences are discussed.

Key words. multiobjective optimal control, Hamiltonian necessary conditions, differential inclusions, preferences, utility function, subdifferential calculus, variational principles

AMS subject classifications. 49K24, 90C29

PII. S0363012999350821

1. Introduction.

Practical decision problems often involve many factors and can be described by a vector-valued decision function whose components describe several competing objectives. The comparison between different values of the decision function is determined by a preference of the decision maker. The main purpose of this paper is to derive Hamiltonian necessary conditions for a nonsmooth multiobjective optimal control problem with the dynamics governed by a differential inclusion.

Historically, the concept of a preference first appeared in the value theory of economics. In the early studies of the value theory a preference is often defined by a utility function. One of the central questions in the value theory was: given a preference, is it always possible to define a utility function (with good analytic properties) that determines the preference? In terms of multiobjective optimal control problems this amounts to asking whether it is possible to reduce a multiobjective optimal control problem to an optimal control problem with a reasonable single objective function. In [13], Debreu proved a celebrated theorem which asserts that a preference $\prec$ is determined by a continuous utility function if and only if $\prec$ is continuous in the sense that, for any x, the sets $\{y : x \prec y\}$ and $\{y : y \prec x\}$ are closed. While this theorem plays a central role in the value theory, it is not of much help to us for the following reasons. First, Debreu's theorem is an existence theorem. It does not provide methods for determining the utility function for a given preference. Second, even if one can find a continuous utility function that determines the preference, an optimal control problem with a continuous decision function and endpoint constraints is not easily subject to analysis. Finally, some useful preference relations (e.g., the preference determined by the lexicographical order of the vectors) are not continuous. For these reasons we shall pursue the multiobjective optimal control problem directly.

In the area of multiobjective optimization and optimal control much research has been devoted to the weak Pareto solution and its generalizations. The preference relation for two vectors $x, y \in R^m$ in a weak Pareto sense is defined by $x \prec y$ if and

[*] Received by the editors January 25, 1999; accepted for publication (in revised form) March 28, 2000; published electronically August 17, 2000.
http://www.siam.org/journals/sicon/39-1/35082.html

[†] Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, MI 49008 (zhu@math-stat.wmich.edu). The author's research was supported by the National Science Foundation under grant DMS-9704203.

only if $x_i \leq y_i, i = 1, \dots, m$, and at least one of the inequalities is strict. In other words, $x \prec y$ if and only if $x - y \in K := \{z \in R^m : z$ has nonpositive components$\}$ and $x \neq y$. More generally, one can use other cones K in the definition of the preference relations. Necessary optimality conditions for (generalized) weak Pareto solutions were derived for optimization problems in [1, 8, 11, 25, 29, 35, 36, 39, 40] (see also the survey paper [14] for more information), for linear-quadratic and H^∞ optimal control problems in [16], and for an abstract optimal control problem in [6]. A common key step in deriving necessary conditions for generalized Pareto solutions is to apply a separation theorem to a tangent cone of the attainable set and a translate of the cone $-K$, where K is the cone that generates the preference. In this paper we take a different approach. We use a normal cone condition similar to that in the extremal principle [24, 26, 28] at the optimal point in terms of the normal cones to the attainable set and a level set of the preference. This approach enables us to handle more general preference relations: they are not necessarily defined by a cone and are not even necessarily continuous. Necessary optimality conditions for the weak Pareto solution and its generalizations can be derived and refined by using our necessary conditions.

The technical implementation of our proof relies on recent results in nonsmooth analysis, in particular, on the calculus for smooth subdifferentials of lower semicontinuous functions [2, 4, 5, 9, 10, 19], the methods for proving the extremal principle [24, 26, 28], and techniques in handling the Hamiltonians for a differential inclusion [10, 19]. To avoid technical distractions we prove here Hamiltonian necessary conditions that extend the classical Hamiltonian necessary conditions for optimal control problems derived by Clarke (see [8]). There are several recent significant refinements of the Hamiltonian necessary conditions for optimal control problems with a single objective function [18, 19, 20, 34]. It is an interesting question to what extent the methods of this paper can be used to generalize these refined Hamiltonian necessary conditions to multiobjective optimal control problems. There are also many other types of necessary conditions for optimal control problems, in particular, those that refine and generalize the maximum principle (see, e.g., [21, 22, 23, 27, 30, 31, 33, 37, 38, 41]). Whether those necessary conditions can be extended to multiobjective optimal control problems in our general setting represents a more challenging open problem.

The remainder of the paper is arranged as follows. Section 2 contains definitions and preliminary results in subdifferential calculus. We state our main result in section 3 along with some examples and discussions. The technical proofs are given in section 4.

2. Preliminaries. Let X be a real reflexive Banach space with closed unit ball B_X and with topological real dual X^*. Note that X has an equivalent Fréchet smooth norm and we will use this norm as the norm of X unless otherwise stated. Let $f : X \to \bar{R} := R \cup \{+\infty\}$ be an extended-valued function. We denote by dom $f := \{x \in X : f(x) \in R\}$ the effective domain of f. We assume all our functions are *proper* in that they take some finite values: dom $f \neq \emptyset$. Let us now recall the definitions of subdifferentials and normal cones (see [5] for greater details and historical comments).

DEFINITION 2.1. *Let $f : X \to \bar{R}$ be a lower semicontinuous function and C a closed subset of X. We say f is Fréchet-subdifferentiable and x^* is a Fréchet-subderivative of f at x if there exists a concave C^1 function g such that $g'(x) = x^*$ and $f - g$ attains a local minimum at x. We denote the set of all Fréchet-subderivatives of f at x by $D_F f(x)$. We define the Fréchet-normal cone of C at x to be $N_F(C, x) := D_F \delta_C(x)$, where δ_C is the indicator function of C defined by $\delta_C(x) := 0$ for $x \in C$*

and ∞ *otherwise.*

In the following definition, $w^* - \lim$ represents the weak-star limit. In a reflexive Banach space it coincides with the weak limit. In this paper, we only use limiting subdifferentials and normal cones in finite dimensional Euclidean spaces where the weak-star limit coincides with the usual limit.

Definition 2.2. *Let* $f : X \to \bar{R}$ *be a lower semicontinuous function. Define*

$$\partial f(x) := \{w^* - \lim_{i\to\infty} v_i : v_i \in D_F f(x_i),\, (x_i, f(x_i)) \to (x, f(x))\}$$

and

$$\partial^\infty f(x) := \{w^* - \lim_{i\to\infty} t_i v_i : v_i \in D_F f(x_i),\, t_i \to 0^+,\, (x_i, f(x_i)) \to (x, f(x))\},$$

and call $\partial f(x)$ *and* $\partial^\infty f(x)$ *the limiting subdifferential and singular subdifferential of* f *at* x*, respectively.*

Now let C *be a closed subset of* X *and*

$$N(C, x) := \{w^* - \lim_{i\to\infty} v_i : v_i \in N_F(C, x_i),\, C \ni x_i \to x\},$$

and call $N(S, x)$ *the limiting normal cone of* S *at* x*.*

We shall also need to use the Clarke subdifferential ∂_C which is derived by taking the weak-star closed convex hull of the sum of the limiting and singular subdifferentials, i.e.,

$$\partial_C f(x) := \mathrm{cl}^* \ \mathrm{co}[\partial f(x) + \partial^\infty f(x)].$$

We conclude this section with a sum rule and a chain rule for the Fréchet subdifferential. They can be viewed as nonsmooth versions of the corresponding calculus rules for derivatives. We start with the sum rule. The prototypes of this result appeared first in [17]. We use the following version, derived in [4], which refines similar results in [2, 19].

Definition 2.3 (uniform lower semicontinuity). *Let* $f_1, \ldots, f_N : X \to \bar{R}$ *be lower semicontinuous functions and* E *a closed subset of* X*. We say that* $(f_1, \ldots, f_n)$ *is uniformly lower semicontinuous on* E *if*

$$\inf_{x\in E} \sum_{n=1}^{N} f_n(x) \le \lim_{\eta\to 0} \inf \left\{ \sum_{n=1}^{N} f_n(x_n) : \|x_n - x_m\| \le \eta, x_n, x_m \in E, n, m = 1, \ldots, N \right\}.$$

We say that $(f_1, \ldots, f_N)$ *is locally uniformly lower semicontinuous at* $x \in \cap_{n=1}^N \mathrm{dom}(f_n)$ *if* $(f_1, \ldots, f_N)$ *is uniformly lower semicontinuous on every closed ball centered at* x *in a neighborhood of* x*.*

Theorem 2.4 (sum rule). *Let* $f_1, \ldots, f_N : X \to \bar{R}$ *be lower semicontinuous functions. Suppose that* $(f_1, \ldots, f_N)$ *is locally uniformly lower semicontinuous at* $\bar{x}$ *and* $\sum_{n=1}^N f_n$ *attains a local minimum at* $\bar{x}$*. Then, for any* $\varepsilon > 0$*, there exist* $x_n \in \bar{x} + \varepsilon B$ *and* $x_n^* \in D_F f_n(x_n), n = 1, \ldots, N$ *such that* $|f_n(x_n) - f_n(\bar{x})| < \varepsilon, n = 1, 2, \ldots, N,$ $\mathrm{diam}(x_1, \ldots, x_N) \cdot \max(\|x_1^*\|, \ldots, \|x_N^*\|) < \varepsilon$*, and*

$$\left\| \sum_{n=1}^{N} x_n^* \right\| < \varepsilon.$$

Following the argument of [10, Theorem 9.1], one can deduce the following Fréchet subdifferential chain rule for Lipschitz functions from the sum rule of Theorem 2.4.

THEOREM 2.5 (chain rule). *Let X and Y be reflexive Banach spaces. Let $\Phi : X \to Y$ and $f : Y \to R$ be locally Lipschitz mappings. Then for any $x^* \in D_F(f \circ \Phi)(\bar{x})$ and any $\varepsilon > 0$ there exist $x \in \bar{x} + \varepsilon B_X$, $y \in \Phi(\bar{x}) + \varepsilon B_Y$, and $y^* \in D_F f(y)$ such that $\|\Phi(x) - \Phi(\bar{x})\| < \varepsilon$ and*

$$x^* \in D_F\langle y^*, \Phi\rangle(x) + \varepsilon B_X.$$

3. The main results. Let $\prec$ be a (nonreflexive) preference for vectors in R^m. We consider the following multiobjective optimization problem with endpoint constraints.

$$\mathcal{P} \quad \text{Minimize } \phi(y(1))$$

$$\text{subject to } \dot{y}(t) \in F(y(t)) \text{ almost everywhere (a.e.) in } [0,1], y(0) = \alpha_0, \tag{3.1}$$

$$y(1) \in E. \tag{3.2}$$

Here, $\phi = (\phi_1, \dots, \phi_m)$ is a Lipschitz vector function on R^n, E is a closed subset of R^n, and F is a multifunction from R^n to R^n satisfying the following conditions.

(H1) For every x, $F(x)$ is a nonempty compact convex set.

(H2) F is Lipschitz with rank L_F, i.e., for any x, y,

$$F(x) \subset F(y) + L_F\|x - y\|B_{R^n}.$$

We say that y is a feasible trajectory for problem $\mathcal{P}$ if y is absolutely continuous and satisfies relations (3.1) and (3.2). We say x is a solution to problem $\mathcal{P}$ provided that it is a feasible trajectory for $\mathcal{P}$ and there exists no other feasible trajectory y such that $\phi(y(1)) \prec \phi(x(1))$. For any $r \in R^m$, we write $l(r) := \{s \in R^m : s \prec r\}$. We will need the following regularity assumptions on the preference.

DEFINITION 3.1. *We say that a preference $\prec$ is* closed *provided that*

(A1) *for any $r \in R^m$, $r \in \overline{l(r)}$;*

(A2) *for any $r \prec s$, $t \in \overline{l(r)}$ implies that $t \prec s$.*

We say that $\prec$ is regular *at $\bar{r} \in R^m$ provided that*

(A3) *for any sequences $r_k, \theta_k \to \bar{r}$ in R^m,*

$$\limsup_{k\to\infty} N(\overline{l(r_k)}, \theta_k) \subset N(\overline{l(\bar{r})}, \bar{r}).$$

Our main result is the following theorem.

THEOREM 3.2. *Let x be a solution to the multiobjective optimal control problem $\mathcal{P}$. Suppose that the preference $\prec$ is regular at $\phi(x(1))$. Then there exist an absolutely continuous mapping $p : [0,1] \to R^n$, a vector $\lambda \in N(\overline{l(\phi(x(1)))}, \phi(x(1)))$ with $\|\lambda\| = 1$, and a scalar $\lambda_0 = 0$ or 1 satisfying $\lambda_0 + \|p(t)\| \neq 0 \ \forall\ t \in [0,1]$ such that*

$$(-\dot{p}(t), \dot{x}(t)) \in \partial_C H(x(t), p(t)) \quad \textit{a.e. in } [0,1],$$
$$-p(1) \in \lambda_0 \partial\langle\lambda, \phi\rangle(x(1)) + N(E, x(1)).$$

Moreover, one can always choose $\lambda_0 = 1$ when $x(1) \in \operatorname{int} E$.

Here H is the Hamiltonian corresponding to F defined by

$$H(x,p) := \max\{\langle p, v\rangle : v \in F(x)\}.$$

Remark 3.3. Observing that H is positive homogeneous in p, we can scale p or, alternatively, λ in Theorem 3.2 by a positive constant.

In the remainder of this section we will examine a few examples. The proof of Theorem 3.2 is postponed to the next section.

Example 3.4 (a single objective problem). When $m = 1$ and $r \prec s \iff r < s$, Theorem 3.2 reduces to the classical Hamiltonian necessary conditions for an optimal control problem [8]. Thus, the necessary conditions in Theorem 3.2 are true generalizations of the Hamiltonian necessary conditions for single objective optimal control problems.

Example 3.5 (the weak Pareto optimal control problem). In a weak Pareto optimal control problem we define the preference by $r \prec s$ if and only if $r_i \leq s_i, i = 1, \ldots, m$, and at least one of the inequalities is strict. It is easy to check that $\prec$ defined this way satisfies assumptions (A1) and (A2) in Definition 3.1 at any $r \in R^m$. Moreover, for any $r \in R^m$, $\overline{l(r)} = r + R^m_-$, where $R^m_- := \{s \in R^m : s_i \leq 0, i = 1, \ldots, m\}$. It follows that, for any $r, \theta \in R^m$, we have $N(\overline{l(r)}, \theta) \subset R^m_+ := -R^m_-$. Since $N(\overline{l(r)}, r) = R^m_+$, we can see that $\prec$ also satisfies assumption (A3) at any $r \in R^m$ and, therefore, it is regular at any $r \in R^m$. Combining Theorem 3.2 and Remark 3.3, we obtain the following corollary.

Corollary 3.6. *Let x be a weak Pareto solution to the multiobjective optimal control problem $\mathcal{P}$. Then there exist an absolutely continuous mapping $p : [0,1] \to R^n$, a vector $\lambda \in R^m_+$ with $\sum_{i=1}^m \lambda_i = 1$, and a scalar $\lambda_0 = 0$ or 1 satisfying $\lambda_0 + \|p(t)\| \neq 0 \ \forall\ t \in [0,1]$ such that*

$$(-\dot{p}(t), \dot{x}(t)) \in \partial_C H(x(t), p(t)) \quad a.e.\ in\ [0,1],$$
$$-p(1) \in \lambda_0 \partial \langle \lambda, \phi \rangle (x(1)) + N(E, x(1)).$$

Moreover, one can always choose $\lambda_0 = 1$ when $x(1) \in \operatorname{int} E$.

Remark 3.7 (the strong Pareto optimal control problem). Problem $\mathcal{P}$, with the preference defined by $r \prec s$ if and only if $r_i < s_i$, is a strong Pareto optimal control problem. We can also check that this preference is regular at any $r \in R^m$. Thus, Theorem 3.2 can also yield a necessary condition for the strong Pareto optimal control problem. However, simple calculation yields $\overline{l(r)} = r + R^m_-$, which is the same as the corresponding result for a weak Pareto optimal problem. This means that the necessary conditions deduced from Theorem 3.2 for both weak and strong Pareto optimal control problems are the same. Clearly this loss of precision is due to the closure operation on the level sets $l(r)$.

We point out that if x is a weak Pareto optimal solution to problem $\mathcal{P}$, then it is a solution to the following single objective optimal control problem: minimize $\max(\phi_1(y(1)) - \phi_1(x(1)), \ldots, \phi_m(y(1)) - \phi_m(x(1)))$ subject to constraints (3.1) and (3.2). Then we can deduce Corollary 3.6 by combining the Hamiltonian necessary conditions for a single objective problem and the subdifferential chain rule for the max function. However, this method does not apply without additional assumptions to the following generalized weak Pareto optimal solution [1].

Example 3.8 (a generalized weak Pareto optimal control problem). Let $K \subset R^m$ be a closed cone. We now define the preference by $r \prec s$ if and only if $r - s \in K$ and $r \neq s$. Multiobjective optimal control problems with this preference are called generalized weak Pareto optimal control problems. When $K = R^m_-$, we get the weak Pareto problem. Note that we do not assume any convexity on K. Similarly to the last example, we can check that the preference $\prec$ defined here is regular at any

$r \in R^m$. Moreover, $N(\overline{l(r)}, r) = K^- := \{s \in R^m : \langle s, t\rangle \leq 0, t \in K\}$. In particular, $N(\overline{l(\phi(x(1)))}, \phi(x(1))) = K^-$. Thus, we have the following corollary.

COROLLARY 3.9. *Let x be a solution to the generalized weak Pareto multiobjective optimal control problem $\mathcal{P}$ with the preference defined by a closed cone K. Then there exist an absolutely continuous mapping $p : [0,1] \to R^n$, a vector $\lambda \in K^-$ with $\|\lambda\| = 1$, and a scalar $\lambda_0 = 0$ or 1 satisfying $\lambda_0 + \|p(t)\| \neq 0$ $\forall$ $t \in [0,1]$ such that*

$$(-\dot{p}(t), \dot{x}(t)) \in \partial_C H(x(t), p(t)) \quad a.e. \text{ in } [0,1],$$
$$-p(1) \in \lambda_0 \partial\langle\lambda, \phi\rangle(x(1)) + N(E, x(1)).$$

Moreover, one can always choose $\lambda_0 = 1$ when $x(1) \in \text{int } E$.

Example 3.10 (a preference determined by a utility function). Let u be a continuous utility function that determines the preference, i.e., $s \prec r$ if and only if $u(s) < u(r)$. We need an additional assumption to ensure the regularity of $\prec$ which we summarize in the following lemma. We will use $d(S, r) := \inf\{\|s - r\| : s \in S\}$ to denote the distance between a set S and a point r.

LEMMA 3.11. *Let u be a continuous utility function determining the preference $\prec$. Suppose that*

$$\liminf_{s \to r} d(D_F u(s), 0) > 0. \tag{3.3}$$

Then $\prec$ is regular at r and

$$N(\overline{l(r)}, r) = \partial^\infty u(r) \bigcup \left(\bigcup_{a>0} a\partial u(r) \right).$$

Proof. It follows from (3.3) that $l(r)$ is nonempty. Then conditions (A1) and (A2) in Definition 3.1 follow from the continuity of u. It remains to show that $\prec$ satisfies assumption (A3). First we observe that, for r' sufficiently close to r, $\overline{l(r')} = \{s \in R^m : u(s) - u(r') \leq 0\}$. Thus, $D_F u(r') \subset N_F(\overline{l(r')}, r')$. Taking limits, we have

$$\partial^\infty u(r) \bigcup \left(\bigcup_{a>0} a\partial u(r) \right) \subset N(\overline{l(r)}, r). \tag{3.4}$$

Let r_k, θ_k, and ξ_k be sequences satisfying $r_k, \theta_k \to r$, $\xi_k \in N(\overline{l(r_k)}, \theta_k)$, and $\xi_k \to \xi$. We need to show that $\xi \in N(\overline{l(r)}, r)$. By the definition of the limiting normal cone, and without loss of generality, we may assume that $\xi_k \in N_F(\overline{l(r_k)}, \theta_k)$. Since $N(\overline{l(r)}, r)$ always contains 0, we consider the interesting case when $\xi \neq 0$. Then, when n is sufficiently large, we have $\xi_k \neq 0$. Since $N_F(\overline{l(r_k)}, \theta_k)$ is empty when $u(\theta_k) > u(r_k)$ and $\{0\}$ when $u(\theta_k) < u(r_k)$, we must have $u(\theta_k) = u(r_k)$, i.e., $N_F(\overline{l(r_k)}, \theta_k) = N_F(\overline{l(\theta_k)}, \theta_k) = N_F(\{s : u(s) - u(\theta_k) \leq 0\}, \theta_k)$. Applying [5, Theorem 3.4] (see also [3, 32]), we conclude that there exist $a_k > 0$ and $\zeta_k \in D_F u(\theta_k)$ such that $\|a_k\zeta_k - \xi_k\| < 1/n$. It follows that

$$\lim_{k \to \infty} a_k \zeta_k = \xi.$$

Since ζ_k is bounded away from 0, a_k is bounded. Passing to a subsequence if necessary, we may assume that $a_k \to a$. If $a \neq 0$, then ζ_k converges to an element of $\partial u(r)$ and, therefore, $\xi \in a\partial u(r)$. If $a = 0$, then, by definition, $\xi \in \partial^\infty u(r)$. In view of (3.4) we have shown that $\prec$ is regular at r. The formula for $N(\overline{l(r)}, r)$ also follows. □

Using Lemma 3.11 and Remark 3.3, we have the following corollary of Theorem 3.2.

COROLLARY 3.12. *Let $\prec$ be a preference determined by a utility function u. Suppose that u satisfies the condition of Lemma* 3.11. *Let x be a solution to the multiobjective optimal control problem $\mathcal{P}$. Then there exist an absolutely continuous mapping $p : [0,1] \to R^n$, a nonzero vector $\lambda \in \partial^\infty u(\phi(x(1))) \cup \partial u(\phi(x(1)))$, and a scalar $\lambda_0 = 0$ or* 1 *satisfying $\lambda_0 + \|p(t)\| \neq 0 \ \forall\ t \in [0,1]$ such that*

$$\begin{gathered}(-\dot{p}(t), \dot{x}(t)) \in \partial_C H(x(t), p(t)) \quad a.e.\ in\ [0,1], \\ -p(1) \in \lambda_0 \partial\langle \lambda, \phi\rangle(x(1)) + N(E, x(1)).\end{gathered}$$

Moreover, one can always choose $\lambda_0 = 1$ when $x(1) \in \operatorname{int} E$.

Here we derived necessary conditions for an optimal control problem with a continuous decision function. This example also shows that, under favorable conditions, necessary optimality conditions in terms of a preference and its utility function are the same. However, the condition in terms of the normal cone of the level sets of the preference is intrinsic. In fact, if u is a (smooth) utility function corresponding to preference $\prec$, then so is $v(r) = (u(r) - u(x(1)))^3$. But v has a derivative 0 at $x(1)$. Thus, using v as a decision function, the necessary optimality conditions in Corollary 3.12 will yield no useful information.

Our next example considers the preference determined by the lexicographical order. This preference does not correspond to any real utility function [12, p. 72].

Example 3.13 (the preference determined by the lexicographical order). Write $r \prec s$ if there exist an integer $q \in \{0, 1, \dots, m-1\}$ such that $r_i = s_i, i = 1, \dots, q$, and $r_{q+1} < s_{q+1}$. It is easy to check that $\prec$ satisfies assumptions (A1) and (A2) in Definition 3.1. Straightforward calculation yields $\overline{l(r)} = \{s = (s_1, \dots, s_m) \in R^m : s_1 \leq r_1\}$. It follows that, for any $r, \theta \in R^m$, $N(\overline{l(r)}, \theta) = \{ae_1 : a \geq 0\}$. Here $e_1 = (1, 0, \dots, 0) \in R^m$. Thus, $\prec$ is regular at any $r \in R^m$. Moreover, $N(\overline{l(\phi(x(1)))}, \phi(x(1))) = \{ae_1 : a \geq 0\}$. Combining Theorem 3.2 and Remark 3.3, we have the following corollary.

COROLLARY 3.14. *Let x be a solution to the multiobjective optimal control problem $\mathcal{P}$ with the lexicographical preference. Then there exist an absolutely continuous mapping $p : [0,1] \to R^n$ and a scalar $\lambda_0 = 0$ or* 1 *satisfying $\lambda_0 + \|p(t)\| \neq 0 \ \forall\ t \in [0,1]$ such that*

$$\begin{gathered}(-\dot{p}(t), \dot{x}(t)) \in \partial_C H(x(t), p(t)) \quad a.e.\ in\ [0,1], \\ -p(1) \in \lambda_0 \partial\phi_1(x(1)) + N(E, x(1)).\end{gathered}$$

Moreover, one can always choose $\lambda_0 = 1$ when $x(1) \in \operatorname{int} E$.

Intuitively this tells us that since objective ϕ_1 is much more important than the other objectives, the necessary conditions for a multiobjective optimal control problem with the lexicographical preference are the same as the necessary conditions for an optimal control problem with a single objective function ϕ_1. To get further information one can add an additional endpoint constraint $D = \{y : \phi_1(y) = \phi_1(x(1))\}$ to obtain the following necessary conditions: there exist an absolutely continuous mapping $p : [0,1] \to R^n$ and a scalar $\lambda_0 = 0$ or 1 satisfying $\lambda_0 + \|p(t)\| \neq 0 \ \forall\ t \in [0,1]$ such that

$$\begin{gathered}(-\dot{p}(t), \dot{x}(t)) \in \partial_C H(x(t), p(t)) \quad \text{a.e. in } [0,1], \\ -p(1) \in \lambda_0 \partial\phi_2(x(1)) + N(E \cap D, x(1)).\end{gathered}$$

This process can be continued. Note that the adjoint arcs p and scalars λ_0 in this sequence of necessary conditions are not necessarily the same. Of course, these necessary conditions can also be derived by directly combining the above argument with necessary conditions for single-valued optimal control problems. What is interesting here is that this problem can be put into the framework of Theorem 3.2 along with many other problems although its preference cannot be characterized by a utility function.

We should point out that while Theorem 3.2 provides a uniform treatment of many different kinds of multiobjective optimal control problems, the cost we pay is the loss of precision in some special cases. We have briefly discussed this point in Remark 3.7. Now we can see that a similar loss of precision also occurs in the necessary conditions derived for multiobjective optimal control problems defined by a lexicographical order. As discussed in Remark 3.7, to improve the precision one has to use a normal cone concept that can distinguish level sets with an identical closure. This appears to be an interesting direction for further investigation.

4. Proof of Theorem 3.2.

We divide the proof into several steps.

Step 1. *Converting the multiobjective optimal control problem into an abstract optimization problem.* The method we use here develops a conversion for the single objective problem that can be traced back to [7] (see also [8, 10]). The way we handle the multiobjective preference is suggested by the proof of the extremal principle in [24, 26, 28]. Let $A := \{y(1) : y \text{ is a solution to (3.1)}\}$, and let $\Gamma := \{\phi(w) : w \in A \cap E\}$. We can see that Γ is the set of "the attainable values" of this multiobjective optimal control problem. Since A is compact (see [8, section 3.1]) and E is closed, Γ is a closed (compact) set. Let ε be an arbitrary positive number, and let $\eta \in (0, \min(1, \varepsilon)/8(L_\phi + 1))$, where L_ϕ is the Lipschitz rank of ϕ. Choose $\gamma_\eta \prec \phi(x(1))$ such that $\|\gamma_\eta - \phi(x(1))\| < \eta^2$ and write $\Theta := \overline{l(\gamma_\eta)}$. Here Θ is the closure of the level set of $\prec$ at γ_η. This is an approximation of $l(\phi(x(1))$, the level set of the optimal solution x. We need this approximation because the intersection of $l(\phi(x(1))$ and Γ is nonempty (contains at least $\phi(x(1))$), yet

$$\Gamma \cap \Theta = \emptyset. \tag{4.1}$$

In fact, it follows from condition (A2) on $\prec$ that $\Gamma \cap \Theta \neq \emptyset$ implies that there exists a solution y of (3.1) with $y(1) \in E$ such that $\phi(y(1)) \prec \phi(x(1))$, which is a contradiction.

Next we use a method similar to that in [24, 26, 28] for proving the extremal principle to derive a necessary condition for a series of abstract minimization problems that approximate our original multiobjective optimal control problem. Note that the extremal principle in the above references cannot be directly applied here for two reasons: (a) the separation in (4.1) is derived by moving (the closure of) the level sets of $\prec$. In order to apply the extremal principle, the move of the level sets of $\prec$ must be a translate, which only occurs in some special cases such as in a single objective problem or in a weak Pareto optimal problem. (b) Even in the cases when the extremal principle is applicable, applying it to the sets Γ and $l(\phi(x(1))$ will not give us necessary control on the locations of the "approximate" optimal solutions. In what follows, the definition of the auxiliary function f is similar to that in the proof of an extremal principle. We bring the solution of the differential inclusion into the picture by using two sets C and K defined below. Their intersection $C \cap K$ describes the solution set of (3.1) in $L^2 \times L^2$.

Let

$$f(\gamma, \theta) := \|\gamma - \theta\| + \delta_\Gamma(\gamma) + \delta_\Theta(\theta).$$

Then, for any $(\gamma,\theta) \in R^{2m}$, $f(\gamma,\theta) > 0$ and $f(\phi(x(1)),\gamma_\eta) = \|\phi(x(1)) - \gamma_\eta\| < \eta^2$. Let

$$C := \{(u,v) \in L^2([0,1],R^n) \times L^2([0,1],R^n) : v(t) \in F(u(t)) \text{ a.e. in } [0,1]\}$$

and

$$K := \left\{(\alpha,u,v) \in R^n \times L^2([0,1],R^n) \times L^2([0,1],R^n) : \right.$$
$$\left. \alpha = \alpha_0 + \int_0^1 v(s)ds \text{ and } u(t) = \alpha_0 + \int_0^t v(s)ds \right\}.$$

Then for any $\gamma = (\gamma_1,\dots,\gamma_m), \alpha, u$, and v we have

$$\delta_\Gamma(\gamma) \le \sum_{i=1}^m \delta_{\text{graph } \phi_i}(\alpha,\gamma_i) + \delta_E(\alpha) + \delta_K(\alpha,u,v) + \delta_C(u,v).$$

In fact, if $\gamma \in \Gamma$, this inequality is trivial. When $\gamma \notin \Gamma$ it is clear that the indicator functions on the right-hand side cannot be all zero. Moreover, since x is a solution of problem $\mathcal{P}$, we have

$$0 = \delta_\Gamma(\phi(x(1))) = \sum_{i=1}^m \delta_{\text{graph } \phi_i}(x(1),\phi(x(1))) + \delta_E(x(1)) + \delta_K(x(1),x,\dot{x}) + \delta_C(x,\dot{x}).$$

Now we introduce another auxiliary function similar to f which brings the solution of (3.1) into the picture. Let

$$\psi(\gamma,\theta,\alpha,u,v) := \|\gamma-\theta\| + \delta_\Theta(\theta) + \sum_{i=1}^m \delta_{\text{graph } \phi_i}(\alpha,\gamma_i) + \delta_E(\alpha) + \delta_K(\alpha,u,v) + \delta_C(u,v).$$

It is easy to check that $\psi > 0$ and $\psi(\phi(x(1)),\gamma_\eta,x(1),x,\dot{x}) = \|\phi(x(1)) - \gamma_\eta\| < \eta^2$. Moreover, since all the functions in ψ are either Lipschitz functions or indicator functions for a closed set, ψ is lower semicontinuous. By virtue of the Ekeland variational principle [15], there exist $\tilde{\gamma} \in \phi(x(1)) + \eta B_{R^m}$, $\tilde{\theta} \in (\gamma_\eta + \eta B_{R^m}) \cap \Theta \subset (\phi(x(1)) + 2\eta B_{R^m}) \cap \Theta$, $\tilde{\alpha} \in (x(1) + \eta B_{R^n}) \cap E$, $\tilde{u} \in x + \eta B_{L^2([0,1],R^n)}$, and $\tilde{v} \in \dot{x} + \eta B_{L^2([0,1],R^n)}$ such that

$$\psi(\gamma,\theta,\alpha,u,v) + \eta\|(\gamma,\theta,\alpha,u,v) - (\tilde{\gamma},\tilde{\theta},\tilde{\alpha},\tilde{u},\tilde{v})\|$$

attains a minimum at $(\gamma,\theta,\alpha,u,v) = (\tilde{\gamma},\tilde{\theta},\tilde{\alpha},\tilde{u},\tilde{v})$.

We turn to the task of decoupling information. To simplify notation we write $z := (\gamma,\theta,\alpha,u,v)$ and $Z := R^{2m+n} \times L^2([0,1],R^n) \times L^2([0,1],R^n)$. Define functions

$$f_1(z) := \|\gamma - \theta\| + \delta_\Theta(\theta) + \sum_{i=1}^m \delta_{\text{graph } \phi_i}(\alpha,\gamma_i) + \delta_E(\alpha),$$

$$f_2(z) := \delta_K(\alpha,u,v),$$

$$f_3(z) := \delta_C(u,v),$$

and

$$f_4(z) := \eta\|(\gamma,\theta,\alpha,u,v) - (\tilde{\gamma},\tilde{\theta},\tilde{\alpha},\tilde{u},\tilde{v})\|.$$

Then, f_1, f_2, f_3, f_4 are lower semicontinuous and $f_1+f_2+f_3+f_4$ attains a minimum at $\tilde{z}$ over a closed neighborhood U of $\tilde{z}$ in Z. Our next step is to apply a fuzzy sum rule to convert this into subdifferential information on $f_i, i=1,2,3,4$.

Step 2. *Applying the sum rule.* To do so, we need to check that (f_1,f_2,f_3,f_4) is uniformly lower semicontinuous around $\tilde{z}$. The argument is similar to that of [19] but somewhat simpler because of the weaker condition required in the sum rule of Theorem 2.4 (see [42] for the case of single objective problems). Let $z_1^k, z_2^k, z_3^k, z_4^k \in U$ be four sequences satisfying

$$\text{diam}(z_1^k, z_2^k, z_3^k, z_4^k) \to 0, \text{ as } k\to\infty, \tag{4.2}$$

such that

$$\lim_{n\to\infty}\sum_{i=1}^4 f_i(z_i^k) = \liminf_{h\to 0}\left\{\sum_{i=1}^4 f_i(z_i), \text{diam}(z_1,z_2,z_3,z_4)\le h\right\}.$$

Then we must have $(u_3^k, v_3^k)\in C$, i.e.,

$$v_3^k(t)\in F(u_3^k(t)) \text{ a.e. in } [0,1]. \tag{4.3}$$

Since $z_3^k\in U$, u_3^k is a bounded sequence in $L^2([0,1],R^n)$. Since F is Lipschitz with values that are compact sets, v_3^k is also a bounded sequence in $L^2([0,1],R^n)$. Without loss of generality, we may assume that v_3^k converges weakly to $\bar{v}$ in $L^2([0,1],R^n)$. Then it follows from relation (4.2) that v_2^k also converges weakly to $\bar{v}$ in $L^2([0,1],R^n)$. Since $(\alpha_2^k, u_2^k, v_2^k)\in K$, we have $\alpha_2^k = \alpha_0 + \int_0^1 v_2^k(s)ds$ and $u_2^k(t) = \alpha_0 + \int_0^t v_2^k(s)ds$. Thus, we may assume that α_2^k converges to $\bar{\alpha} = \alpha_0 + \int_0^1 \bar{v}(s)ds$ and u_2^k converges to $\bar{u}(t) = \alpha_0 + \int_0^t \bar{v}(s)ds$ in $L^2([0,1],R^n)$. By (4.2), u_3^k converges to $\bar{u}$ in $L^2([0,1],R^n)$. It follows from (4.3) that, for almost all $t\in[0,1]$,

$$\langle v^*, v_3^k(t)\rangle \le \sup\{\langle v^*, v\rangle : v\in F(u_3^k(t))\}\ \forall v^*\in R^n.$$

Taking limits as $k\to\infty$ yields, for almost all $t\in[0,1]$,

$$\langle v^*, \bar{v}(t)\rangle \le \sup\{\langle v^*, v\rangle : v\in F(\bar{u}(t))\}\ \forall v^*\in R^n.$$

Thus, $\bar{v}(t)\in F(\bar{u}(t))$ a.e. in $[0,1]$. The convergence of α_2^k to $\bar{\alpha}$ combined with (4.2) implies that α_1^k also converges to $\bar{\alpha}$. Passing to a subsequence, if necessary, we may assume that the bounded sequences γ_1^k and θ_1^k converge to $\bar{\gamma}$ and $\bar{\theta}$, respectively. Write $\bar{z} := (\bar{\gamma},\bar{\theta},\bar{\alpha},\bar{u},\bar{v})$. Since $f_2(z_2^k) = f_3(z_3^k) = 0$ for sufficiently large k, since f_1 is lower semicontinuous, and since f_4 is weakly lower semicontinuous, we have

$$\lim_{k\to\infty}\sum_{i=1}^4 f_i(z_i^k) \ge f_1(\bar{z}) + f_4(\bar{z}) = \sum_{i=1}^4 f_i(\bar{z}) \ge \sum_{i=1}^4 f_i(\tilde{z}).$$

This verifies that (f_1,f_2,f_3,f_4) is uniformly lower semicontinuous at $\tilde{z}$.

Now we can apply the fuzzy sum rule of Theorem 2.4 to $\sum_{i=1}^4 f_i$ at $\tilde{z}$. Noticing that f_4 is Lipschitzian with rank η, we conclude that there exist $z_1, z_2, z_3 \in \tilde{z} + \eta B_Z$ and $z_i^* \in D_F f_i(z_i), i=1,2,3$, such that

$$\|z_1^* + z_2^* + z_3^*\| < 2\eta. \tag{4.4}$$

Since f_1 does not depend on u and v, we have $z_1^* = (\gamma^*, \theta^*, \alpha^*, 0, 0)$. Similarly, we may write $z_2^* = (0, 0, \beta^*, u^*, v^*)$ and $z_3^* = (0, 0, 0, q, p)$. Our next task is to calculate $z_i^*, i = 1, 2, 3$.

Step 3. *Calculating* z_1^*.

Let $z_1 = (\gamma, \theta, \alpha, u_1, v_1)$, and let g be a concave C^1 function on $R^m \times R^m \times R^n$ such that $g'(\gamma, \theta, \alpha) = (\gamma^*, \theta^*, \alpha^*)$ and $f_1 - g$ attains a minimum at (γ, θ, α). In particular, let $\gamma' = \phi(\alpha')$, and we have that

$$\|\phi(\alpha') - \theta'\| - g(\phi(\alpha'), \theta', \alpha') + \delta_{E \times \Theta}(\alpha', \theta') \tag{4.5}$$

attains a minimum at (θ, α).

We apply the sum rule of Theorem 2.4 and the chain rule of Theorem 2.5 to the functions in (4.5). With some straightforward (yet somewhat tedious) calculation we conclude that there exist $\theta_0 \in (\theta + \eta B_{R^m}) \cap \Theta$, $\alpha_1 \in \alpha + \eta B_{R^n}$, $\alpha_2 \in (\alpha + \eta B_{R^n}) \cap E$, and $\lambda \in N_F(\Theta, \theta_0)$ with $\|\lambda\| \in (1 - 3\eta, 1 + 3\eta)$ such that

$$\alpha^* \in D_F \langle \lambda, \phi \rangle (\alpha_1) + N_F(E, \alpha_2) + \eta B_{R^n}. \tag{4.6}$$

Step 4. *Calculating* z_2^*.

Let $z_2^* = (0, 0, \beta^*, u^*, v^*)$ and $z_2 = (\gamma_2, \theta_2, \beta, u, v)$. Then $(\beta^*, u^*, v^*) \in N_F(K, (\beta, u, v))$. The useful information for us is summarized in the following lemma.

LEMMA 4.1. *Let* $(\beta^*, u^*, v^*) \in N_F(K, (\beta, u, v))$. *Then*

$$\beta^* + v^*(t) + \int_t^1 u^*(s) ds = 0 \ a.e. \ in \ [0, 1].$$

Proof. Since K is convex, the Fréchet normal cone of K coincides with the convex normal cone. Thus,

$$\langle \beta' - \beta, \beta^* \rangle + \langle u' - u, u^* \rangle + \langle v' - v, v^* \rangle \leq 0 \ \forall (\beta', u', v') \in K.$$

By the definition of K we have that, for any $v' \in L^2([0, 1], R^n)$,

$$\left\langle \beta^*, \int_0^1 (v'(t) - v(t)) dt \right\rangle + \int_0^1 \left\langle u^*(t), \int_0^t (v'(s) - v(s)) ds \right\rangle dt$$
$$+ \int_0^1 \langle v^*(t), v'(t) - v(t) \rangle dt \leq 0.$$

Integration by parts yeilds that, for any $v' \in L^2([0, 1], R^n)$,

$$\int_0^1 \left\langle \beta^* + \int_t^1 u^*(s) ds + v^*(t), v'(t) - v(t) \right\rangle dt \leq 0.$$

Thus, $\beta^* + \int_t^1 u^*(s) ds + v^*(t) = 0$ a.e. in $[0, 1]$. □

Step 5. *Calculating* $N_F(C, (u, v))$.

Combining (4.4), (4.6), and Lemma 4.1, we conclude that, for any $\varepsilon > 0$, there exist $\gamma_\varepsilon \in \phi(x(1)) + \varepsilon B_{R^m}$, $\theta_0 \in (\theta + \eta B_{R^m}) \cap \overline{l(\gamma_\varepsilon)} \subset (\phi(x(1)) + \varepsilon B_{R^m}) \cap \overline{l(\gamma_\varepsilon)}$, $\alpha_1 \in \alpha + \eta B_{R^n} \subset x(1) + \varepsilon B_{R^n}$, $\alpha_2 \in (\alpha + \eta B_{R^n}) \cap E \subset (x(1) + \varepsilon B_{R^n}) \cap E$, $\lambda \in N_F(\overline{l(\gamma_\varepsilon)}, \theta_0)$ with $\|\lambda\| \in (1 - \varepsilon, 1 + \varepsilon)$, $\alpha_2^* \in N_F(E, \alpha_2)$, and

$$\alpha^* \in D_F \langle \lambda, \phi \rangle (\alpha_1) + \alpha_2^* + \varepsilon B_{R^n}, \tag{4.7}$$

such that there exist $(u,v)\in(x,\dot{x})+\varepsilon B_{L^2([0,1],R^n)\times L^2([0,1],R^n)}$, $(q,p)\in N_F(C,(u,v))$, and $u^*\in L^2([0,1],R^n)$ satisfying

$$\left\|\left(u^*,-\alpha^*-\int_{\cdot}^{1}u^*(s)ds\right)-(q,p)\right\|<\varepsilon. \tag{4.8}$$

Now we need to calculate the normal cone $N_F(C,(u,v))$. This calculation is similar to [10, Lemma 9.4].

Lemma 4.2. *Let* $(q,p)\in N_F(C,(u,v))$. *Then*

$$(-q(t),v(t))\in\partial_C H(u(t),p(t)) \ \textit{a.e. in}\ [0,1].$$

Proof. Since $(q,p)\in N_F(C,(u,v))$, there exists a C^1 concave function w on $L^2([0,1],R^n)\times L^2([0,1],R^n)$ with $w(u,v)=0$ and $w'(u,v)=0$ such that, for any $(u',v')\in C$,

$$\langle q,u'-u\rangle+\langle p,v'-v\rangle+w(u',v')\le 0. \tag{4.9}$$

Let $B:=(0,1)\cap\{$the Lebesgue points of u and $v\}$. Then B has measure 1. For any $t\in B$, $(x,\nu)\in$ Graph F, and $h>0$, let

$$u_h'(s):=\begin{cases}x & \text{if } s\in[t-h,t+h],\\ u(t) & \text{otherwise,}\end{cases}$$

and

$$v_h'(s):=\begin{cases}\nu & \text{if } s\in[t-h,t+h],\\ v(t) & \text{otherwise.}\end{cases}$$

Then $\|u_h'-u\|=O(h)$, $\|v_h'-v\|=O(h)$, and $w(u_h',v_h')=o(h)$. Setting $(u',v')=(u_h',v_h')$ in (4.9), dividing by $2h$, and taking limits yields

$$\langle q(t),x-u(t)\rangle+\langle p(t),\nu-v(t)\rangle\le 0\ \forall(x,\nu)\in\text{Graph } F. \tag{4.10}$$

In particular, setting $x=u(t)$, we have

$$\langle p(t),\nu-v(t)\rangle\le 0\ \forall\nu\in F(u(t)). \tag{4.11}$$

That is to say

$$\langle p(t),v(t)\rangle=\sup_{\nu\in F(u(t))}\langle p(t),\nu\rangle=H(u(t),p(t)). \tag{4.12}$$

Let

$$\begin{aligned}g(x,p):=&\ \langle p(t)-p,v(t)\rangle+\|p(t)-p\|^2\\ &+\langle q(t),x-u(t)\rangle+H(x,p).\end{aligned}$$

Then g is Lipschitz and strictly convex in p for each x. Let U be a ball around $u(t)$, and let K be a uniform bound for $F(x)$ over U. Then $|H(x,p)|\le K\|p\|$ for $x\in U$. Thus for all $x\in U$, the function $p\to g(x,p)$ attains a unique minimum at $p=p(x)$ and $\|p(x)\|\le c$ for some constant c. We claim that

(i) $p\to g(u(t),p)$ attains a local minimum at $p=p(t)$, and

(ii) $x\to\min_p g(x,p)$ attains a local maximum at $x=u(t)$.

Then it follows from [10, Lemma 9.5] that $(0,0) \in \partial_C g(u(t), p(t))$, i.e.,

$$(-q(t), v(t)) \in \partial_C H(u(t), p(t)).$$

It remains to verify claims (i) and (ii). By the minimax theorem we have

$$\begin{aligned}\min_p g(x,p) &= \min_p \{\langle p(t)-p, v(t)\rangle + \|p(t)-p\|^2 \\ &\qquad + \langle q(t), x-u(t)\rangle + H(x,p)\} \\ &= \min_p \max_{\nu \in F(x)} \{\langle p(t)-p, v(t)\rangle + \|p(t)-p\|^2 \\ &\qquad + \langle q(t), x-u(t)\rangle + \langle p, \nu\rangle\} \\ &= \max_{\nu\in F(x)} \min_p \{\langle p, \nu - v(t)\rangle + \|p(t)-p\|^2 \\ &\qquad + \langle q(t), x-u(t)\rangle + \langle p(t), v(t)\rangle\} \\ &= \max_{\nu \in F(x)} \{\langle p(t), \nu - v(t)\rangle - \|\nu - v(t)\|^2/4 \\ &\qquad + \langle q(t), x-u(t)\rangle + \langle p(t), v(t)\rangle\}.\end{aligned} \tag{4.13}$$

In particular, when $x = u(t)$, we have, by (4.11) and (4.12),

$$\begin{aligned}\min_p g(u(t),p) &= \max_{\nu \in F(u(t))} \{\langle p(t), \nu - v(t)\rangle - \|\nu - v(t)\|^2/4 + \langle p(t), v(t)\rangle\} \\ &= \langle p(t), v(t)\rangle = g(u(t), p(t)).\end{aligned}$$

This verifies (i). On the other hand, combining (4.10) and (4.13), we have

$$\min_p g(x,p) \le \langle p(t), v(t)\rangle = g(u(t),p(t)) = \min_p g(u(t),p),$$

which verifies (ii).

Step 6. *Taking limits.*

Let $\varepsilon = 1/k$ for $k = 1, 2, \ldots$. By (4.7) and (4.8) there exist sequences $\gamma^k, \theta_0^k \to \phi(x(1))$, $\alpha_1^k, \alpha_2^k \to x(1)$, $\lambda^k \in N_F(\overline{l(\gamma^k)}, \theta_0^k)$ with $\|\lambda^k\| \to 1$, $\alpha_2^{*k} \in N_F(E, \alpha_2^k)$,

$$\alpha^{*k} \in D_F\langle\lambda^k, \phi\rangle(\alpha_1^k) + \alpha_2^{*k} + (1/k)B_{R^n}, \tag{4.14}$$

$(u^k, v^k) \to (x, \dot{x})$ in $L^2([0,1], R^n) \times L^2([0,1], R^n)$, $(q^k, p^k) \in N_F(C, (u^k, v^k))$, and $u^{*k} \in L^2([0,1], R^n)$ such that

$$\left\| \left(u^{*k}, -\alpha^{*k} - \int_\cdot^1 u^{*k}(s)ds \right) - (q^k, p^k) \right\| < 1/k. \tag{4.15}$$

We consider the limiting processes for the following two cases.

The Good Case: $\|\alpha_2^{*k}\|$ is bounded. Passing to a subsequence, we may assume that α_2^{*k} converges to $\alpha_2^* \in N(E, x(1))$. Since ϕ is Lipschitzian and $\|\lambda^k\| \to 1$, taking subsequences if necessary, we may assume that α^{*k} converges to

$$\alpha^* \in \partial\langle\lambda, \phi\rangle(x(1)) + N(E, x(1)),$$

where $\lambda \in N(\overline{l(\phi(x(1)))}, \phi(x(1)))$ and $\|\lambda\| = 1$ by (A3).

By Lemma 4.2, $(q^k, p^k) \in N_F(C, (u^k, v^k))$ implies that

$$(-q^k(t), v^k(t)) \in \partial_C H(u^k(t), p^k(t)) \text{ a.e. in } [0,1]. \tag{4.16}$$

Since F is Lipschitz of rank L_F, $H(u,p)$ is Lipschitz with respect to u of rank $L_F\|p\|$. It follows from (4.16) that

$$\|q^k(t)\| \le l\|p^k(t)\|. \tag{4.17}$$

Combining (4.15) and (4.17), we have

$$\|u^{*k}(t)\| \le L_F \left(\int_t^1 \|u^{*k}(s)\|ds + \|\alpha^{*k}\| + 2/k \right).$$

Invoking Gronwall's inequality, we may conclude that $u^{*k}(t)$ is uniformly bounded on $[0,1]$, and, therefore, u^{*k} is a bounded sequence in $L^2([0,1],R^n)$. Again, taking a subsequence if necessary, we may assume that u^{*k} converges weakly to, say, q, in $L^2([0,1],R^n)$. Then, by (4.15), q^k weakly converges to q and p^k strongly converges to $p = -\alpha^* - \int_\cdot^1 q(s)ds$ in $L^2([0,1],R^n)$. Taking limits in (4.16) as $k \to \infty$ yields

$$(-\dot{p}(t), \dot{x}(t)) \in \partial_C H(x(t), p(t)) \text{ a.e. in } [0,1].$$

It is obvious that $-p(1) = \alpha^* \in \partial\langle\lambda,\phi\rangle(x(1)) + N(E,x(1))$. Thus, we derived the necessary condition in Theorem 3.2 corresponding to the case when $\lambda_0 = 1$.

The Bad Case: $\|\alpha_2^{*k}\|$ is unbounded. Without loss of generality, we may assume that $\|\alpha_2^{*k}\| \to \infty$. Dividing sequences $\alpha^{*k}, \alpha_2^{*k}, u^{*k}, q^k$, and p^k by $\|\alpha_2^{*k}\|$ and taking limits as before yield that there exists an absolutely continuous function p satisfying

$$(-\dot{p}(t), \dot{x}(t)) \in \partial_C H(x(t), p(t)) \text{ a.e. in } [0,1]$$

with $-p(1) = \alpha^* = \lim_{k\to\infty} \alpha^{*k}/\|\alpha_2^{*k}\| = \lim_{k\to\infty} \alpha_2^{*k}/\|\alpha_2^{*k}\| \in N(E,x(1))$. Observing that $\|\alpha^*\| = 1$, we have $\|p(t)\| > 0$ for all $t \in [0,1]$. This corresponds to the necessary condition in Theorem 3.2 when $\lambda_0 = 0$.

Finally, we observe that if $x(1) \in \text{int } E$, then when k is sufficiently large $\alpha_2^{*k} = 0$, so that the good case always applies. □

Acknowledgments. I thank Yu. S. Ledyaev and R. B. Vinter for helpful discussions and J. Warga for his help in improving the exposition.

REFERENCES

[1] J. Borwein, *Proper efficient points for maximizations with respect to cones*, SIAM J. Control Optim., 15 (1977), pp. 57–63.

[2] J. M. Borwein and A. Ioffe, *Proximal analysis in smooth spaces*, Set-Valued Anal., 4 (1996), pp. 1–24.

[3] J. M. Borwein, J. S. Treiman, and Q. J. Zhu, *Necessary conditions for constrained optimization problems with semicontinuous and continuous data*, Trans. Amer. Math. Soc., 350 (1998), pp. 2409–2429.

[4] J. M. Borwein and Q. J. Zhu, *Viscosity solutions and viscosity subderivatives in smooth Banach spaces with applications to metric regularity*, SIAM J. Control Optim., 34 (1996), pp. 1568–1591.

[5] J. M. Borwein and Q. J. Zhu, *A survey of subdifferential calculus with applications*, Nonlinear Anal., 38 (1999), pp. 687–773.

[6] W. W. Breckner *Derived sets for weak multiobjective optimization problems with state and control variables*, J. Optim. Theory Appl., 93 (1997), pp. 73–102.

[7] F. H. Clarke, *Necessary Conditions for Nonsmooth Problems in Optimal Control and the Calculus of Variations*, Ph.D. thesis, University of Washington, Seattle, WA, 1973.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983, Russian edition MIR, Moscow, 1988. Classics in Applied Mathematics 5, SIAM, Philadelphia, 1990.

[9] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conference Series in Applied Mathematics 57, SIAM, Philadelphia, 1989.

[10] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.

[11] B. D. Craven, *Nonsmooth multiobjective programming*, Numer. Funct. Anal. Optim., 10 (1989), pp. 49–64.

[12] G. Debreu, *Theory of Value*, John Wiley and Sons, New York, 1959.

[13] G. Debreu, *Mathematical Economics: Twenty Papers of Debreu*, Cambridge University Press, Cambridge, UK, 1983, pp. 163–172.

[14] J. Dong, *Nondifferentiable multiobjective optimization*, Adv. Math. (China), 23 (1994), pp. 517–528.

[15] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[16] Z. Hu, S. E. Salcudean, and P. D. Loewen, *Multiple objective control problems via nonsmooth analysis*, Optimal Control Appl. Methods, 19 (1998), pp. 411–422.

[17] A. D. Ioffe, *Calculus of Dini subdifferentials of functions and contingent derivatives of set-valued maps*, Nonlinear Anal., 8 (1984), pp. 517–539.

[18] A. D. Ioffe, *Euler-Lagrange and Hamiltonian formalisms in dynamic optimization*, Trans. Amer. Math. Soc., 349 (1997), pp. 2871–2900.

[19] A. D. Ioffe and R. T. Rockafellar, *The Euler and Weierstrass conditions for nonsmooth variational problems*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 59–87.

[20] P. D. Loewen and R. T. Rockafellar, *New necessary conditions for the generalized problem of Bolza*, SIAM J. Control Optim., 34 (1996), pp. 1496–1511.

[21] B. Kaskosz, *A maximum principle in relaxed controls*, Nonlinear Anal., 14 (1990), pp. 357–367.

[22] B. Kaskosz and S. Lojasiewicz, Jr., *A maximum principle for generalized control systems*, Nonlinear Anal., 9 (1985), pp. 109–130.

[23] B. Kaskosz and S. Lojasiewicz, Jr., *Lagrange-type extremal trajectories in differential inclusions*, Systems Control Lett., 19 (1992), pp. 241–247.

[24] A. Y. Kruger and B. S. Mordukhovich, *Extremal points and Euler equations in nonsmooth optimization*, Dokl. Akad. Nauk. BSSR, 24 (1980), pp. 684–687 (in Russian).

[25] M. Minami, *Weak Pareto-optimal necessary conditions in a nondifferential multiobjective program on a Banach space*, J. Optim. Theory Appl., 41 (1983), pp. 451–461.

[26] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.

[27] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian). (English translation to appear in Wiley-Interscience.)

[28] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[29] C. Singh, *Optimality conditions in multiobjective differentiable programming*, J. Optim. Theory Appl., 53 (1987), pp. 115–123.

[30] H. J. Sussmann, *A strong maximum principle for systems of differential inclusions*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 1809–1814.

[31] H. J. Sussmann, *Transversality conditions and a strong maximum principle for systems of differential inclusions*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 1–6.

[32] J. S. Treiman, *Lagrange multipliers for nonconvex generalized gradients with equality, inequality, and set constraints*, SIAM J. Control Optim., 37 (1999), pp. 1313–1329.

[33] H. D. Tuan, *On controllability extremality in nonconvex differential inclusions*, J. Optim. Theory Appl., 85 (1995), pp. 435–472.

[34] R. B. Vinter and H. Zheng, *Necessary conditions for optimal control problems with state constraints*, Trans. Amer. Math. Soc., 350 (1998), pp. 1181–1204.

[35] L. Wang, J. Dong, and Q. Liu, *Optimality conditions in nonsmooth multiobjective programming*, Systems Sci. Math. Sci., 7 (1994), pp. 250–255.

[36] L. Wang, J. Dong, and Q. Liu, *Nonsmooth multiobjective programming*, Systems Sci. Math. Sci., 7 (1994), pp. 362–366.

[37] J. Warga, *Optimization and controllability without differentiability assumptions*, SIAM J. Control Optim., 21 (1983), pp. 837–855.

[38] J. Warga, *An extension of the Kaskosz maximum principle*, Appl. Math. Optim., 22 (1990), pp. 61–74.

[39] X. Q. Yang and V. Jeyakumar, *First and second-order optimality conditions for convex composite multiobjective optimization*, J. Optim. Theory Appl., 95 (1997), pp. 209–224.

[40] M. Ying, *The nondominated solution and the proper efficient of nonsmooth multiobjective programming*, J. Systems Sci. Math. Sci., 5 (1985), pp. 269–278.

[41] Q. J. Zhu, *Necessary optimality conditions for nonconvex differential inclusion with endpoint constraints*, J. Differential Equations, 124 (1996), pp. 186–204.

[42] Q. J. Zhu, *Optimal control problem and nonsmooth analysis*, in Proceedings of the 38th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 15–18.

SIAM J. Control Optim.
Vol. 39, No. 1, pp. 113–132

ADAPTIVE FINITE ELEMENT METHODS FOR OPTIMAL CONTROL OF PARTIAL DIFFERENTIAL EQUATIONS: BASIC CONCEPT[*]

ROLAND BECKER[†], HARTMUT KAPP[†], AND ROLF RANNACHER[†]

Abstract. A new approach to error control and mesh adaptivity is described for the discretization of optimal control problems governed by elliptic partial differential equations. The Lagrangian formalism yields the first-order necessary optimality condition in form of an indefinite boundary value problem which is approximated by an adaptive Galerkin finite element method. The mesh design in the resulting reduced models is controlled by residual-based a posteriori error estimates. These are derived by duality arguments employing the cost functional of the optimization problem for controlling the discretization error. In this case, the computed state and costate variables can be used as sensitivity factors multiplying the local cell-residuals in the error estimators. This results in a generic and simple algorithm for mesh adaptation within the optimization process. This method is developed and tested for simple boundary control problems in semiconductor models.

Key words. optimal control problem, finite elements, a posteriori error estimates, mesh adaptation, model reduction

AMS subject classifications. 65K10, 65N30, 49K20

PII. S0363012999351097

1. Introduction.

In this article, we develop an adaptive Galerkin finite element method for optimal control problems governed by elliptic partial differential equations. The main goal is the derivation of a posteriori error estimates as basis for guiding the mesh adaptation and for controlling the error in this model reduction. The problems considered have the form

$$J(u,q) \;\to\; \min!, \qquad A(u) = f + B(q), \tag{1.1}$$

where A is an elliptic differential operator for the state variable u, B an impact operator for the control variable q, and J is a cost functional. As prototypical examples, we will consider problems of boundary control in semiconductor models. Our approach utilizes the classical Lagrangian framework for reformulating the optimal control problem (1.1) as a boundary value problem for stationary points of the associated first-order necessary optimality condition. Introducing the Lagrangian functional

$$\mathcal{L}(u,\lambda,q) := J(u,q) + \langle \lambda, A(u,q) - B(q) - f\rangle, \tag{1.2}$$

with the costate variable λ (Lagrangian multiplier), the solutions of (1.1) are among the stationary points of $\mathcal{L}$, determined by the system of equations

$$\nabla\mathcal{L}(u,\lambda,q) = 0. \tag{1.3}$$

[*] Received by the editors January 27, 1999; accepted for publication (in revised form) May 3, 2000; published electronically August 17, 2000. This work was supported by the Deutsche Forschungsgemeinschaft (DFG), SFB 359 "Reaktive Strömungen, Diffusion und Transport," Universität Heidelberg.

http://www.siam.org/journals/sicon/39-1/35109.html

[†] Institut für Angewandte Mathematik, INF 294, Universität Heidelberg, D-69120 Heidelberg, Germany (roland.becker@iwr.uni-heidelberg.de, hartmut.kapp@iwr.uni-heidelberg.de, rolf.rannacher@iwr.uni-heidelberg.de).

We use a standard finite element method for discretizing this saddle-point problem which results in finite dimensional problems

$$\nabla\mathcal{L}(u_h, \lambda_h, q_h) = 0, \tag{1.4}$$

for the "discrete" states u_h, controls q_h, and costates λ_h. As long as the discretization procedure uses a pure Galerkin approach the discrete problem actually corresponds to a formulation of the original minimization problem on the discrete state space. Since discretization in partial differential equations is expensive, at least for praxis-relevant models, the question of how this "model reduction" affects the quality of the optimization result is crucial for a cost-efficient computation. The need for a posteriori error control is therefore evident.

The discretization of the state equation generally leads to approximate solutions $\{u_h, q_h\}$ which are *not admissible* in the strict sense for the original constrained minimization problem. Let S denote the solution operator which associates the state variable $u = u(q)$ to a given control function. The *optimal* control minimizes the functional $j(q) := J(S(q), q)$ for all controls. Then, discretization of the state equation also changes the functional. Denoting by S_h the discrete solution operator, the discrete optimal control q_h solves

$$j_h(q_h) := J(S_h(q_h), q_h) \ \rightarrow \ \min!. \tag{1.5}$$

If we want to perform numerical computation with controlled accuracy, we have to substitute the notion of an "admissible solution" by an error estimate for the state equation. Of course, the distance between the numerical and the exact solution should be measured with respect to the specific needs of the optimization problem, i.e., its effect on the functional to be minimized. This asks for a sensitivity analysis for the optimization problem with respect to perturbations in the state equation, particularly perturbations resulting from discretization. In this sense, our a posteriori error estimation aims to control the error due to replacing the infinite dimensional problem (1.1) by its finite dimensional analogue. The crucial question is now which quality measure is appropriate for controlling the discretization error. In general, forcing this error to be small uniformly in the whole computational domain, as is often required in ODE models, is not feasible for partial differential equations. Therefore, we need to develop control of the discretization error in accordance with the sensitivity properties of the optimization problem.

Our approach to this problem uses the general method developed in [3] and [4] for error control in the Galerkin finite element discretization of differential equations of the general form $A(u) = f$. Employing the linearized dual problem

$$A'(u_h)^* z = F(\cdot) \tag{1.6}$$

for an arbitrary output functional F, an a posteriori error estimate

$$|F(u) - F(u_h)| \leq \eta_\omega(u_h) := \sum_{T\in\mathbb{T}_h} \rho_T(u_h)\, \omega_T(z) \tag{1.7}$$

can be derived. On a computational mesh $\mathbb{T}_h = \{T\}$ consisting of cells T, the local consistency errors, expressed in terms of residuals $\rho_T(u_h)$, are multiplied by weights $\omega_T(z)$ involving the "dual solution" z. These weights describe the dependence of the error on variations of the local residuals, i.e., on the local mesh size. In general the estimate (1.7) has to be approximated by numerically solving the dual problem (1.6). This results in a feedback process for generating successively more and more accurate error bounds and solution-adapted meshes. In applying this approach to saddle-point problems arising from optimal control problems, it seems natural to base the error control on the given cost functional, i.e., to choose $F := J$. In this particular case the corresponding approximate dual solution can be expressed in terms of the computed solution $\{u_h, \lambda_h, q_h\}$. Hence, the evaluation of the corresponding a posteriori error estimate

$$|J(u,q) - J(u_h,q_h)| \le \eta_\omega(u_h,\lambda_h,q_h) \tag{1.8}$$

does not require much extra work and a posteriori error estimation is almost for free. This leads to a generic and simple strategy for mesh adaptation in discretizing optimal control problems.

It may be seen as a drawback that in this approach the accuracy in the discretization of the state equation is only controlled with respect to its effect on the cost functional. This can lead to discrete models which approximate the original optimization problem with minimal cost but the obtained discrete states and controls are “admissible” only in a very weak sense, possibly insufficient for the particular application. If satisfaction of the state equation is desired in a stronger sense, we can combine our method with traditional “energy-error control” leading to an a posteriori error estimate of the form

$$|J(u,q) - J(u_h,q_h)| + \beta\,\|A(u_h) - f - B(q_h)\|_* \;\le\; \eta_{\omega,E}(u_h,\lambda_h,q_h), \tag{1.9}$$

where $\|\cdot\|_*$ denotes the dual of the natural “energy norm” corresponding to the operator $A'(u)$, and β is a tuning factor. For a discussion of adaptive finite element methods using residual-based a posteriori error estimates, we refer to the survey papers [1], [14], [8], and [12].

First, we develop our approach within a general setting in order to abstract from inessential technicalities. Then, all steps are made concrete for a linear model problem of boundary control. Despite its simplicity this problem represents the main structure of optimal control and is chosen in order to clarify the idea underlying the proposed procedure. Some numerical results illustrate the main features of the adaptive algorithm particularly in comparison to more conventional methods based on global error control for the state equation. At the end, we extend our method to problems with nonlinear state equations with an example of boundary control in semiconductor models.

2. A linear model situation. We consider an abstract setting for optimal control: Let Q, V, and H be Hilbert spaces for the control variable $q \in Q$, the state variable $u \in V$, and given observations $c_0 \in H$. The inner product and norm of H are $(\cdot,\cdot)$ and $\|\cdot\|$, respectively. The state equation is given in the form

$$a(u,\varphi) = (f,\varphi) + b(q,\varphi) \quad \forall \varphi \in V, \tag{2.1}$$

where the bilinear form $a(\cdot,\cdot)$ represents a linear elliptic operator and the bilinear form $b(\cdot,\cdot)$ expresses the action of the control. The goal is to minimize the cost functional

$$J(u,q) = \tfrac{1}{2}\|cu - c_0\|^2 + \tfrac{1}{2}n(q,q), \tag{2.2}$$

where $c : V \to H$ is a linear bounded observation operator. For simplicity, we assume that $a(\cdot,\cdot)$ and $n(\cdot,\cdot)$ induce norms on the spaces V and Q denoted by $\|\cdot\|_a$ and $\|\cdot\|_n$, respectively. This guarantees the existence of a unique solution of the optimal control problem and the classical regularity theory for elliptic equations applies (see, e.g., [11]).

Introducing a Lagrangian parameter $\lambda \in V$ and the Lagrangian function $\mathcal{L}(u,q,\lambda)$,

$$\mathcal{L}(u,q,\lambda) := J(u,q) + a(u,\lambda) - b(q,\lambda) - (f,\lambda),$$

the first-order necessary conditions (Euler–Lagrange equations) of the optimization problem,

$$\nabla\mathcal{L}(u,q,\lambda)(v,\mu,r) = 0 \quad \forall\{v,\mu,r\} \in V \times V \times Q \tag{2.3}$$

have the explicit form

$$\begin{aligned} a(v,\lambda) + (cu - c_0, cv) &= 0 \quad \forall v \in V, \\ a(u,\mu) - b(q,\mu) &= (f,\mu) \quad \forall \mu \in V, \\ -b(r,\lambda) + n(q,r) &= 0 \quad \forall r \in Q. \end{aligned} \tag{2.4}$$

This system has the usual saddle-point structure

$$\begin{aligned} (cu,cv) + a(v,\lambda) &= (c_0,cv) \quad \forall v \in V, \\ a(u,\mu) - b(q,\mu) &= (f,\mu) \quad \forall \mu \in V, \\ -b(r,\lambda) + n(q,r) &= 0 \quad \forall r \in Q. \end{aligned} \tag{2.5}$$

Introducing operators A, B, C, N which represent the corresponding bilinear forms, system (2.5) can also be written in matrix form as

$$\begin{bmatrix} C & A^T & 0 \\ A & 0 & -B \\ 0 & -B^T & N \end{bmatrix} \begin{bmatrix} u \\ \lambda \\ q \end{bmatrix} = \begin{bmatrix} c_0 \\ f \\ 0 \end{bmatrix}. \tag{2.6}$$

Below, we will consider the following realization of the foregoing abstract setting which represents the case of an elliptic linear state equation subjected to boundary control. Let $\Omega \subset \mathbb{R}^2$ be an open bounded domain with Lipschitz boundary $\partial\Omega$ which is decomposed into a homogeneous Neumann part Γ_N and a control part Γ_C on which the control acts ($\partial\Omega = \Gamma_C \cup \Gamma_N$),

$$\begin{gathered} -\Delta u + u = f \quad \text{in } \Omega, \\ \partial_n u = 0 \ \text{ on } \Gamma_N, \quad \partial_n u = q \ \text{ on } \Gamma_C. \end{gathered} \tag{2.7}$$

The observations are given on a part Γ_O of the boundary and the associated cost functional is

$$J(u,q) = \tfrac{1}{2}\|u - c_0\|^2_{\Gamma_O} + \tfrac{\alpha}{2}\|q\|^2_{\Gamma_C} \tag{2.8}$$

with a regularization parameter $\alpha > 0$. In this case the natural choice for the function spaces is $V = H^1(\Omega)$, the first-order Sobolev Hilbert-space over Ω, and $H = L^2(\Gamma_O)$, $Q = L^2(\Gamma_C)$, the usual Lebesgue Hilbert-spaces over Γ_C and Γ_O, respectively. The bilinear forms $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $n(\cdot,\cdot)$ are given by

$$a(u,v) = (\nabla u, \nabla v)_\Omega + (u,v)_\Omega, \quad b(q,v) = (q,v)_{\Gamma_C}, \quad n(q,r) = \alpha(q,r)_{\Gamma_C},$$

where $(\cdot,\cdot)_\Sigma$ denotes the L^2-inner product on the set Σ. The operator c in the cost functional is the trace operator, $cu = u_{|\Gamma_O}$. Then, the necessary optimality condition $\nabla\mathcal{L}(u,\lambda,q) = 0$ reads as follows:

$$\begin{aligned} (u,v)_{\Gamma_O} - (c_0,v)_{\Gamma_O} + (\nabla v,\nabla\lambda)_\Omega + (v,\lambda)_\Omega &= 0 \quad \forall v \in V,\\ (\nabla u,\nabla\mu)_\Omega + (u,\mu)_\Omega - (f,\mu)_\Omega - (q,\mu)_{\Gamma_C} &= 0 \quad \forall \mu \in V,\\ \alpha(q,r)_{\Gamma_C} - (\lambda,r)_{\Gamma_C} &= 0 \quad \forall r \in Q. \end{aligned} \tag{2.9}$$

3. A priori error estimate.

3. A priori error estimate. For simplicity of notation, we introduce the product space $X = V\times V\times Q$, with elements of the form $x = \{u,\lambda,q\}$, which is equipped with the natural norm

$$\|x\|_X := \big(\|u\|_a^2 + \|\lambda\|_a^2 + \|q\|_n^2\big)^{1/2}.$$

Furthermore, we define a bilinear form $A(\cdot,\cdot)$ on X by

$$A(x,y) = A(\{u,\lambda,q\},\{v,\mu,r\}) := (cu,c\mu) + a(u,v) - b(q,v) + a(\mu,\lambda) - b(r,\lambda) + n(q,r).$$

Using this notation, system (2.5) can be written in compact form as

$$A(x,y) = F(y) \quad \forall y \in X, \tag{3.1}$$

with the linear functional $F(\cdot)$ defined by

$$F(y) = F(\{v,\mu,r\}) := (c_0,c\mu) + (f,v).$$

In order to simplify the analysis, we impose the following conditions:

$$|A(x,y)| \le c_A\|x\|_X\|y\|_X, \tag{3.2}$$

$$|b(r,v)| \le c_b\|r\|_n\|v\|_a. \tag{3.3}$$

The second condition, which relies on the presence of the regularization term $n(\cdot,\cdot)$ (requiring that $\alpha > 0$ in the above example), is rather strong. It can be substituted by an "inf-sup" condition for $b(\cdot,\cdot)$ under which the regularization could be omitted; see Remark 3.1, below. The bilinear form $A(\cdot,\cdot)$ satisfies the following stability condition.

PROPOSITION 3.1. *Under the assumptions* (3.2) *and* (3.3)*, there exists a constant* $\gamma > 0$ *such that*

$$\inf_{x\in X}\left\{\sup_{y\in X}\frac{A(x,y)}{\|x\|_X\|y\|_X}\right\} \ge \gamma. \tag{3.4}$$

Proof. For any fixed $x = \{u,\lambda,q\}$, we choose the test triple $y = \{v,\mu,r\} := \{u,\lambda,q\}$ to obtain

$$\begin{aligned} A(x,y) &= \|cu\|^2 + \|u\|_a^2 + \|\lambda\|_a^2 + \|q\|_n^2 - b(q,\lambda) - b(q,u)\\ &\ge \|cu\|^2 + \|u\|_a^2 + \|\lambda\|_a^2 + \|q\|_n^2 - \tfrac14\|q\|_n^2 - \tfrac34\|\lambda\|_a^2 - \tfrac14\|q\|_n^2 - \tfrac34\|u\|_a^2\\ &\ge \|cu\|^2 + \tfrac14\|u\|_a^2 + \tfrac14\|\lambda\|_a^2 + \tfrac12\|q\|_n^2. \end{aligned}$$

We conclude the asserted estimate by noting that $\|y\| = \|x\|$. □

We consider the discretization of the variational equation (3.1) by a standard Galerkin method using trial spaces $X_h := V_h\times V_h\times Q_h \subset X$. For each $x \in X$, there shall exist an interpolation $i_h x \in X_h$, such that $\|x - i_h x\|_X \;\to\; 0 \; (h \to 0)$. Then, approximations $x_h \in X_h$ are determined by

$$A(x_h, y_h) = F(y_h) \quad \forall y_h \in X_h. \tag{3.5}$$

This discretization is automatically stable since a discrete analogue of (3.4) is fulfilled by the same argument as used above,

$$\inf_{x_h \in X_h} \left\{ \sup_{y_h \in X_h} \frac{A(x_h, y_h)}{\|x_h\|_X \|y_h\|_X} \right\} \geq \gamma > 0. \tag{3.6}$$

Combining (3.5) and (3.1), we get the Galerkin orthogonality

$$A(x - x_h, y_h) = 0, \quad y_h \in X_h. \tag{3.7}$$

This leads us to the following abstract a priori error estimate.

Proposition 3.2. *For the Galerkin approximation in $X_h \subset X$, there holds*

$$\begin{aligned} \|u - u_h\|_a + &\|\lambda - \lambda_h\|_a + \|q - q_h\|_n \\ &\leq c\Big\{ \inf_{\mu_h \in V_h} \|u - \mu_h\|_a + \inf_{v_h \in V_h} \|\lambda - v_h\|_a + \inf_{r_h \in Q_h} \|q - r_h\|_n \Big\}. \end{aligned} \tag{3.8}$$

Proof. The stability estimate (3.6) implies that

$$\gamma \|i_h x - x_h\| \leq \sup_{y_h \in X_h} \frac{A(i_h x - x_h, y_h)}{\|y_h\|_X} = \sup_{y_h \in X_h} \frac{A(i_h x - x, y_h)}{\|y_h\|_X} \leq c_A \|i_h x - x\|_X \,.$$

Here, we have used the Galerkin relation (3.7) and the continuity estimate (3.2). □

Remark 3.1. *Of course, more precise error estimates could be given exploiting the structure of the underlying problem. For instance, it may be possible to equip the space Q with a different norm than the one induced by $n(\cdot,\cdot)$, in order to get robustness with respect to the regularization. This requires us to replace* (3.3) *by the following weaker inf-sup condition*

$$\inf_{q \in Q} \left\{ \sup_{v \in V} \frac{b(q, v)}{\|v\|_a} \right\} \geq \kappa > 0.$$

We note that for the model example with boundary control and boundary observations given above the conditions (3.2) *and* (3.3) *are satisfied.*

4. A posteriori error estimate. In this section, we derive an a posteriori error estimate for the model control problem. As discussed above, the error estimator to be derived should control the value of the cost functional. First, we carry out the analysis in the abstract functional analytic setting. Recalling the definition of the Lagrange functional $\mathcal{L}$,

$$\mathcal{L}(x) = \mathcal{L}(u, \lambda, q) = J(u, q) + a(u, \lambda) - (f, \lambda) - b(q, \lambda),$$

the continuous and discrete optimal control solutions $x = \{u, \lambda, q\} \in X$ and $x_h = \{u_h, \lambda_h, q_h\} \in X_h$ satisfy

$$\nabla\mathcal{L}(x)(\varphi) = 0, \quad \varphi \in X, \qquad \nabla\mathcal{L}(x_h)(\varphi_h) = 0, \quad \varphi_h \in X_h. \tag{4.1}$$

This implies the Galerkin orthogonality relation

$$\nabla^2\mathcal{L}(x-x_h,\varphi_h)=0,\quad \varphi_h\in X_h, \tag{4.2}$$

for which it is essential that $\mathcal{L}$ is quadratic. Since the solutions u and u_h satisfy the corresponding state equations, the cost functional and the Lagrangian functional are related by

$$J(u,q)-J(u_h,q_h)=\mathcal{L}(x)-\mathcal{L}(x_h). \tag{4.3}$$

Further, there holds

$$\mathcal{L}(x)-\mathcal{L}(x_h)=\nabla\mathcal{L}(x)(x-x_h)-\tfrac{1}{2}\nabla^2\mathcal{L}(x-x_h,x-x_h). \tag{4.4}$$

The first term on the right-hand side vanishes since x is a stationary point of $\mathcal{L}$. Using the Galerkin orthogonality relation (4.2) in the second term, we obtain for arbitrary $\varphi_h\in X_h$ that

$$\nabla^2\mathcal{L}(x-x_h,x-x_h)=\nabla^2\mathcal{L}(x-x_h,x-x_h-\varphi_h)=-\nabla\mathcal{L}(x_h)(x-x_h-\varphi_h). \tag{4.5}$$

Therefore, choosing $\varphi_h=i_hx-x_h$, we get the following error representation.

PROPOSITION 4.1. *For the abstract model problem with linear state equation and quadratic cost functional, the following error identity holds:*

$$J(u,q)-J(u_h,q_h)=\tfrac{1}{2}\nabla\mathcal{L}(x_h)(x-i_hx). \tag{4.6}$$

In order to convert the abstract error identity (4.6) into a form which can be evaluated, we need to be more specific about the setting of the underlying problem and its discretization. As an example, we demonstrate this for the linear Neumann control problem described by (2.7) and (2.8). Here, the Galerkin finite element discretization of the saddle-point problem (2.9) uses subspaces $V_h\subset V=H^1(\Omega)$ and $Q_h\subset Q=L^2(\Gamma_C)$ of piecewise polynomial functions defined on regular decompositions $\mathbb{T}_h=\{T\}$ of the domain Ω into cells T (triangles or quadrilaterals); for a detailed description of such a setting see, e.g., Brenner and Scott [5]. Here, we use quadrilateral meshes where on each cell T the local shape functions are constructed by mapping bilinear functions defined on a reference square to the cell T. This ansatz is referred to as the "isoparametric" bilinear finite element. We assume that the space Q_h of discrete controls is given by the traces along Γ_C of the finite element functions of V_h. This is not necessary for our results but simplifies the notation. In order to avoid the technicalities caused by curved boundaries, we suppose the domain Ω to be polygonal. We use the notation $h_T:=diam(T)$ and $h_\Gamma:=diam(\Gamma)$ for the width of a cell $T\in\mathbb{T}_h$ and a corresponding cell edge $\Gamma\subset\partial T$. In order to ease local mesh refinement and coarsening, "hanging nodes" are allowed, but at most one per cell edge; see Figure 4.1. The degree of freedom at such a hanging node is eliminated by interpolation in order to keep the discretization "conforming" (see, e.g., [6] and [3]).

First, we turn (4.6) into a residual based a posteriori estimate as follows.

PROPOSITION 4.2. *For error control with respect to the cost functional J, there holds the weighted a posteriori error estimate*

$$|J(u,q)-J(u_h,q_h)|\ \le\ \eta_\omega(u_h,\lambda_h,q_h)=\sum_{T\in\mathbb{T}_h}\eta_T(u_h,\lambda_h,q_h), \tag{4.7}$$

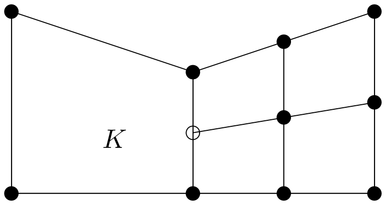


FIG. 4.1. *Quadrilateral mesh patch with a "hanging node."*

with the local error indicators

$$\eta_T(u_h,\lambda_h,q_h) := \rho_T^{(u)}\,\omega_T^{(\lambda)} + \rho_{\partial T}^{(u)}\,\omega_{\partial T}^{(\lambda)} + \rho_T^{(\lambda)}\,\omega_T^{(u)} + \rho_{\partial T}^{(\lambda)}\,\omega_{\partial T}^{(u)} + \rho_{\partial T}^{(q)}\,\omega_{\partial T}^{(q)}$$

and the cellwise residuals and weights

$$\begin{aligned}
\rho_T^{(u)} &:= \|R_h^{(u)}\|_T, & \omega_T^{(\lambda)} &:= \|\lambda - i_h\lambda\|_T,\\
\rho_{\partial T}^{(u)} &:= \|r_h^{(u)}\|_{\partial T}, & \omega_{\partial T}^{(\lambda)} &:= \|\lambda - i_h\lambda\|_{\partial T},\\
\rho_T^{(\lambda)} &:= \|R_h^{(\lambda)}\|_T, & \omega_T^{(u)} &:= \|u - i_h u\|_T,\\
\rho_{\partial T}^{(\lambda)} &:= \|r_h^{(\lambda)}\|_{\partial T}, & \omega_{\partial T}^{(u)} &:= \|u - i_h u\|_{\partial T},\\
\rho_{\partial T}^{(q)} &:= \|r_h^{(q)}\|_{\partial T\cap\Gamma_C}, & \omega_{\partial T}^{(q)} &:= \|q - j_h q\|_{\partial T\cap\Gamma_C}.
\end{aligned}$$

The "cell residuals" $R_h^{(u)}$, $R_h^{(\lambda)}$ and the "edge residuals" $r_h^{(u)}$, $r_h^{(\lambda)}$, $r_h^{(q)}$ are on cells T and cell edges Γ defined by

$$R_{h|T}^{(u)} := -\Delta u_h + u_h - f, \quad R_{h|T}^{(\lambda)} := -\Delta\lambda_h + \lambda_h, \qquad r_{h|\Gamma}^{(q)} := \alpha q_h - \lambda_h \ \text{ if } \Gamma\subset\Gamma_C,$$

$$r_{h|\Gamma}^{(u)} := \begin{cases} \frac{1}{2}h_\Gamma^{-1/2}[\partial_n\varphi_h] & \text{if } \Gamma\subset\partial T\setminus\partial\Omega,\\ h_\Gamma^{-1/2}\partial_n u_h & \text{if } \Gamma\subset\partial\Omega\backslash\Gamma_C,\\ h_\Gamma^{-1/2}(\partial_n u_h - q_h) & \text{if } \Gamma\subset\Gamma_C, \end{cases} \qquad r_{h|\Gamma}^{(\lambda)} := \begin{cases} \frac{1}{2}h_\Gamma^{-1/2}[\partial_n\varphi_h] & \text{if } \Gamma\subset\partial T\setminus\partial\Omega,\\ h_\Gamma^{-1/2}\partial_n\lambda_h & \text{if } \Gamma\subset\partial\Omega\backslash\Gamma_O,\\ h_\Gamma^{-1/2}(c_0 - u_h + \partial_n\lambda_h) & \text{if } \Gamma\subset\Gamma_O. \end{cases}$$

Here, $[\partial_n\varphi_h]$ denotes the jump of the normal derivative of φ_h across the interelement edges Γ, the boundary components Γ_C, Γ_O are the control and observation boundary, respectively, and i_h, j_h denote some local interpolation operators into the finite element spaces.

Proof. From the abstract error identity (4.6), we obtain that

$$\begin{aligned}
J(u,q) - J(u_h,q_h) &= \tfrac{1}{2}\nabla\mathcal{L}(x_h)(x - i_h x)\\
&= \tfrac{1}{2}(c_0 - u_h, u - i_h u)_{\Gamma_O} + \tfrac{1}{2}a(u - i_h u, \lambda_h)\\
&\quad + \tfrac{1}{2}a(u_h, \lambda - i_h\lambda) - \tfrac{1}{2}(f, \lambda - i_h\lambda) - \tfrac{1}{2}b(q_h, \lambda - i_h\lambda)\\
&\quad + \tfrac{1}{2}n(q_h, q - j_h q) - \tfrac{1}{2}b(q - j_h q, \lambda_h)\\
&=: I_\lambda + I_u + I_q.
\end{aligned}$$

Recalling the definition of the bilinear forms and integrating cellwise by parts, the first term I_λ is rewritten as follows:

$$\begin{aligned}
2I_\lambda &= (c_0-u_h,u-i_hu)_{\Gamma_O}+a(u-i_hu,\lambda_h)\\
&= (c_0-u_h,u-i_hu)_{\Gamma_O}+(\nabla(u-i_hu),\nabla\lambda_h)+(u-i_hu,\lambda_h)\\
&= (c_0-u_h+\partial_n\lambda_h,u-i_hu)_{\Gamma_O}+(u-i_hu,\partial_n\lambda_h)_{\partial\Omega\setminus\Gamma_O}\\
&\quad+\sum_{T\in\mathbb{T}_h}\Big\{(u-i_hu,-\Delta\lambda_h+\lambda_h)_T+(u-i_hu,\partial_n\lambda)_{\partial T\setminus\partial\Omega}\Big\}\\
&= (c_0-u_h+\partial_n\lambda_h,u-i_hu)_{\Gamma_O}+(u-i_hu,\partial_n\lambda_h)_{\partial\Omega\setminus\Gamma_O}\\
&\quad+\sum_{T\in\mathbb{T}_h}\Big\{(u-i_hu,-\Delta\lambda_h+\lambda_h)_T+\tfrac12(u-i_hu,[\partial_n\lambda])_{\partial T\setminus\partial\Omega}\Big\}.
\end{aligned}$$

Hence using the definition of the residuals $r_h^{(\lambda)}$ and $R_h^{(\lambda)}$, we find

$$\begin{aligned}
2I_\lambda &= (u-i_hu,r_h^{(\lambda)})_{\partial\Omega}+\sum_{T\in\mathbb{T}_h}\Big\{(u-i_hu,R_h^{(\lambda)})_T+(u-i_hu,r_h^{(\lambda)})_{\partial T\setminus\partial\Omega}\Big\}\\
&= \sum_{T\in\mathbb{T}_h}\Big\{(u-i_hu,R_h^{(\lambda)})_T+(u-i_hu,r_h^{(\lambda)})_{\partial T}\Big\},
\end{aligned}$$

and, consequently by the Cauchy–Schwarz inequality,

$$2|I_\lambda| \le \sum_{T\in\mathbb{T}_h}\Big\{\|u-i_hu\|_T\,\|R_h^{(\lambda)}\|_T+\|u-i_hu\|_{\partial T}\,\|r_h^{(\lambda)}\|_{\partial T}\Big\}.$$

In the same way, we get for the other terms I_λ and I_q:

$$\begin{aligned}
2I_u &= a(u_h,\lambda-i_h\lambda)-(f,\lambda-i_h\lambda)-b(q_h,\lambda-i_h\lambda)\\
&= (\nabla u_h,\nabla(\lambda-i_h\lambda))+(u_h,\lambda-i_h\lambda)-(f,\lambda-i_h\lambda)-(q_h,\lambda-i_h\lambda)_{\Gamma_C}\\
&= \sum_{T\in\mathbb{T}_h}\Big\{(-\Delta u_h+u_h-f,\lambda-i_h\lambda)_T+\tfrac12([\partial_nu_h],\lambda-i_h\lambda)_{\partial T\setminus\partial\Omega}\Big\}\\
&\quad+(\partial_nu_h,\lambda-i_h\lambda)_{\partial\Omega\setminus\Gamma_C}+(\partial_nu_h-q_h,\lambda-i_h\lambda)_{\Gamma_C}\\
&= \sum_{T\in\mathbb{T}_h}\Big\{(R_h^{(u)},\lambda-i_h\lambda)_T+(r_h^{(u)},\lambda-i_h\lambda)_{\partial T\setminus\partial\Omega}\Big\}\\
&\quad+(r_h^{(u)},\lambda-i_h\lambda)_{\partial\Omega\setminus\Gamma_C}+(r_h^{(u)},\lambda-i_h\lambda)_{\Gamma_C}\\
&= \sum_{T\in\mathbb{T}_h}\Big\{(R_h^{(u)},\lambda-i_h\lambda)_T+(r_h^{(u)},\lambda-i_h\lambda)_{\partial T}\Big\},
\end{aligned}$$

$$\begin{aligned}
2I_q &= n(q_h,q-j_hq)-b(q-j_hq,\lambda_h)\;=\;(\alpha q_h-\lambda_h,q-j_hq)_{\Gamma_C}\\
&= \sum_{\Gamma\subset\Gamma_C}(r_h^{(q)},q-j_hq)_\Gamma,
\end{aligned}$$

and, consequently,

$$\begin{aligned}
2|I_u| &\le \sum_{T\in\mathbb{T}_h}\Big\{\|R_h^{(u)}\|_T\,\|\lambda-i_h\lambda\|_T+\|r_h^{(u)}\|_{\partial T}\,\|\lambda-i_h\lambda\|\|_{\partial T}\Big\},\\
2|I_q| &\le \sum_{\Gamma\subset\Gamma_C}\|r_h^{(q)}\|_\Gamma\,\|q-j_hq\|_\Gamma\,.
\end{aligned}$$

Collecting these estimates implies the asserted result. □

REMARK 4.1. *We note that in the a posteriori error estimate* (4.7)*, the residual of the state equation is weighted by terms involving the adjoint variable* λ *from the original equation* (2.5)*. This has a natural interpretation as it is well known from sensitivity analysis that the adjoint variable is a measure for the influence of perturbations on the cost functional. Since discretization can be interpreted as a special kind of perturbation, the appearance of* λ *in the estimator is not surprising. The special form of the weights involving the interpolation* $i_h z$ *is a characteristic feature of the Galerkin discretization.*

REMARK 4.2. *The a posteriori error estimate* (4.7) *is easier to understand if one recalls the model situation of approximating the boundary value problem*

$$-\Delta u = f \ \ in\, \Omega, \quad u = 0 \ \ on\, \partial\Omega,$$

by a Galerkin finite element method. The natural variational formulation of this problem is equivalent to an unconstraint optimization problem, namely, the minimization of the "energy functional" $J(u) := \frac{1}{2}\|\nabla u\|_\Omega^2 - (f,u)_\Omega$ *over the solution space* $V := H_0^1(\Omega)$*. In this context, in view of the identity*

$$J(u) - J(u_h) = -\tfrac{1}{2}\|\nabla e\|_\Omega^2,$$

error control with respect to the energy functional is equivalent to control of the error in the energy norm, $\|\nabla e\|_\Omega$*. It is well known that the latter can be achieved without referring to an additional dual problem since in this case the corresponding dual solution coincides with the error* e *itself; see* [3] *for a discussion of energy-error control in the context of "duality techniques."*

Evaluation of error estimators. The a posteriori error estimate (4.7) still involves the *continuous* solutions $\{u, \lambda, q\}$. As proposed in [4], we use the computed solutions $\{u_h, \lambda_h, q_h\}$ for approximating the weights $\omega_{\partial T}^{(\cdot)}$ and $\omega_T^{(\cdot)}$. To this end, we recall the well-known local approximation properties of finite elements, e.g.,

$$\|u - i_h u\|_T + h_\Gamma^{1/2}\|u - i_h u\|_\Gamma \le c_I h_T\, \|\nabla^2 u\|_T, \tag{4.8}$$

where $\Gamma \subset \partial T$, and i_h is the generic operator of cellwise nodal interpolation into V_h, with interpolation constant usually in the range $c_I \sim 0.1 - 1$; for details of the interpolation theory for finite elements, we refer to [5]. Analogous estimates hold for the terms involving q and λ. Then, the derivatives are approximated by suitable difference quotients, e.g.,

$$\|\nabla^2 u\|_T \approx \|\nabla_h^2 u_h\|_T. \tag{4.9}$$

For a more detailed discussion of this evaluation of weights and some of its alternatives, we refer to [4]. We emphasize that the proposed procedure for evaluating the a posteriori error estimate (4.7) uses only information in terms of the already computed solution $\{u_h, \lambda_h, q_h\}$.

Error control for the state variable. The estimate (4.7) provides control of the error with respect to the cost functional which is the quantity of primary interest in the optimization problem. But this does not include control of the error in the state equation. Though the corresponding residuals $\rho_T^{(u)}$ and $\rho_{\partial T}^{(u)}$ are present in the error estimator, they are weighted according to their effect on the cost functional. In case that the discrete state u_h is required to be admissible in a stronger sense, the error estimate (4.7) can be extended to also include control of the error in satisfying the state equation measured in the natural energy norm

$$\|u\|_E := \left(\|u\|^2 + \|\nabla u\|^2\right)^{1/2}.$$

The standard a posteriori error analysis for the boundary value problem

$$-\Delta u + u = f \ \text{ in } \Gamma_C, \quad \partial_n u = 0 \ \text{ on } \Gamma_N, \quad \partial_n u = q \ \text{ on } \Gamma_C, \tag{4.10}$$

for frozen control q, yields the following bound for the "energy error":

$$\|u-u_h\|_E^2 \ \le\ \eta_E(u_h) := c \sum_{T\in\mathbb{T}_h} \left\{\rho_T^{(u)2} + \rho_{\partial T}^{(u)2}\right\}, \tag{4.11}$$

with the residuals as defined in Proposition 4.2. For the derivation of this estimate see, for example, [14], [4], and the literature cited therein. We see that all terms of $\eta_E(u_h)$ appear also in $\eta_\omega(u_h,\lambda_h,q_h)$, but are weighted in terms of the adjoint variable. The effect of this modification will be illustrated below by a numerical test. In practice, we may use a combination of our weighted error estimator and the energy-error estimator $\eta_{\omega,E}(u_h,\lambda_h,q_h) := \eta_\omega(u_h,\lambda_h,q_h) + \beta\eta_E(u_h)$, with a suitable weighting factor $\beta \ge 0$.

Strategies for mesh adaptation. Several strategies are possible for mesh adaptation on the basis of a posteriori error estimators η as developed above. Usually, a certain tolerance TOL for the error in the quantity $J(u,q)$ and an upper bound for the complexity of the discrete model, i.e. the maximum number of mesh cells $N_{\max}$, are given. It is assumed that an "optimal" mesh-size distribution is achieved if the local error indicators η_T are equilibrated over the mesh $\mathbb{T}_h$. This suggests use of the "error-balancing strategy"; i.e., we cycle through the mesh and try to equilibrate the local error indicators η_T according to $\eta_T \approx TOL/N$. This process requires iteration with respect to the number of mesh cells N and eventually results in $\eta \approx TOL$. However, this strategy may lead to very slow mesh refinement and is very delicate to use. More robust is the "fixed-fraction strategy" in which we order the cells according to the size of η_T and refine a certain threshold $X\%$ of cells with largest η_T (or those cells which contribute to a certain percentage of the error estimator η). A certain fraction $Y\%$ of cells with small η_T may be coarsened. By this strategy, one can achieve a prescribed rate of increase of N (or keep it constant as may be desirable in nonstationary computations). In the test computations described below the second version of the "Fixed-Fraction Strategy" has been used with threshold 30%.

5. Numerical results—linear case. We present a linear model problem as described in (2.7), where Ω is a T-shaped domain with maximum side length one; see Figure 5.1 (left). In this example the control acts along the lower boundary Γ_C, whereas the observation is taken along the upper boundary Γ_O. The cost functional is chosen as

$$J(u,q) := \tfrac{1}{2}\|u-c_0\|_{\Gamma_O}^2 + \tfrac{\alpha}{2}\|q\|_{\Gamma_C}^2,$$

with $c_0 \equiv 1.0$ and $\alpha = 1.0$. In this case, the regularization term $\frac{\alpha}{2}\|q\|_{\Gamma_C}^2$ may be viewed as part of the cost functional with its own physical meaning. We perform computations on a series of locally refined meshes. On each mesh, the system of the first-order necessary condition (2.5) is discretized by the Galerkin finite element

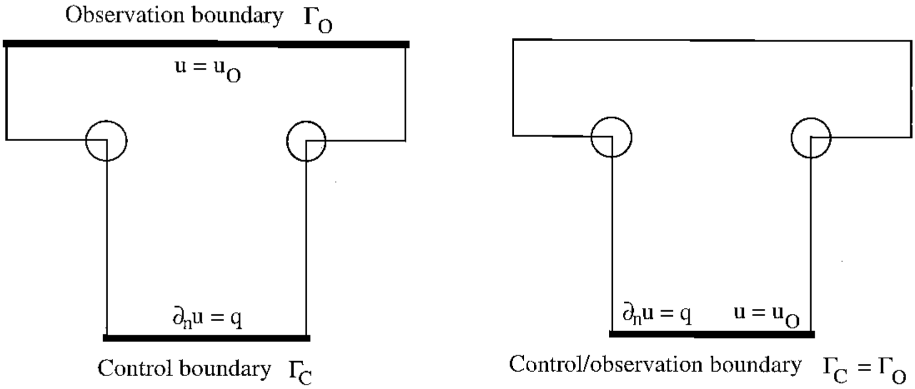

Observation boundary Γ_O
$u = u_O$
$\partial_n u = q$
Control boundary Γ_C
$\partial_n u = q$
$u = u_O$
Control/observation boundary $\Gamma_C = \Gamma_O$

FIG. 5.1. *Configuration of the boundary control model problem on a T-domain (Ginzburg–Landau model): configuration* 1 *(left), configuration* 2 *(right).*

TABLE 5.1
Linear test (*configuration* 1): *Efficiency of the weighted error estimator.*

N	320	1376	4616	11816	23624	48716
E_h	$1.0e-3$	$3.5e-4$	$3.2e-5$	$1.6e-5$	$6.4e-6$	$2.8e-6$
I_{eff}	1.1	0.7	0.7	1.0	0.8	0.7

method described above. The resulting discrete saddle-point problems are solved iteratively by a GMRES method with multigrid preconditioning. The adaptive mesh refinement is based on an a posteriori error estimator. The weights in the error estimator (4.7) are evaluated using the strategy indicated in (4.8) and (4.9), with an interpolation constant set to $C_I = 0.1$. The mesh refinement uses the "fixed-fraction strategy" described above.

Table 5.1 shows the quality of the error estimator (4.7) for quantitative error control. The *efficiency index* is defined by $I_{eff} := E_h/\eta_h$, where $E_h := |J(u,q) - J(u_h, q_h)|$ is the error in the cost functional and $\eta_h := \eta(u_h, q_h)$ the value of the error estimator used. The reference value is obtained on a mesh with more than 200,000 cells. We compare the weighted error estimator with a simple ad hoc approach based on the standard energy-error estimator (4.11) for the state equation. Figure 5.2 shows the computed "optimal" states over the meshes generated by the two different error estimators.

The two meshes are quite different: The energy-error estimator overemphasizes the steep gradients near the control boundary and it leaves the mesh too coarse along the observation boundary. The more selective *weighted* error estimator concentrates the mesh cells where they are needed for the optimization process. The quantitative effects on the mesh efficiency of these two different refinement criteria is shown in Figure 5.3 (E_h versus N in log/log-scale).

Finally, we check how well the approximation $\{u_h, \lambda_h, q_h\}$ obtained by the weighted error estimator (4.7) actually satisfies the state equation; for this the *global* energy-error estimator (4.11) is taken as quality measure. Table 5.2 shows a comparison of the two sequences of meshes generated by the weighted error estimator $\eta_\omega = \eta_\omega(u_h, \lambda_h, q_h)$ ("ω-meshes") and the energy-error estimator $\eta_E = \eta_E(u_h)$ ("E-meshes"). The first and second columns contain the values of η_ω and η_E on ω-meshes, while the third and fourth columns contain the values of η_ω and η_E on E-meshes.

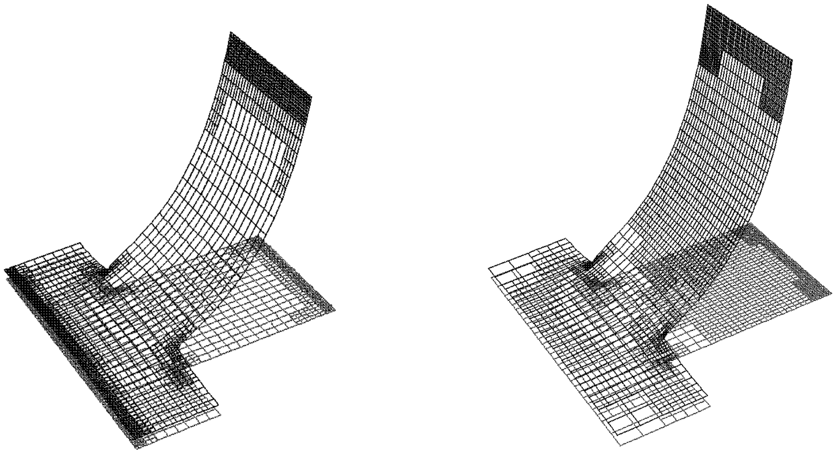


FIG. 5.2. *Linear test: Comparison of discrete solutions obtained by the weighted error estimator (left,* $N \sim 1600$ *cells) and the energy-error estimator (right,* $N \sim 1700$ *cells).*

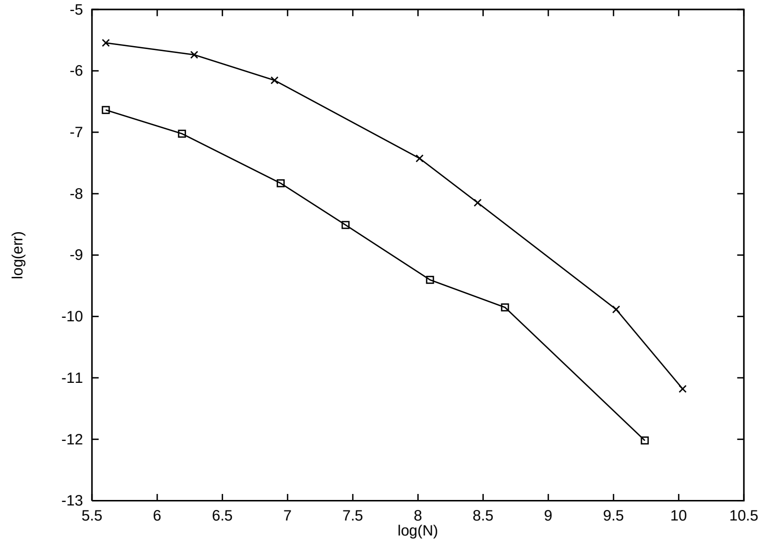


FIG. 5.3. *Linear test* (*configuration* 1): *Comparison of the efficiency of the meshes generated by the weighted error estimator (symbol* □ *) and the energy-error estimator (symbol* × *) in* log / log *scale.*

We see that the energy-norm error bound η_E for the state equation on the ω-meshes is slightly larger than on the E-meshes. This is not surprising since the ω-meshes are not so much refined in the regions where the state variable has a steep gradient. The cells are rather concentrated along the control and observation boundaries which seems to be more effective for the optimization process. Indeed, the approximate solution $\{u_h, \lambda_h, q_h\}$ obtained by the weighted error estimator η_ω achieves a much smaller value (factor ~ 0.1) of the cost functional. However, for other data, e.g., $c_0 = \cos(2x)$ and $\alpha = 0.0001$, the discrepancy between the two kinds of meshes with respect to the satisfaction of the state equation may be more significant. In those cases it would be advisable to use the combined error estimator $\eta_{\omega,E}$ described

TABLE 5.2
Linear test (*configuration* 1): *Values of the two error estimators* η_ω *and* η_E *obtained on "ω-meshes" and on "E-meshes."*

$N \approx$	η_ω on ω-meshes	η_E on ω-meshes	η_ω on E-meshes	η_E on E-meshes
140	0.0040205	0.0193270	0.0043245	0.0162589
300	0.0022030	0.0157156	0.0026536	0.0112183
750	0.0008330	0.0092718	0.0020437	0.0074801
3700	0.0001660	0.0049598	0.0004870	0.0034197
11000	0.0000532	0.0026208	0.0002199	0.0019036
21000	0.0000317	0.0020740	0.0001189	0.0014285
28000	0.0000239	0.0016294	0.0001088	0.0012403
48000	0.0000108	0.0013373	0.0000722	0.0009399
145000	0.0000037	0.0006950	0.0000328	0.0005466

above, if stronger "admissibility" of the discrete state u_h is required.

6. The nonlinear case. Now, we consider a nonlinear analogue of the abstract linear model problem (2.1), (2.2),

$$a(u)(\varphi) = (f,\varphi) + b(q,\varphi) \quad \forall \varphi \in V, \tag{6.1}$$

where $a(\cdot)(\cdot)$ is a semilinear form on the Hilbert space V. The cost functional $J(\cdot,\cdot)$ is the same as in the linear case. The corresponding Lagrange functional

$$\mathcal{L}(u,q,\lambda) := J(u,q) + a(u)(\lambda) - b(q,\lambda) - (f,\lambda),$$

leads to the first-order necessary optimality condition

$$\begin{aligned} a'(u)(v,\lambda) + (cu - c_0, cv) &= 0 \quad \forall v \in V,\\ a(u)(\mu) - b(q,\mu) &= (f,\mu) \quad \forall \mu \in V,\\ -b(r,\lambda) + n(q,r) &= 0 \quad \forall r \in Q, \end{aligned} \tag{6.2}$$

where $a'(u)(\cdot,\cdot)$ denotes the tangent form of $a(\cdot)(\cdot)$ at u. As in the linear case, the discrete approximations $\{u_h, \lambda_h, q_h\}$ are determined as solutions of the saddle-point problem

$$\begin{aligned} a'(u_h)(v_h,\lambda_h) + (cu_h - c_0, cv) &= 0 \quad \forall v_h \in V_h,\\ a(u_h)(\mu_h) - b(q_h,\mu_h) &= (f,\mu_h) \quad \forall \mu_h \in V_h,\\ -b(r_h,\lambda_h) + n(q_h,r_h) &= 0 \quad \forall r_h \in Q_h, \end{aligned} \tag{6.3}$$

where the discretization is the same as in the linear case. We will use again the notation $x = \{u,\lambda,q\}$ and $x_h = \{u_h,\lambda_h,q_h\}$ for points in the spaces $X := V \times V \times Q$ and $X_h := V_h \times V_h \times Q_h$, respectively.

The a posteriori error estimation in the case of a nonlinear state equation follows the same pattern as in the linear case. First, we state an abstract result.

PROPOSITION 6.1. *For the Galerkin finite element approximation* (6.3) *of the abstract model problem* (6.2) *with nonlinear state equation and quadratic cost functional there holds*

$$J(u,q) - J(u_h,q_h) = \tfrac{1}{2}\nabla\mathcal{L}(x_h)(x - i_h x) + R(x,x_h)\,, \tag{6.4}$$

where the remainder term $R(x,x_h)$ *can be estimated by*

$$|R(x,x_h)| \;\le\; \sup_{\hat{x}\in[x_h,x]} |\nabla^3\mathcal{L}(\hat{x})(x-x_h, x-x_h, x-x_h)|. \tag{6.5}$$

Proof. The Galerkin orthogonality relation now reads

$$\nabla^2\mathcal{L}(\overline{xx_h})(x-x_h,\varphi_h) = \nabla\mathcal{L}(x)(\varphi_h) - \nabla\mathcal{L}(x_h)(\varphi_h) = 0, \quad \varphi_h \in X_h, \tag{6.6}$$

with the abbreviating notation

$$\mathcal{L}(\overline{xx_h}) := \int_0^1 \mathcal{L}(x+t(x_h-x))\,dt.$$

Since the solutions u and u_h satisfy the corresponding state equations, there holds again

$$J(u,q) - J(u_h,q_h) = \mathcal{L}(x) - \mathcal{L}(x_h).$$

By Taylor expansion, there holds

$$\begin{aligned} rcl\mathcal{L}(x) - \mathcal{L}(x_h) = \nabla\mathcal{L}(x)(x-x_h) - \tfrac{1}{2}\nabla^2\mathcal{L}(x)(x-x_h,x-x_h) \\ + \tfrac{1}{6}\nabla^3\mathcal{L}(\tilde{x})(x-x_h,x-x_h,x-x_h), \end{aligned}$$

where $\tilde{x}$ lies between x and x_h. Since x is a stationary point of $\mathcal{L}$, the first term on the right vanishes. In order to relate the second term to the Galerkin relation (6.6), we use again Taylor expansion,

$$\nabla^2\mathcal{L}(x)(x-x_h,x-x_h) = \nabla^2\mathcal{L}(\overline{xx_h})(x-x_h,x-x_h) + \nabla^3\mathcal{L}(\hat{x})(x-x_h,x-x_h,x-x_h),$$

where $\hat{x}$ is another point between x and x_h. In view of the identity

$$\nabla^2\mathcal{L}(\overline{xx_h})(x-x_h,\cdot) = \nabla\mathcal{L}(x)(\cdot) - \nabla\mathcal{L}(x_h)(\cdot) = -\nabla\mathcal{L}(x_h)(\cdot),$$

and the Galerkin relation (6.6), we conclude that

$$\begin{aligned} \mathcal{L}(x) - \mathcal{L}(x_h) &= -\tfrac{1}{2}\nabla^2\mathcal{L}(\overline{xx_h})(x-x_h,x-x_h) + R(x,x_h) \\ &= -\tfrac{1}{2}\nabla^2\mathcal{L}(\overline{xx_h})(x-x_h,x-x_h-\varphi_h) + R(x,x_h) \\ &= \tfrac{1}{2}\nabla\mathcal{L}(x_h)(x-x_h-\varphi_h) + R(x,x_h), \end{aligned}$$

with an arbitrary $\varphi_h \in X_h$, and the remainder term

$$R(x,x_h) = \nabla^3\mathcal{L}(\hat{x})(x-x_h,x-x_h,x-x_h) + \tfrac{1}{6}\nabla^3\mathcal{L}(\tilde{x})(x-x_h,x-x_h,x-x_h).$$

Taking here $\varphi_h = i_hx - x_h$ eventually results in

$$\mathcal{L}(x) - \mathcal{L}(x_h) = \tfrac{1}{2}\nabla\mathcal{L}(x_h)(x-i_hx) + R(x,x_h),$$

which completes the proof. □

We note that if the cost functional $J(\cdot)$ is quadratic and the control form $b(\cdot,\cdot)$ bilinear, then the only nonzero terms in $\nabla^3\mathcal{L}$ are

$$\frac{\partial^3\mathcal{L}}{\partial\lambda\partial^2u}(x) = a''(u)(\cdot,\cdot,\cdot), \qquad \frac{\partial^3\mathcal{L}}{\partial^3u}(x) = a'''(u)(\cdot,\cdot,\cdot,\lambda).$$

Further, if additionally the state equation is linear, then the remainder term $R(x,x_h)$ vanishes.

We will apply this abstract result for a nonlinear problem of optimal control in the "Ginzburg–Landau model" of superconductivity in semiconductors; for references see Du, Gunzburger, and Peterson [7], Itô and Kunish [10], and also Tinkham [13]. It has the same structure as the linear model problem considered above,

$$\begin{gathered} -\Delta u + s(u) = f \quad \text{in } \Omega, \\ \partial_n u = 0 \ \text{ on } \Gamma_N, \quad \partial_n u = q \ \text{ on } \Gamma_C, \end{gathered} \tag{6.7}$$

with the nonlinearity $s(u) := u^3 - u$, and the quadratic cost functional

$$J(u,q) = \tfrac{1}{2}\|u - c_0\|^2_{\Gamma_O} + \tfrac{\alpha}{2}\|q\|^2_{\Gamma_C}.$$

The corresponding first-order necessary condition (6.2) uses the notation

$$a(u)(v) = (\nabla u, \nabla v)_\Omega + (s(u), v)_\Omega, \quad b(q,v) = (q,v)_{\Gamma_C}, \quad n(q,r) = \alpha(q,r)_{\Gamma_C},$$

and is approximated by the scheme (6.3). The well-posedness of this optimization problem, the existence of the adjoint variable λ, as well as a priori error estimates for its discretization have been discussed by Gunzburger and Hou [9]. From Proposition 6.1, we conclude the following a posteriori result.

PROPOSITION 6.2. *For error control with respect to the cost functional J, there holds the weighted a posteriori error estimate*

$$|J(u,q) - J(u_h,q_h)| \ \le \ \eta_\omega(u_h,\lambda_h,q_h) + R(\{u,\lambda,q\},\{u_h,\lambda_h,q_h\}), \tag{6.8}$$

where the local error indicators $\eta_T(u_h,\lambda_h,q_h)$ in the linearized error estimator

$$\eta_\omega(u_h,\lambda_h,q_h) := \sum_{T\in\mathbb{T}_h} \eta_T(u_h,\lambda_h,q_h) \tag{6.9}$$

are defined as in the linear case (Proposition 4.2), here with the "cell residuals"

$$\begin{gathered} R^{(u)}_{h|T} := -\Delta u_h + s(u_h) - f, \quad R^{(\lambda)}_{h|T} := -\Delta\lambda_h + s'(u_h)\lambda_h, \\ r^{(q)}_{h|\Gamma} := \alpha q_h - \lambda_h, \ \ \textit{if } \Gamma \subset \Gamma_C. \end{gathered} \tag{6.10}$$

For the remainder term, there holds the a priori estimate

$$\big|R(\{u,\lambda,q\},\{u_h,\lambda_h,q_h\})\big| \ \le \ 6\int_\Omega \Big\{ \max\{|u|,|u_h|\}|u-u_h|^3 + |u-u_h|^2|\lambda-\lambda_h| \Big\}\,dx. \tag{6.11}$$

As in the linear case, the weights are evaluated numerically using the approximations $\{u_h,\lambda_h,q_h\}$, but now the weighted error estimator contains an additional linearization error represented by the remainder R. Theory as well as practical experience show that, in the present case, this additional error is of higher order on well-adapted meshes and can therefore be neglected. In fact, assuming sufficient smoothness of the solution $\{u,\lambda,q\}$, there holds

$$\big|R(\{u,\lambda,q\},\{u_h,\lambda_h,q_h\})\big| \le c(u,u_h)\, h^6_{\max}, \tag{6.12}$$

with the maximum step size $h_{\max}$ of the mesh. The proof of this order-optimal estimate employing techniques from L^∞-error analysis of finite elements would be

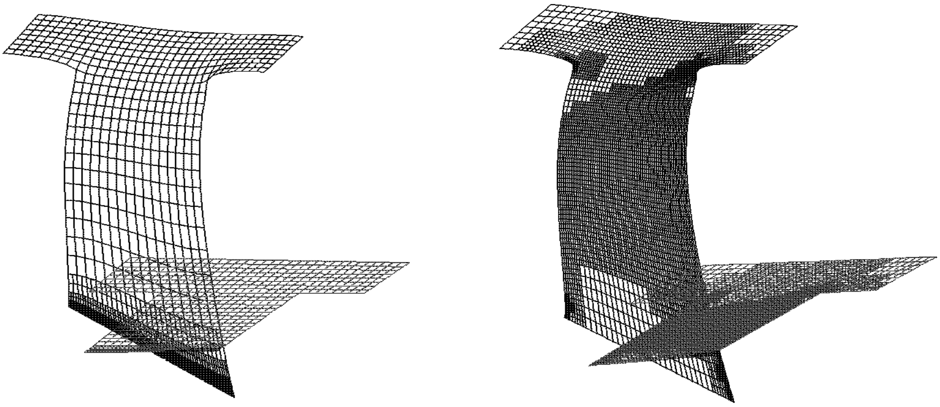


FIG. 7.1. *Nonlinear test (configuration* 2, $\alpha=0$*): Comparison of discrete solutions obtained by the weighted error estimator (left,* $N\sim 5000$ *cells) and the energy-error estimator (right,* $N\sim 4800$ *cells).*

TABLE 7.1
Nonlinear test (configuration 2, $\alpha=0$*): Efficiency of the weighted error estimator in the case* $\Gamma_C=\Gamma_O$.

N	596	1616	5084	8648	15512
E_h	2.56e-04	2.38e-04	8.22e-05	4.21e-05	3.99e-05
I_{eff}	0.34	0.81	0.46	0.29	0.43

rather lengthy and is therefore omitted. In view of this observation, we neglect the remainder term in the a posteriori error estimate (6.8) and base the mesh adaptation on its main part $\eta_\omega(u_h,\lambda_h,q_h)$.

The discrete problems (6.3) are solved by a quasi-Newton iteration which is derived from a corresponding scheme formulated on the continuous level. On each discrete level the Newton iteration is carried to the limit before the error estimator is applied for mesh refinement. The results of this process may significantly differ from those obtained if each Newton step is discretized separately, mixing iteration and discretization errors together; see the preceding publication [2] for the latter approach.

7. Numerical results—nonlinear case. We again compare the weighted error estimator with a simple ad hoc energy-error estimator of the form (4.11) using the modified cell residuals (6.2). For illustrating our approach, we consider two different choices for the boundaries of control and observation shown in Figure 5.1 as configuration 1 and configuration 2. The notation I_{eff}, E_h, and η_h is as defined above.

(i) *First test:* First, we consider configuration 2 in which the same boundary is taken for control and observation, $\Gamma_C=\Gamma_O$ (lower boundary of the T-shaped domain) and set $\alpha=0$. In this configuration, we do not expect any need for strong mesh refinement "far away" from this boundary if we only want to deal with the optimization problem.

The observations are taken as $c_0(x)=\sin(0.19x)$. Table 7.1 shows the quality of the weighted error estimator for quantitative error control for this nonlinear test case. The reference value for the objective function $J(u,q)$ is computed on a refined mesh with about $130,000$ cells. Due to the special choice $\Gamma_C=\Gamma_O$, the adjoint variable λ equals zero almost everywhere away from Γ_C, i.e., the weighted error

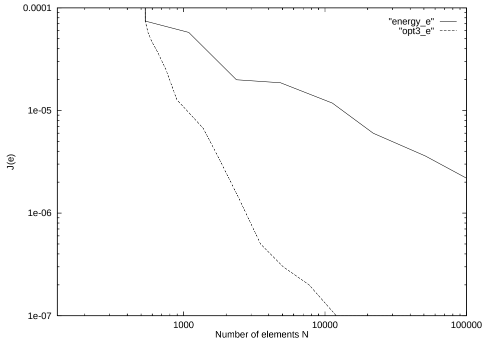


Fig. 7.2. *Nonlinear test (configuration* 2*,* $\alpha = 0$*): Comparison of efficiency of meshes generated by the two estimators* η_ω *(broken line) and* η_E *(solid line) in* $\log/log$ *scale.*

Table 7.2
Nonlinear test (configuration 1*,* $\alpha = 0.1$*): Efficiency of the weighted error estimator for computing a secondary stationary point.*

N	512	15368	27800	57632	197408
E_h	9.29e-05	8.14e-07	4.86e-07	2.31e-07	4.58e-08
I_{eff}	1.32	0.56	0.35	0.42	0.32

Table 7.3
Nonlinear test (configuration 1*,* $\alpha = 1$*): Efficiency of the weighted error estimator for computing a secondary stationary point.*

N	512	8120	25544	42608	126284
E_h	2.08e-03	4.35e-05	9.26e-06	5.95e-06	8.94e-07
I_{eff}	0.52	0.73	0.88	1.21	0.98

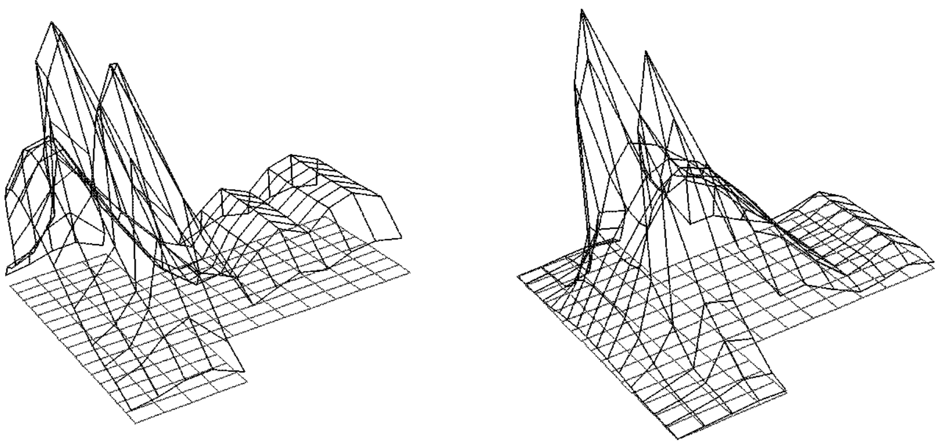


Fig. 7.3. *Nonlinear test (configuration* 1*,* $\alpha = 0.1$*): Distributions of local error indicators in the weighted error estimator* η_ω *(left) and the energy-error estimator* η_E *(right).*

estimator pays attention only to the neighborhood of the control boundary. The energy-error estimator mainly recognizes the singularities in the primal solution at the two reentrant corners (see Figure 7.1).

In Figure 7.2, we compare the efficiency of the meshes generated by the two

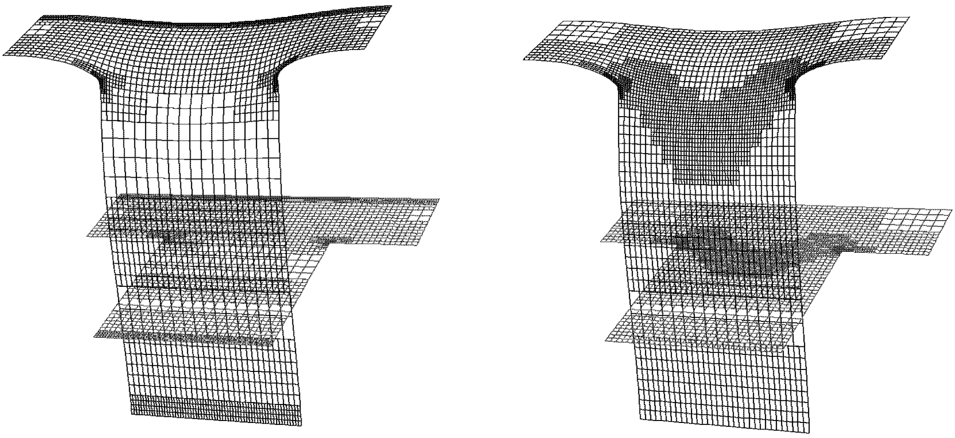


FIG. 7.4. *Nonlinear test (configuration* 1*,* $\alpha = 0.1$*): Comparison of discrete solutions obtained by the weighted error estimator* η_ω *(left,* $N \sim 3000$ *cells) and the energy-error estimator* η_E *(right,* $N \sim 3300$ *cells).*

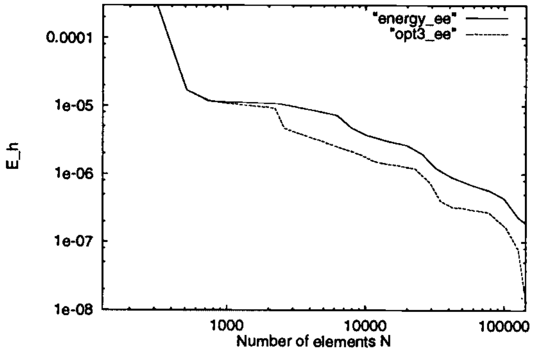


FIG. 7.5. *Nonlinear test (configuration* 1*,* $\alpha = 0.1$*): Comparison of efficiency of meshes generated by the two error estimators:* η_ω *(broken line) and* η_E *(solid line) in* log*/log scale.*

estimators. We see that in this situation, the solution of the optimization problem is approximated with significantly less cells using the weighted error estimator which exploits the "extreme" feature $\Gamma_C = \Gamma_O$ of this problem.

(ii) *Second test:* Now, we consider configuration 1 in which the control and the observation are taken on opposite boundaries, $\Gamma_C \cap \Gamma_O = \emptyset$. In this case, we expect better results for the energy-error estimator than in configuration 1, because the information must pass from the control to the observation boundary and the corner singularities will have a stronger effect. Nevertheless, the weighted error estimator should perform better since it also considers the critical control and observation boundaries.

We take the observation as $c_0 \equiv 1$, as in the linear case, and set $\alpha = 0.1$. In this configuration, there exist several stationary points of $\mathcal{L}(u, q, \lambda)$, which can be obtained by varying the starting values for the Newton iteration. One trivial solution (actually the global minimum) is a constant equal to c_0, with $q \equiv 0$. In this case, we have $J(u, q) = 0$ (up to round-off error) and match the observations already on a very coarse mesh. We do not show the results of this computation. The two other stationary points are symmetric to each other with respect to the center plane $\{x = 0\}$. Tables 7.2 and 7.3 show the quality of the weighted error estimator for quantitative error control for one of these local minima. The reference values for $J(u, q)$ are obtained on an adaptive mesh with about $550,000$ cells. The numerical results demonstrate the correct qualitative behavior of the weighted error estimator. For the choice $\alpha = 1$, we get slightly better results than for $\alpha = 0.1$ because of higher stability in the optimization problem.

Next, Figure 7.3 shows the distribution of the cell error indicators η_T in the weighted error estimator η_ω and in the energy-error estimator η_E for $\alpha = 0.1$. We clearly see the different ways in which these error estimators put their weight: η_ω observes the control and observation boundary which is critical for the optimization process while η_E emphasizes the corner singularities. Figure 7.4 shows the resulting meshes together with the computed discrete solutions. Finally, in Figure 7.5, we see the faster convergence toward the minimum of the objective functional using the weighted error estimator compared to the energy-error estimator.

REFERENCES

[1] M. Ainsworth and J.T. Oden, *A posteriori error estimation in finite element analysis*, Comput. Methods Appl. Mech. Engrg., 142 (1997), pp. 1–88.

[2] R. Becker and H. Kapp, *Optimization in PDE models with adaptive finite element discretization*, in Proceedings ENUMATH 97 (Heidelberg), H.G. Bock et al., eds., World Scientific, River Edge, NJ, 1998, pp. 147–155.

[3] R. Becker and R. Rannacher, *Weighted a posteriori error control in FE methods*, ENUMATH-95, Paris, 1995, in Proceedings ENUMATH-97, H.G. Bock et al., eds., World Scientific, River Edge, NJ, 1998, pp. 621–637.

[4] R. Becker and R. Rannacher, *A feed-back approach to error control in finite element methods: Basic analysis and examples*, East-West J. Numer. Math, 4 (1996), pp. 237–264.

[5] S.C. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, Berlin, 1994.

[6] G.F. Carey and J.T. Oden, *Finite Elements, Computational Aspects*, Vol. III, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[7] Q. Du, M.D. Gunzburger, and J.S. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[8] K. Eriksson, D. Estep, P. Hanspo, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numerica 1995, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 1995, pp. 105–158.

[9] M.D. Gunzburger and L.S. Hou, *Finite-dimensional approximation of a class of constrained nonlinear optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1001–1043.

[10] K. Itô and K. Kunisch, *Augmented Lagrangian-SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim., 34 (1996), pp. 874–891.

[11] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.

[12] R. Rannacher, *Error control in finite element computations*, in Error Control and Adaptivity in Scientific Computing, H. Bulgak and C. Zenger, eds., NATO Sci. Ser. C Math. Phys. Sci., 536, Kluwer, Dordrecht, the Netherlands, 1999, pp. 247–278.

[13] M. Tinkham, *Introduction to Superconductivity*, McGraw-Hill, New York, 1975.

[14] R. Verfürth, *A Review of A Posteriori Estimation and Adaptive Mesh-Refinement Techniques*, Adv. Numer. Math., Wiley-Teubner, New York, Stuttgart, 1996.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 133–140

PERSISTENCE OF EXCITATION PROPERTIES FOR TIME-VARYING AUTOREGRESSIVE SYSTEMS[*]

SERGIO BITTANTI[†] AND M. C. CAMPI[‡]

Abstract. It is well known that a crucial property for the effective identification of time-varying systems is that the data carry continual information on the parameters to be estimated. As a matter of fact, only in this case can the identification algorithm rely on fresh information in forming a reliable estimate of the current value of these parameters. This concept has been formalized in the system identification literature under the name of *persistence of excitation*.

In this paper, the persistence of excitation property is studied for a class of time-varying systems (that includes the standard autoregressive model as a particular case) and conditions for it to hold are derived.

Key words. time-varying models, persistence of excitation, autoregressive models, system identification

AMS subject classifications. Primary, 93E12; Secondary, 93B30

PII. S0363012998343483

1. Introduction.

In the last two decades, a considerable effort has been put into the comprehension of identification methods for the estimation of time-varying systems.

A huge stream of research has been devoted to situations that somehow reduce to the problem of estimating constant unknown parameters. This is, for instance, the case of the so-called random coefficient autoregressive models; see, e.g., Nicholls and Quinn (1982), Chow (1983), and Beran and Hall (1992). These models are characterized by parameters which are randomly fluctuating according to the law $\vartheta(t) = \bar{\vartheta} + \delta(t)$, $\delta(t)$ being an independent sequence. In this framework the main concern is the consistent estimation of the mean value $\bar{\vartheta}$. Another kind of time-varying systems which has attracted interest in recent years are the so-called nearly nonstationary autoregressive models. In this case, the time-varying parameters are asymptotically convergent and the corresponding asymptotic invariant model exhibits singularities on the unit circle. The limiting distribution of the estimation error when the identification is performed via the standard least squares algorithm is studied, e.g., in Cox and Llatas (1991); see also Cox (1991).

In the above literature, the fact that the estimated parameters are in fact constant makes the estimation task simpler than in truly time-varying situations. As a matter of fact, when the parameters are constant, the same unknowns are estimated through time and it is expected that a consistent estimate can be formed under the sole condition that data carry enough information in the long run. On the other hand, when the goal is that of estimating truly time-varying parameters, one has to somehow guarantee that a certain amount of information is available over any finite interval of time. As a matter of fact, only in this way can the identification algorithm rely on fresh information in forming a reliable estimate of the current value of the parameters.

[*] Received by the editors August 12, 1998; accepted for publication (in revised form) February 16, 2000; published electronically August 17, 2000.

http://www.siam.org/journals/sicon/39-1/34348.html

[†] Dipartimento di Elettronica e Informazione, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy (bittanti@elet.polimi.it).

[‡] Dipartimento di Elettronica per l'Automazione, Universita' di Brescia, via Branze 38, 25123 Brescia, Italy (campi@ing.unibs.it).

This idea is well known in the identification literature under the name of *persistence of excitation.*

Letting $\varphi(\cdot)$ be the observation vector, a persistence of excitation condition which has been widely used in the literature takes the form

$$\text{pr}\left(\lambda_{\min}\left\{\sum_{i=t}^{t+s-1}\frac{\varphi(i)\varphi(i)'}{1+\|\varphi(i)\|^2}\right\}\geq k_1 \,\middle|\, \sigma_{t-1}\right)\geq k_2 \qquad \forall t, \tag{1}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue and σ_t is the so-called σ-algebra of the past, that is, the σ-algebra generated by all system processes up to time t. Roughly, this condition requires that, whatever the past evolution of the system might have been, the information carried by data over the next s time points spans the entire parameter space with a finite nonzero probability.

Condition (1) was first introduced in Guo (1990) in a form that is slightly different from but equivalent to (1), and has henceforth been used in many different contributions.

Under (1), Guo (1990) proves stability and convergence results for a Kalman filter based algorithm used in the estimation of time-varying parameters generated by a random-walk–type equation. In the paper of Bittanti and Campi (1994) it is proven that a forgetting factor least squares identification algorithm provides bounded estimates if condition (1) is met and the forgetting factor is chosen to be larger than a certain threshold. Another contribution using the persistence of excitation condition (1) is Campi (1994). There, an explicit expression for the asymptotic estimation error is given for a forgetting factor based least squares algorithm. This bound shows the dependence of the estimation error on the speed of the time variability of the parameters and the variance of noise.

There are many more contributions on system identification where significant properties are proven under conditions related to (1). Among others, we cite Bittanti and Campi (1991a, 1991b); Guo, Ljung, and Priouret (1993); Guo and Ljung (1995a, 1995b); and Campi (1997). An additional interesting paper is Ravikanth and Meyn (1999), where a lower bound for the estimation error valid for any identification algorithm is worked out.

In all of the above-mentioned contributions, condition (1) is taken for granted or proven only in certain specific situations. In the present paper we address the problem of verifying that such a condition is in fact satisfied for a class of time-varying systems which includes, but is not limited to, autoregressive systems. In this way, all the results proven in these contributions can in fact be applied to this class of models.

The paper is organized as follows. In section 2 the system class is introduced. The persistence of excitation condition is then discussed in section 3.

2. The system. Let us consider a time-varying state variable system described by the equation

$$\varphi(t)=G(t)\varphi(t-1)+v(t). \tag{2}$$

In (2), $\varphi(t)\in\mathbb{R}^n$ is the so-called observation vector and it is a measurable signal, and $v(t)$ is a remote unmeasurable noise that plays the role of a latent variable in the generation of $\varphi(t)$. Throughout the paper, it is assumed that matrices $G(t)$ form a strictly stationary stochastic process.

The transition matrix associated with $G(t)$ is defined as

$$\Phi(t,s) := G(s)G(s+1)\cdots G(t).$$

A typical example of system (2) is a time-varying scalar autoregressive model of the form

$$y(t) = a_1(t)y(t-1) + a_2(t)y(t-2) + \cdots + a_n(t)y(t-n) + d(t). \tag{3}$$

In this case, by letting

$$G(t) = \begin{bmatrix} a_1(t) & a_2(t) & \cdots & a_n(t) \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix}, \quad \varphi(t) = \begin{bmatrix} y(t) \\ y(t-1) \\ \vdots \\ y(t-n+1) \end{bmatrix}, \quad v(t) = \begin{bmatrix} d(t) \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

system (2) is immediately recovered. Clearly, system (2) can accommodate many other specific situations than the autoregressive system (3).

The following assumptions are made on system (2).

Assumption 1. $v(\cdot)$ is a zero-mean, bounded independently and identically distributed (i.i.d.) sequence, independent of $G(\cdot)$.

Assumption 2. $\exists\rho$: $\rho^{-t}\|\Phi(t,0)\| \le \alpha\ \forall t$ almost surely.

Clearly, Assumption 2 is an exponential stability condition. It is worthwhile pointing out that there is a milder stability condition that could be considered.

Assumption 2$'$. $\exists\rho$: $\limsup_{t\to\infty}\rho^{-t}\|\Phi(t,0)\| = 0$ almost surely.

Assumption 2$'$ is a stability assumption of stochastic type that requires $\|\Phi(t,0)\|$ to go to zero exponentially fast with *asymptotic* deterministic rate ρ. On the other hand, Assumption 2 imposes restrictions for any *finite* t. It is in fact a truly deterministic stability assumption.

It is easy to see that Assumption 2$'$ is equivalent to

$$\limsup_{t\to\infty} t^{-1}\log\|\Phi(t,0)\| \le -\gamma < 0 \quad \text{almost surely.} \tag{4}$$

This last condition has been discussed (in a continuous-time setting) by Solo (1994). Among other things, Solo provides conditions on the eigenvalues of the stochastic matrix $G(t)$ such that (4) holds true.

Finally, notice that, since $G(\cdot)$ is strictly stationary, Assumption 2 is equivalent to

$$\|\Phi(t,s)\| \le \alpha\rho^{t-s} \quad \forall t,s \qquad \text{almost surely.}$$

3. Main result: Persistence of excitation condition. In this section, the persistence of excitation condition is discussed and necessary conditions for it to hold are derived.

For subsequent use, we introduce the σ-algebra generated by the past of $v(\cdot)$ and the past, present, and future of $G(\cdot)$:

$$\zeta_t = \sigma(v(i), \quad i \le t; \quad G(\cdot)).$$

Notice that $\varphi(t)$ is measurable with respect to ζ_t.

For the sake of clarity, we point out that the σ-algebra of the past in condition (1) is given by

$$\sigma_t = \sigma(v(j), G(j), j \leq t).$$

We start by proving the following proposition which is a law of large numbers of conditional type for system (2).

PROPOSITION 3.1. *Under Assumptions* 1 *and* 2,

$$E\left[\left\| \frac{1}{k} \sum_{i=t}^{t+k-1} (\varphi(i)\varphi(i)' - E[\varphi(i)\varphi(i)' \,|\, \zeta_t]) \right\|^2 \,\middle|\, \zeta_t \right] \longrightarrow 0 \quad \textit{as } k \to \infty,$$

uniformly with respect to both time t and probability outcome.

Proof. The following chain of inequalities holds true:

$$\begin{aligned} &E\left[\left\| \frac{1}{k} \sum_{i=t}^{t+k-1} (\varphi(i)\varphi(i)' - E[\varphi(i)\varphi(i)' \,|\, \zeta_t]) \right\|^2 \,\middle|\, \zeta_t \right] \\ &\quad \leq \frac{1}{k^2} n \left\| E\left[\sum_{i,j=t}^{t+k-1} (\varphi(i)\varphi(i)' - E[\varphi(i)\varphi(i)' \,|\, \zeta_t]) \right.\right. \\ &\qquad\qquad \left.\left. \times \; (\varphi(j)\varphi(j)' - E[\varphi(j)\varphi(j)' \,|\, \zeta_t]) \,\middle|\, \zeta_t \right] \right\| \end{aligned}$$

(since, for any stochastic matrix $M \geq 0$ of dimension n, $E[\|M\|] \leq n\|E[M]\|$)

$$\begin{aligned} &\leq \frac{1}{k^2} n \sum_{i,j=t}^{t+k-1} \|E[(\varphi(i)\varphi(i)' - E[\varphi(i)\varphi(i)' \mid \zeta_t]) \\ &\qquad\qquad \times (\varphi(j)\varphi(j)' - E[\varphi(j)\varphi(j)' \mid \zeta_t]) \mid \zeta_t]\| \\ &\leq \frac{1}{k^2} 2n \sum_{\substack{i,j=t \\ j \geq i}}^{t+k-1} \|E[(\varphi(i)\varphi(i)' - E[\varphi(i)\varphi(i)' \mid \zeta_t]) \\ &\qquad\qquad \times (E[\varphi(j)\varphi(j)' \mid \zeta_i] - E[\varphi(j)\varphi(j)' \mid \zeta_t]) \mid \zeta_t]\|. \end{aligned}$$

In this last expression, the norm of $(\varphi(i)\varphi(i)' - E[\varphi(i)\varphi(i)' \mid \zeta_t])$ is deterministically bounded in view of the boundedness of $v(\cdot)$ (Assumption 1) and the exponential stability of the system (Assumption 2).

Therefore, to complete the proof it suffices to prove that the norm of $(E[\varphi(j)\varphi(j)' \mid \zeta_i] - E[\varphi(j)\varphi(j)' \mid \zeta_t])$, $j \geq i \geq t$, is bounded by a deterministic function of $j - i$ only, which tends exponentially to zero as $j - i \to \infty$.

Set $\varphi(r \mid s) := E[\varphi(r) \mid \zeta_s]$, $r \geq s$. Since $v(\cdot)$ is an independent sequence, we have

$$\varphi(r \mid s) = \sum_{k=-\infty}^{s+1} \Phi(r,k)v(k-1).$$

Taking into account the exponential stability assumption (Assumption 2) and that the noise $v(\cdot)$ is bounded (Assumption 1), this last expression shows that $\|\varphi(r \mid s)\|$ is bounded by a deterministic function of $r-s$ only, which tends exponentially to zero as $r-s\to\infty$. The term $(E[\varphi(j)\varphi(j)' \mid \zeta_i]-E[\varphi(j)\varphi(j)' \mid \zeta_t])$ can now be handled as follows:

$$\begin{aligned}
&E[\varphi(j)\varphi(j)' \mid \zeta_i]-E[\varphi(j)\varphi(j)' \mid \zeta_t]\\
&\quad= E[(\varphi(j \mid i)+(\varphi(j)-\varphi(j \mid i)))(\varphi(j \mid i)+(\varphi(j)-\varphi(j \mid i)))' \mid \zeta_i]\\
&\qquad - E[(\varphi(j \mid t)+(\varphi(j)-\varphi(j \mid t)))(\varphi(j \mid t)+(\varphi(j)-\varphi(j \mid t)))' \mid \zeta_t]\\
&\quad= \varphi(j \mid i)\varphi(j \mid i)'-\varphi(j \mid t)\varphi(j \mid t)'-\sum_{k=t+2}^{i+1}\Phi(j,k)\Delta\phi(j,k)',
\end{aligned}$$

where $\Delta := E[v(t)v(t)']$. The thesis follows by observing that the norm of each of these three terms is bounded by a deterministic function of $j-i$ only, which tends exponentially to zero as $j-i\to\infty$. □

Notice that, up to now, no conditions have been introduced guaranteeing that vector $\varphi(\cdot)$ is somehow exciting (in fact, under Assumptions 1 and 2, $v(\cdot)$ and/or $G(\cdot)$ may well be identically zero). We now introduce an extra condition (Assumption 3 below) which can be interpreted as an excitation condition. We anticipate that, in view of Proposition 1, Assumption 3 immediately leads to concluding that $\varphi(\cdot)$ is persistently exciting in the sense of definition (1) (see Theorem 1 below). The fact that Assumption 3 holds true in many situations of interest (e.g., for the autoregressive system (2)) is discussed immediately after the theorem.

Assumption 3. $E[\varphi(i)\varphi(i)' \mid \zeta_t]\geq H>0$ $\forall i\geq t+\bar{n}$, for some integer $\bar{n}$.

THEOREM 3.2. *Under Assumptions* 1–3*, there exist an integer s and two positive real numbers k_1 and k_2 such that the persistence of excitation condition* (1) *is satisfied.*

Proof. Recalling that, for any pair of positive semidefinite matrices C and D, $\lambda_{\min}[C]\geq\lambda_{\min}[D]-\|C-D\|$, one obtains

$$\begin{aligned}
\lambda_{\min}\left\{\frac{1}{k}\sum_{i=t}^{t+k-1}\varphi(i)\varphi(i)'\right\}\geq{}&\lambda_{\min}\left\{\frac{1}{k}\sum_{i=t}^{t+k-1}E[\varphi(i)\varphi(i)' \mid \zeta_t]\right\}\\
&-\left\|\frac{1}{k}\sum_{i=t}^{t+k-1}(\varphi(i)\varphi(i)'-E[\varphi(i)\varphi(i)' \mid \zeta_t])\right\|.
\end{aligned}$$

Take now conditional expectation of this last equation with respect to ζ_t. Thanks to Assumption 3 and Proposition 1, it is then apparent that there exist an integer s and a real number β such that

$$E\left[\lambda_{\min}\left\{\frac{1}{s}\sum_{i=t}^{t+s-1}\varphi(i)\varphi(i)'\right\}\middle|\zeta_t\right]\geq\beta>0\quad\forall t.$$

Then in view of the boundedness of $\varphi(\cdot)$ (Assumptions 1 and 2), we can conclude that there exist two positive real numbers k_1 and k_2 such that

$$\mathrm{pr}\left(\lambda_{\min}\left\{\sum_{i=t}^{t+s-1}\frac{\varphi(i)\varphi(i)'}{1+\|\varphi(i)\|^2}\right\}\geq k_1 \mid \zeta_t\right)\geq k_2\quad\forall t.$$

Since the σ-algebra generated by $v(j)$ and $G(j)$, $j\leq t-1$, is coarser than ζ_t, the thesis follows. □

Next, we show that Assumption 3 holds true in the case of the autoregressive system (2). Take $\bar{n} = n$. Recalling that $\varphi(r \mid s) = E[\varphi(r) \mid \zeta_s]$, for any $j \in [i-n, i-1]$ one has

$$\begin{aligned} &E[\varphi(i)\varphi(i)' \mid \zeta_t] \\ &\quad = E[E[(\varphi(i \mid j) + (\varphi(i) - \varphi(i \mid j)))(\varphi(i \mid j) + (\varphi(i) - \varphi(i \mid j)))' \mid \zeta_j] \mid \zeta_t] \\ &\qquad \text{(since } j \geq t) \\ &\quad \geq E[(\varphi(i) - \varphi(i \mid j))(\varphi(i) - \varphi(i \mid j))' \mid \zeta_t]. \end{aligned}$$

Since $\varphi(i) - \varphi(i \mid j) = \sum_{k=j+2}^{i+1} \Phi(i,k)v(k-1)$, we have ($\sigma^2 := E[d(t)^2]$)

- for $j = i-1$,

$$E[\varphi(i)\varphi(i)' \mid \zeta_t] \geq \operatorname{diag}(\sigma^2, 0, \dots, 0) = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix};$$

- for $j = i-2$,

$$\begin{aligned} E[\varphi(i)\varphi(i)' \mid \zeta_t] &\geq \Phi(i,i) \operatorname{diag}(\sigma^2, 0, \dots, 0)\phi(i,i)' \\ &= \begin{bmatrix} \star & \star & 0 & \dots & 0 \\ \star & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}; \end{aligned}$$

$(\cdots)$

- for $j = i-n$,

$$\begin{aligned} E[\varphi(i)\varphi(i)' \mid \zeta_t] &\geq \Phi(i, i-n+2) \operatorname{diag}(\sigma^2, 0, \dots, 0)\Phi(i, i-n+2)' \\ &= \begin{bmatrix} \star & \star & \star & \dots & \star \\ \star & \star & \star & \dots & \star \\ \star & \star & \star & \dots & \star \\ \vdots & \vdots & \vdots & & \vdots \\ \star & \star & \star & \dots & \sigma^2 \end{bmatrix}, \end{aligned}$$

where the $\star$'s are random entries, whose value is bounded uniformly with respect to time t and probability outcome. From the above relations, Assumption 3 easily follows with $\bar{n} = n$.

It is interesting to note that Assumption 3 holds in many extra situations. As a simple example, if $G(\cdot)$ is deterministic such that Assumption 2 holds, then Assumption 3 is met provided that the very minimal condition $E[\varphi(i)\varphi(i)'] > 0$ is satisfied.

The analysis has been conducted so far under the stability assumption, Assumption 2. It is, of course, of interest to investigate whether the persistence of excitation condition (1) still holds under the weaker stability assumption, Assumption 2$'$. Unfortunately, this is not the case, as the following simple example shows.

Example. Suppose that $\varphi(t)$ has two components and let $G(t) = \operatorname{diag}(g(t), 0)$. $g(\cdot)$ is an i.i.d. sequence such that $g(t) = 2$ with probability 0.5 and $g(t) = 0.25$ with probability 0.5. Finally, $v(t) = [v_1(t)v_2(t)]'$, where $v_1(\cdot)$ and $v_2(\cdot)$ are i.i.d. sequences independent of each other and of $g(\cdot)$ and take on values -1 and $+1$ with probability 0.5.

Assumptions 1 and 3 are trivially satisfied in this case. Assumption $2'$ is also satisfied. This is seen as follows. Since $\|\Phi(t,0)\| = g(0)g(1)\cdots g(t)$, for any given $\rho \in (0.1)$, we have

$$\frac{1}{t}\log(\rho^{-t}\|\Phi(t,0)\|) = \log\frac{1}{\rho} + \frac{1}{t}\sum_{s=0}^{t}\log g(s).$$

The second term in the left-hand side tends almost surely to $E[\log g(t)] = 0.5[\log 2 + \log 0.25] = -0.5\log 2$. Then, by taking ρ to be a real number such that $\log\frac{1}{\rho} - 0.5\log 2 < 0$, we have that $\frac{1}{t}\log(\rho^{-t}\|\Phi(t,0)\|)$ tends almost surely to a negative number. From this, we conclude that $\limsup_{t\to\infty}\rho^{-t}\|\Phi(t,0)\| = 0$ almost surely, that is, Assumption $2'$.

Next, we show that the persistence of excitation condition (1) is not satisfied in this case.

Given any real number h, let $\mathcal{A}_h := \{|\varphi_i(0)| > h\}$, where $\varphi_1(0)$ is the first component of $\varphi(0)$. Since $g(t)$ takes on value 2 with probability 0.5, $\varphi(0)$ has an unbounded distribution and so $\mathrm{pr}(\mathcal{A}_h) \neq 0 \quad \forall h$. Moreover, note that if $|\varphi_1(0)| > h$, then $|\varphi_1(t)| > (0.25)^t h - 5/4$. ($g(t)$ is either 2 or 0.25 and $|v_1(t)| = 1$.) Now, suppose by contradiction that (1) holds for certain fixed k_1, k_2, and s. Since $\varphi_2(i) = v_2(i)$ keeps bounded and $\|\varphi(i)\| \geq (0.25)^s h - 5/4(i \in [1,s])$, a real number h exists such that condition $\lambda_{\min}\{\sum_{i=1}^{s}\frac{\varphi(i)\varphi(i)'}{1+\|\varphi(i)\|^2}\} \geq k_1$ is not satisfied on $\mathcal{A}_h$. So

$$E\left[\left\{\lambda_{\min}\left\{\sum_{i=1}^{s}\frac{\varphi(i)\varphi(i)'}{1+\|\varphi(i)\|^2}\right\} \geq k_1\right\}\cdot 1(\mathcal{A}_h)\right] = 0 < k_2 \cdot 1(\mathcal{A}_h)$$

(where $1(\mathcal{A}_h)$ is the indicator function of set $\mathcal{A}_h$) and this contradicts condition (1).

The above example shows that the persistence of excitation condition (1) does not hold under Assumption $2'$. On the other hand, almost all results in the identification literature (like those in Guo (1990), Bittanti and Campi (1994), or Campi (1994)) have been worked out under this condition (1). Consequently, at the present state of the art, it is not clear how to handle situations where the system is only characterized by a mild stability condition like Assumption $2'$. The above observation raises an interesting conceptual question: one may ask if it is possible to work out a persistence of excitation condition milder than (1), that holds true under Assumption $2'$ and still permits one to prove boundedness results for the identification algorithms. This issue is certainly worthy of further investigation.

REFERENCES

R. Beran and P. Hall (1992), *Estimating coefficient distributions in random coefficient regressions*, Ann. Statist., 20, pp. 1970–1984.

S. Bittanti and M.C. Campi (1991a), *Adaptive RLS algorithms under stochastic excitation—L^2 convergence analysis*, IEEE Trans. Automat. Control, 36, pp. 963–967.

S. Bittanti and M.C. Campi (1991b), *Adaptive RLS algorithms under stochastic excitation—strong consistency analysis*, Systems Control Lett., 17, pp. 3–8.

S. Bittanti and M.C. Campi (1994), *Bounded error identification of time-varying parameters by RLS techniques*, IEEE Trans. Automat. Control, 39, pp. 1106–1110.

M.C. Campi (1994), *Exponentially weighted least squares identification of time-varying systems with white disturbances*, IEEE Trans. Signal Process., 42, pp. 2906–2914.

M.C. Campi (1997), *Performance of RLS identification algorithms with forgetting factor: A Φ-mixing approach*, J. Math. Systems Estim. Control, 7, pp. 29–53.

G.C. Chow (1983), *Random and changing coefficient models*, in Handbook of Econometrics, Z. Griliches and M.D. Intriligator, eds., Handbooks in Econom. 2, North-Holland, Amsterdam, pp. 1213–1245.

D.D. Cox (1991), *Gaussian likelihood estimation for nearly nonstationary AR* (1) *processes*, Ann. Statist., 19, pp. 1129–1142.

D.D. Cox and I. Llatas (1991), *Maximum likelihood type estimation for nearly nonstationary autoregressive time series*, Ann. Statist., 19, pp. 1109–1128.

L. Guo (1990), *Estimating time-varying parameters by the Kalman filter based algorithm: Stability and convergence*, IEEE Trans. Automat. Control, 35, pp. 141–147.

L. Guo and L. Ljung (1995a), *Exponential stability of general tracking algorithms*, IEEE Trans. Automat. Control, 39, pp. 1376–1387.

L. Guo and L. Ljung (1995b), *Performance analysis of general tracking algorithms*, IEEE Trans. Automat. Control, 39, pp. 1388–1401.

L. Guo, L. Ljung, and P. Priouret (1993), *Performance analysis of the forgetting factor RLS algorithm*, Internat J. Adapt. Control Signal Process., 7, pp. 525–537.

D.F. Nicholls and B.G. Quinn (1982), *Random Coefficient Autoregressive Models: An Introduction*, Springer-Verlag, New York.

R. Ravikanth and S.P. Meyn (1999), *Bounds on achievable performance in the identification and adaptive control of time-varying systems*, IEEE Trans. Automat. Control, 44, pp. 670–682.

V. Solo (1994), *On the stability of slowly time-varying linear systems*, Math. Control Signals Systems, 7, pp. 331–350.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 141–158

AN EXPLICIT FORMULA FOR THE DERIVATIVE OF A CLASS OF COST FUNCTIONALS WITH RESPECT TO DOMAIN VARIATIONS IN STOKES FLOW[*]

THOMAS SLAWIG[†]

Abstract. Domain optimization problems for the two-dimensional stationary Stokes equations are studied. Fréchet differentiability of a class of cost functionals with respect to the variation of the shape of the computational domain is established. An embedding domain technique provides an equivalent formulation of the problem on a fixed domain and, moreover, a simply computable formula for the derivative of the cost functional with respect to the domain. Existence of a solution to the class of domain optimization problems is proved. Numerical examples show the reliability of the derivative formula.

Key words. domain optimization, Stokes equations, embedding domain technique, finite element method

AMS subject classifications. 35Q30, 49J50, 65N30, 76D07

PII. S0363012998347870

1. Introduction.

In this paper we show the differentiability of a certain class of cost functionals with respect to variations in the shape of the computational domain under the constraint of the stationary Stokes equations. We consider domain optimization problems in incompressible viscous flow at low velocity or high viscosity, i.e., low Reynolds number. In this case the Stokes equations are an appropriate linear approximation of the full nonlinear Navier–Stokes equations.

Many optimization techniques are based on gradient information, i.e., they require information about the derivative of the cost functional with respect to the control parameters, i.e., here the shape of the domain. Usually this information is obtained via an adjoint equation. In domain optimization problems this equation usually incorporates the normal derivative of the state variable along the boundary which is numerically unstable. In this work we use an embedding domain technique which provides a formula for the derivative of the cost functional with respect to domain variations which avoids the evaluation of normal derivatives, is efficient and numerically stable. Moreover, the embedding domain technique reduces the effort of discretization and assembling of the discrete systems for the solution of the state equations on the changing domains during the optimization process.

Embedding (or "fictitious") domain techniques have been widely applied in the treatment of PDEs. For Stokes and Navier–Stokes equations on complicated shaped domains they were studied, e.g., by Börgers [1] and Glowinski, Pan, and Periaux [2], [3]. Our Lagrange multiplier approach is similar to Glowinski's. Dankova and Haslinger [4] used a slightly different one by introducing a distributed Lagrange multiplier and applied it on domain optimization problems; see, e.g., [4]. Domain optimization for Stokes equations were studied, for example, by Pironneau [5] who computed the shape of body with minimum drag. Gunzburger and Kim [6] showed existence of an optimal shape for a minimum drag problem in a channel flow. Simon et al. [7]

[*]Received by the editors November 20, 1998; accepted for publication (in revised form) April 14, 2000; published electronically August 17, 2000.
http://www.siam.org/journals/sicon/39-1/34787.html

[†]Potsdam Institute for Climate Impact Research, Telegrafenberg, 14412 Potsdam, Germany (slawig@pik-potsdam.de).

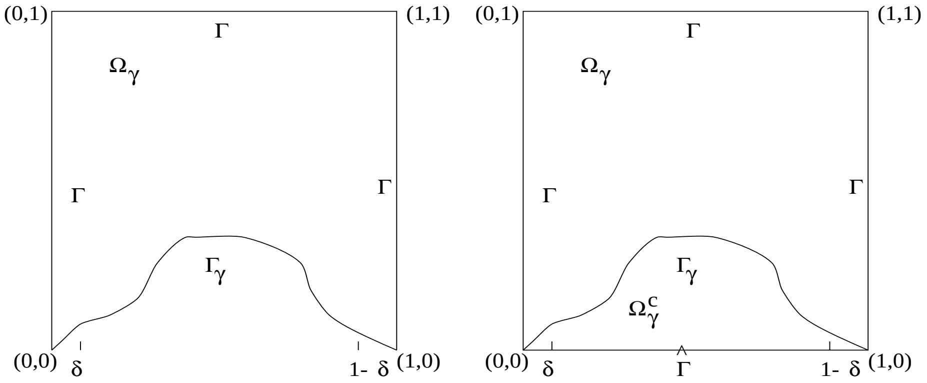


FIG. 2.1. *The domain Ω_γ in original version (left) and embedded into $\hat{\Omega}$ (right).*

proved differentiability of the drag with respect to domain variations in Navier–Stokes flow.

Here we use the embedding domain technique not only to prove differentiability of a class of cost functionals (including drag), but moreover to obtain an explicit formula for the derivative. This formula is a boundary integral which specifically does not involve normal derivatives of the state variables. This is due to a special choice of extension for the inhomogeneity of the state equation in the embedding domain method. Thus the formula is useful and fast in numerical optimization schemes using derivative information as gradient, quasi-Newton, or SQP methods. By the same technique Kunisch and Peichl [8] obtained a derivative formula for the scalar Poisson problem. In this paper we present its analogue for the two-dimensional Stokes problem with its special saddle-point structure. An extension to the full nonlinear Navier–Stokes case is possible and might be presented in another paper.

In the next two sections we define a geometric model configuration and summarize the needed results for the Stokes equations. The considered class of domain optimization problems and the embedding domain technique are presented in the next two sections. In the sixth section we show the continuous dependence of the solution of the Stokes system with respect to the variation of the domain. The main result, the explicit formula for the Fréchet derivative, is presented in section 7. Finally, we summarize the numerical methods we used and show an example.

2. The geometric model configuration. We now present a model configuration of the geometry of the computational domain. We consider domains $\Omega_\gamma := \Omega(\gamma) \subset \mathbb{R}^2$, where γ is the control parameter describing the shape of the domain. For all admissible domains the boundary $\partial\Omega_\gamma$ shall have two parts: a fixed one denoted by Γ which consists of the two lateral and the top side of the unit square, i.e., of the three segments $[(0,0),(0,1)],[(0,1),(1,1)],[(1,1),(1,0)]$, and a variable one denoted by Γ_γ which is the graph of a function $\gamma : [0,1] \to [0,1)$ with $\gamma(0) = \gamma(1) = 0$; compare Figure 2.1 left.

A sufficiently high regularity of the solution of the Stokes equations is one necessary condition for our proof of differentiability of the considered cost functionals with respect to domain variation. We thus need either a smooth or a polygonal boundary, where in the latter case the domain has to be convex. To apply the embedding domain method we have to combine these two boundary types which will become clearer later on. By the above definition Γ is a polygon. To represent the variable part Γ_γ we choose functions $\gamma \in C^2[0,1]$. To preserve the convexity of Ω_γ near the two transition points $(0,0),(1,0)$ we assume that γ is linear in neighborhoods of these points. Working in Sobolev spaces we assure the regularity by choosing $\gamma \in H^3(I)$ with $I := (0,1)$. To show existence of a solution of the domain optimization problems we assume boundedness in this space. We define the set of admissible functions γ defining the variable boundary parts Γ_γ, and thus the admissible domains Ω_γ, by

$$\begin{aligned}(2.1)\quad \mathcal{S} := \{\gamma \in H^3(I) : \|\gamma\|_{H^3(I)} \le c_0, \gamma(0) = \gamma(1) = 0, c_1 \le \gamma|_{(\delta,1-\delta)} \le c_2,\\ \gamma'|_{(0,\delta)} = c_3, \gamma'|_{(1-\delta,1)} = c_4\}.\end{aligned}$$

Here $c_1, c_2 \in (0,1), \delta \in (0,\frac{1}{2}), c_0, c_3 \in \mathbb{R}^+, c_4 \in \mathbb{R}^-$ are fixed.

3. The Stokes equations. The stationary Stokes problem on Ω_γ can be written in the following variational form: Find the pair of velocity vector and pressure $(\mathbf{u}_\gamma, p_\gamma) \in H^1(\Omega_\gamma)^2 \times L^2_0(\Omega_\gamma)$ such that

$$(3.1)\qquad \begin{array}{rcll} \nu(\nabla\mathbf{u}_\gamma, \nabla\mathbf{v})_{\Omega_\gamma} - (p_\gamma, \operatorname{div}\mathbf{v})_{\Omega_\gamma} & = & (\mathbf{f}_\gamma, \mathbf{v})_{\Omega_\gamma} & \text{for all} \quad \mathbf{v} \in H^1_0(\Omega_\gamma)^2, \\ (\operatorname{div}\mathbf{u}_\gamma, q)_{\Omega_\gamma} & = & 0 & \text{for all} \quad q \in L^2_0(\Omega_\gamma), \\ \mathbf{u}_\gamma & = & \Phi & \text{on } \Gamma, \\ \mathbf{u}_\gamma & = & \mathbf{0} & \text{on } \Gamma_\gamma. \end{array}$$

In this dimension-free formulation the parameter $\nu > 0$ represents the inverse of the Reynolds number. For scalar-valued functions $(\cdot,\cdot)_{\Omega_\gamma}$ denotes the $L^2(\Omega_\gamma)$ inner product. Furthermore, we define $(\mathbf{u},\mathbf{v})_{\Omega_\gamma} := \sum_{i=1,2}(u_i,v_i)_{\Omega_\gamma}$ and $(\nabla\mathbf{u},\nabla\mathbf{v})_{\Omega_\gamma} := \sum_{i,j=1,2}(\frac{\partial u_i}{\partial x_j}, \frac{\partial v_i}{\partial x_j})_{\Omega_\gamma}$. For the inhomogeneity in the first equation which represents external forces we assume $\mathbf{f}_\gamma \in L^2(\Omega_\gamma)^2$. The space $L^2_0(\Omega_\gamma) := \{q \in L^2(\Omega_\gamma) : \int_{\Omega_\gamma} q\,dx = 0\}$ is chosen to get uniqueness of the pressure.

The homogeneous or inhomogeneous Dirichlet boundary conditions indicate that physically the boundary represents a wall or a region with prescribed velocity, respectively. The function Φ for the boundary values of the velocity on Γ is assumed to have a divergence-free extension onto Ω_γ which is in $H^2(\Omega_\gamma)^2$. For this purpose we define

$$H(\Gamma) := \left\{\Phi \in L^2(\Gamma)^2 : \text{there is}\, \bar{\mathbf{u}}_\gamma \in H^2(\Omega_\gamma)^2 : \operatorname{div}\bar{\mathbf{u}}_\gamma = 0 \text{ in}\, \Omega_\gamma, \bar{\mathbf{u}}_\gamma|_{\Gamma_\gamma} = \mathbf{0}, \bar{\mathbf{u}}_\gamma|_\Gamma = \Phi\right\}.$$

We now slightly modify the standard existence, uniqueness, and regularity results for the Stokes equations in the following theorem.

THEOREM 3.1. *Let $\gamma \in \mathcal{S}, \mathbf{f}_\gamma \in L^2(\Omega_\gamma)^2$, and $\Phi \in H(\Gamma)$. Then there exists a unique solution $(\mathbf{u}_\gamma, p_\gamma) \in H^2(\Omega_\gamma)^2 \times [H^1(\Omega_\gamma) \cap L^2_0(\Omega_\gamma)]$ to* (3.1) *which satisfies*

$$\|\mathbf{u}_\gamma\|_{H^2(\Omega_\gamma)^2} + \|p_\gamma\|_{H^1(\Omega_\gamma)} \le C\left(\|\mathbf{f}_\gamma\|_{L^2(\Omega_\gamma)^2} + \|\Phi\|_{L^\infty(\Gamma)^2}\right)$$

with $C > 0$ independent of $\gamma, \mathbf{f}_\gamma$, and Φ, i.e., the regularity is uniform in γ.

Proof. Regularity for a smooth boundary and uniqueness are stated, e.g., in [11, Thm. I.5.4]. For a convex polygonal boundary, regularity is shown in [12], from which it can be deduced that the regularity remains valid also in our case where both boundary types are mixed. The uniform regularity for the polygonal part is stated in the same reference, whereas for the smooth part it can be deduced from [10, Sect. IV.5]. □

4. A class of domain optimization problems. We consider domain optimization problems of the following form:

$$(4.1)\qquad \min_{\gamma\in\mathcal{S}} \mathcal{J}(\gamma) := \min_{\gamma\in\mathcal{S}} \frac{1}{2}\|\mathcal{A}\mathbf{u}_\gamma - \mathbf{u}_d\|^2_{L^2(\Omega_C)^k} \quad \text{such that} \quad \mathbf{u}_\gamma \text{ solves (3.1)}$$

with an *observation operator* $\mathcal{A} \in \mathcal{L}(H^1(\Omega_\gamma)^2, L^2(\Omega_C)^k)$ for $k \in \{1,2\}$, some *desired state* $\mathbf{u}_d \in L^2(\Omega_C)^k$, and an *observation region* Ω_C satisfying $\operatorname{dist}(\Gamma_\gamma, \Omega_C) > 0$ for all $\gamma \in \mathcal{S}$. The dependence of $\mathcal{J}$ on γ is implicit due to the fact that $\mathcal{J}$ depends on $\mathbf{u}_\gamma$ which itself depends on γ. The functional $\mathcal{J}$ may also include an additional regularization term which we exclude from our theoretical study because it usually depends directly on γ and not by means of the solution $\mathbf{u}_\gamma$ of the state equations. Therefore its differentiability is obtained easily. The above definition of the cost functional is quite general. It includes typical choices as the *tracking-type* functional

$$\mathcal{J}(\gamma) := \frac{1}{2}\|\mathbf{u}_\gamma - \mathbf{u}_d\|^2_{L^2(\Omega_C)^2}$$

or the *minimum drag problem*; compare Pironneau [5], Gunzburger and Kim [6]:

$$\mathcal{J}(\gamma) := \frac{\nu}{2}\|\nabla \mathbf{u}_\gamma + \nabla^T \mathbf{u}_\gamma\|^2_{L^2(\Omega_C)^{2\times 2}}.$$

A restriction we have to make for our derivative formula is that $\mathcal{J}$ does not depend on the pressure.

5. The embedding domain technique. To solve the domain optimization problem (4.1) by a classical iterative scheme it is necessary to discretize the domain, assemble the system matrices, and solve the linear system in each iteration step. To reduce this effort we introduce a *fictitious domain* $\hat{\Omega}$ in which all admissible domains can be embedded, i.e., $\Omega_\gamma \subset \hat{\Omega}$ for all $\gamma \in \mathcal{S}$. We furthermore assume that the fixed boundary part Γ is a part of $\partial\hat{\Omega}$ whereas Γ_γ is replaced by a partition called $\hat{\Gamma}$ which is now fixed as well. Thus we have $\partial\hat{\Omega} = \bar{\Gamma} \cup \bar{\hat{\Gamma}}$; compare Figure 2.1 right. In our model problem we take $\hat{\Omega}$ as the unit square and define $\Omega^c_\gamma := \hat{\Omega} \setminus \bar{\Omega}_\gamma$.

We now formulate a problem on $\hat{\Omega}$ which is equivalent to (4.1). For this purpose we introduce a *fictitious domain formulation* of the state equations. As motivation we recall that the velocity part $\mathbf{u}_\gamma$ of the solution of the Stokes problem (3.1) is the solution of the constrained minimization problem

$$\min_{\substack{\mathbf{u} \in H^1(\Omega_\gamma)^2 \\ \mathbf{u}|_\Gamma = \Phi, \mathbf{u}|_{\Gamma_\gamma} = \mathbf{0}}} \frac{\nu}{2}\|\nabla \mathbf{u}\|^2_{L^2(\Omega_\gamma)^2} - (\mathbf{f}_\gamma, \mathbf{u})_{\Omega_\gamma} \quad \text{s.t.} \quad \operatorname{div} \mathbf{u} = 0 \ \text{ in } \Omega_\gamma.$$

Equations (3.1) are the necessary conditions for a saddle point $(\mathbf{u}_\gamma, p_\gamma)$ of the associated Lagrangian with p_γ being the Lagrange multiplier corresponding to the constraint of zero divergence.

We now denote by $\tilde{\mathbf{f}}_\gamma$ the extension of $\mathbf{f}_\gamma$ by zero onto $\hat{\Omega}$ and consider the problem

$$\min_{\substack{\hat{\mathbf{u}} \in H^1(\hat{\Omega})^2 \\ \hat{\mathbf{u}}|_\Gamma = \Phi, \hat{\mathbf{u}}|_{\hat{\Gamma}} = \mathbf{0}}} \frac{\nu}{2}\|\nabla \hat{\mathbf{u}}\|^2_{L^2(\hat{\Omega})^2} - (\tilde{\mathbf{f}}_\gamma, \hat{\mathbf{u}})_{\hat{\Omega}} \quad \text{s.t.} \quad \left\{ \begin{array}{rcll} \operatorname{div} \hat{\mathbf{u}} & = & 0 & \text{in } \hat{\Omega}, \\ \hat{\mathbf{u}} & = & \mathbf{0} & \text{on } \Gamma_\gamma. \end{array} \right. \tag{5.1}$$

The second constraint is added since Γ_γ is no longer a part of the boundary of the computational domain $\hat{\Omega}$ but an inner line. Thus we get a second Lagrange multiplier g_γ which is an element of the dual of

$$H_\gamma := H^{1/2}_{00}(\Gamma_\gamma)^2 = \left\{ \mathbf{h} \in H^{1/2}(\Gamma_\gamma)^2 : \text{there is } \tilde{\mathbf{h}} \in H^{1/2}(\partial\Omega_\gamma)^2 : \tilde{\mathbf{h}}|_{\Gamma_\gamma} = \mathbf{h}, \tilde{\mathbf{h}}|_\Gamma = \mathbf{0} \right\};$$

compare, e.g., [9, VII, Sect. 2.1, Rem. 1]. The necessary conditions for a saddle point of the Lagrangian associated with problem (5.1) result in the following equations:

Find $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma, g_\gamma) \in H^1(\hat{\Omega})^2 \times L^2_0(\hat{\Omega}) \times H^*_\gamma$ such that

$$\begin{aligned}
\nu(\nabla\hat{\mathbf{u}}_\gamma, \nabla\hat{\mathbf{v}})_{\hat{\Omega}} - (\hat{p}_\gamma, \operatorname{div}\hat{\mathbf{v}})_{\hat{\Omega}} - \langle g_\gamma, \tau_\gamma\hat{\mathbf{v}}\rangle_{H^*_\gamma, H_\gamma} &= (\tilde{\mathbf{f}}_\gamma, \hat{\mathbf{v}})_{\hat{\Omega}} && \text{for all} \quad \hat{\mathbf{v}} \in H^1_0(\hat{\Omega})^2,\\
(\operatorname{div}\hat{\mathbf{u}}_\gamma, \hat{q})_{\hat{\Omega}} &= 0 && \text{for all} \quad \hat{q} \in L^2_0(\hat{\Omega}),\\
\tau_\gamma\hat{\mathbf{u}}_\gamma &= \mathbf{0}, \\
\mathbf{u}_\gamma &= \Phi && \text{on } \Gamma,\\
\mathbf{u}_\gamma &= \mathbf{0} && \text{on } \hat{\Gamma}.
\end{aligned} \tag{5.2}$$

Here τ_γ denotes the trace operator onto Γ_γ which is a linear continuous mapping from $H^1_0(\hat{\Omega})^2$ onto H_γ. The dual pairing between H_γ and its dual is denoted by $\langle\cdot,\cdot\rangle_{H^*_\gamma, H_\gamma}$. We can now prove the equivalence of problems (3.1) and (5.2) as follows.

THEOREM 5.1. *Let $\gamma \in \mathcal{S}, \mathbf{f}_\gamma \in L^2(\Omega_\gamma)^2$, and $\Phi \in H(\Gamma)$. Then $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma, g_\gamma) \in H^1(\hat{\Omega})^2 \times L^2_0(\hat{\Omega}) \times H^*_\gamma$ is a solution of* (5.2) *if and only if*

- $(\mathbf{u}_\gamma, p_\gamma) := (\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)|_{\Omega_\gamma} \in H^2(\Omega_\gamma)^2 \times [H_1(\Omega_\gamma) \cap L^2_0(\Omega_\gamma)]$ *solves* (3.1),
- $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)|_{\Omega^c_\gamma} = (\mathbf{0}, 0)$,
- $\langle g_\gamma, \mathbf{h}\rangle_{H^*_\gamma, H_\gamma} = (\nu\frac{\partial\mathbf{u}_\gamma}{\partial\mathbf{n}_\gamma} - p_\gamma\mathbf{n}_\gamma, \mathbf{h})_{\Gamma_\gamma}$ *for all* $\mathbf{h} \in H_\gamma$.

Here $\mathbf{n}_\gamma$ denotes the outer (with respect to Ω_γ) normal vector on Γ_γ and

$$(\mathbf{g}, \mathbf{h})_{\Gamma_\gamma} := \int_I \mathbf{g}(x, \gamma(x)) \cdot \mathbf{h}(x, \gamma(x))\sqrt{1 + \gamma'(x)^2}\, dx$$

is the inner product on $L^2(\Gamma_\gamma)^2$.

Proof. The result is proved by testing (3.1) with appropriate functions that vanish on Ω^c_γ, applying the uniqueness result for the Stokes equations and Green's formula. Regularity of g_γ follows from Theorem 3.1. For the existence of the solution to (5.2) we need the nonsmooth transition between Γ and Γ_γ. Otherwise Ω^c_γ would not even be Lipschitz. For details see [15, Thms. 2.5 and 3.5]. □

By the regularity of $(\mathbf{u}_\gamma, p_\gamma)$ the functional $g_\gamma \in H^*_\gamma$ can be extended onto $L^2(\Gamma_\gamma)^2$ and we get

$$g_\gamma = \left(\nu\frac{\partial\mathbf{u}_\gamma}{\partial\mathbf{n}_\gamma} - p_\gamma\mathbf{n}_\gamma\right)\Bigg|_{\Gamma_\gamma} \qquad \text{in } L^2(\Gamma_\gamma)^2. \tag{5.3}$$

This relation is due to the extension of the inhomogeneity $\mathbf{f}_\gamma$ by zero onto $\hat{\Omega}$. For an arbitrary L^2 extension g_γ equals the jump of the right-hand side of (5.3) on Γ_γ; see also [3]. The equality in (5.3) is essential for the derivative formula presented below.

As a consequence of Theorems 3.1 and 5.1 we now obtain for the solutions on $\hat{\Omega}$.

COROLLARY 5.2. *Let for $\gamma \in \mathcal{S}$ denote $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma, g_\gamma)$ a solution of* (5.2). *Then the families $\{(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma)\}_{\gamma\in\mathcal{S}}$, $\{\|g_\gamma\|_{H^*_\gamma}\}_{\gamma\in\mathcal{S}}$, and $\{\|g_\gamma\|_{L^2(\Gamma_\gamma)^2}\}_{\gamma\in\mathcal{S}}$ are uniformly bounded in $H^1(\hat{\Omega})^2 \times L^2_0(\hat{\Omega})$ and $\mathbb{R}$, respectively.*

Proof. The estimates for $\hat{\mathbf{u}}_\gamma$ and $\hat{p}_\gamma$ follow from Theorem 3.1 and the fact that the solutions vanish on Ω^c_γ; those for the Lagrange multiplier follow from its representation given in Theorem 5.1 and the uniform boundedness of $\mathbf{u}_\gamma$ and p_γ. □

6. Continuous dependence of the solution on the shape of the domain.

To study convergence with respect to γ of the Lagrange multipliers $g_\gamma \in H^*_\gamma$ we introduce for $\mathbf{h} \in H_\gamma$ the mapping $\mathcal{I}_\gamma\mathbf{h}(x) := \mathbf{h}(x, \gamma(x)), x \in I$, which is an isomorphism between H_γ and

$$H_I := \left\{\mathbf{g} \in H^{1/2}(I)^2 : \int_I \frac{\|\mathbf{g}(x)\|_2^2}{x(1-x)}\, dx < \infty\right\}.$$

A function $\mathbf{h} \in H_\gamma$ has to satisfy additional conditions at the transition points between Γ and Γ_γ such that its extension $\tilde{\mathbf{h}}$ by zero onto $\partial\Omega_\gamma$ is in $H^{1/2}(\partial\Omega_\gamma)^2$: The integral

$$\int_{\tilde{I}}\int_{\tilde{I}} \frac{\|\tilde{\mathbf{h}}(x,\gamma(x)) - \tilde{\mathbf{h}}(\xi,\gamma(\xi))\|_2^2}{|x-\xi|^2}\, d\tilde{\gamma}(x) d\tilde{\gamma}(\xi)$$

has to be finite for arbitrary functions $\tilde{\gamma} : \tilde{I} \to \tilde{\gamma}(\tilde{I}) \subset \partial\Omega_\gamma$ that parameterize a part of the boundary. Critical are parameterizations which contain one of the transition points between Γ and Γ_γ. To guarantee the existence of the above integral the additional integral condition in the definition of H_I has to be satisfied; compare [15, Thm. 2.4].

We note a result concerning convergence of the transformed trace operators below.

Lemma 6.1. *Let $\gamma_n, \gamma \in \mathcal{S}$. Then $\gamma_n \to \gamma$ in $W^{1,\infty}(I)$ implies $\mathcal{I}_{\gamma_n}\tau_{\gamma_n} \to \mathcal{I}_\gamma\tau_\gamma$ as linear operators from $\{\hat{\mathbf{v}} \in H^1(\hat{\Omega})^2 : \hat{\mathbf{v}}|_{\hat{\Gamma}} = \mathbf{0}\}$ onto H_I.*

Proof. See [15, Lem. 2.1]. □

We define $\left(\mathcal{I}_\gamma^{-1}\right)^* : H_\gamma^* \to H_I^*$, i.e., the adjoint of $\mathcal{I}_\gamma^{-1}$, by

$$\langle \left(\mathcal{I}_\gamma^{-1}\right)^* g, \mathbf{g}\rangle_{H_I^*,H_I} := \langle g, \mathcal{I}_\gamma^{-1}\mathbf{g}\rangle_{H_\gamma^*,H_\gamma}, \qquad g \in H_\gamma^*,\ \mathbf{g} \in H_I.$$

Now we can formulate the following result of continuous dependence.

Theorem 6.2. *Let $\gamma, \gamma_n \in \mathcal{S}$ with $\gamma_n \to \gamma$ in $W^{1,\infty}(I), \mathbf{f} \in L^\infty(\hat{\Omega})^2$, and $\mathbf{f}_\gamma := \mathbf{f}|_{\Omega_\gamma}, \mathbf{f}_{\gamma_n} := \mathbf{f}|_{\Omega_{\gamma_n}}$. Then the corresponding solutions of problem (5.2) satisfy*

$$\begin{array}{rcll} (\hat{\mathbf{u}}_{\gamma_n}, \hat{p}_{\gamma_n}) & \to & (\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma) & \text{in } H^1(\hat{\Omega})^2 \times L_0^2(\hat{\Omega}), \\ \left(\mathcal{I}_{\gamma_n}^{-1}\right)^* g_{\gamma_n} & \overset{*}{\rightharpoonup} & \left(\mathcal{I}_\gamma^{-1}\right)^* g_\gamma & \text{in } H_I^*. \end{array}$$

Moreover, the mapping $\gamma \mapsto \hat{\mathbf{u}}_\gamma$ is Lipschitz continuous, i.e., there exists L independent of $\gamma, \bar{\gamma}$ such that

$$\|\hat{\mathbf{u}}_{\bar{\gamma}} - \hat{\mathbf{u}}_\gamma\|_{H^1(\hat{\Omega})^2} \le L\|\bar{\gamma} - \gamma\|_{L^\infty(I)} \qquad \text{for all } \bar{\gamma}, \gamma \in \mathcal{S}.$$

Proof. By Corollary 5.2 we have for a subsequence $(\hat{\mathbf{u}}_{\gamma_n}, \hat{p}_{\gamma_n}) \rightharpoonup (\hat{\mathbf{u}}, \hat{p})$ weakly in $H^1(\hat{\Omega})^2 \times L_0^2(\hat{\Omega})$ and thus using the first equation of (5.2),

$$\lim_{n\to\infty} \langle g_{\gamma_n}, \tau_{\gamma_n}\hat{\mathbf{v}}\rangle_{H_{\gamma_n}^*,H_{\gamma_n}} = \nu(\nabla\hat{\mathbf{u}}, \nabla\hat{\mathbf{v}})_{\hat{\Omega}} - (\hat{p}, \operatorname{div}\hat{\mathbf{v}})_{\hat{\Omega}} - (\tilde{\mathbf{f}}_\gamma, \hat{\mathbf{v}})_{\hat{\Omega}}$$

for all $\hat{\mathbf{v}} \in H_0^1(\hat{\Omega})^2$. For every $\mathbf{g} \in H_I$ there exists $\hat{\mathbf{v}} \in H_0^1(\hat{\Omega})^2$ with $\mathcal{I}_\gamma\tau_\gamma\hat{\mathbf{v}} = \mathbf{g}$ and

$$\langle (\mathcal{I}_{\gamma_n}^{-1})^* g_{\gamma_n}, \mathbf{g}\rangle_{H_I^*,H_I} = \langle (\mathcal{I}_{\gamma_n}^{-1})^* g_{\gamma_n}, \mathcal{I}_\gamma\tau_\gamma\hat{\mathbf{v}} - \mathcal{I}_{\gamma_n}\tau_{\gamma_n}\hat{\mathbf{v}}\rangle_{H_I^*,H_I} + \langle g_{\gamma_n}, \tau_{\gamma_n}\hat{\mathbf{v}}\rangle_{H_{\gamma_n}^*,H_{\gamma_n}},$$

where the first term on the right tends to zero by Corollary 5.2 and Lemma 6.1. Thus

$$\lim_{n\to\infty} \langle (\mathcal{I}_{\gamma_n}^{-1})^* g_{\gamma_n}, \mathbf{g}\rangle_{H_I^*,H_I} = \nu(\nabla\hat{\mathbf{u}}, \nabla\hat{\mathbf{v}})_{\hat{\Omega}} - (\hat{p}, \operatorname{div}\hat{\mathbf{v}})_{\hat{\Omega}} - (\tilde{\mathbf{f}}_\gamma, \hat{\mathbf{v}})_{\hat{\Omega}} =: \langle G, \mathbf{g}\rangle_{H_I^*,H_I}.$$

We define $g := \mathcal{I}_\gamma^* G \in H_\gamma^*$. For $\hat{\mathbf{v}} \in H_0^1(\hat{\Omega})^2$ we have $\tau_\gamma\hat{\mathbf{v}} \in H_\gamma$ and thus $(\hat{\mathbf{u}}, \hat{p}, g)$ satisfies the first equation of (5.2). Green's formula implies

$$0 = \lim_{n\to\infty} (\operatorname{div}\hat{\mathbf{u}}_{\gamma_n}, \hat{\varphi})_{\hat{\Omega}} = -\lim_{n\to\infty} (\hat{\mathbf{u}}_{\gamma_n}, \nabla\hat{\varphi})_{\hat{\Omega}} = -(\hat{\mathbf{u}}, \nabla\hat{\varphi})_{\hat{\Omega}} = (\operatorname{div}\hat{\mathbf{u}}, \hat{\varphi})_{\hat{\Omega}}$$

for all $\hat{\varphi} \in C_0^\infty(\hat{\Omega}) \cap L_0^2(\hat{\Omega})$. Thus $\hat{\mathbf{u}}$ solves the second equation of (5.2). Since it can be shown that also $\tau_\gamma\hat{\mathbf{u}} = \mathbf{0}$ (see [15, Lem. 2.12]) this implies that $(\hat{\mathbf{u}}, \hat{p}, g)$ is the unique solution to problem (5.2) and thus equals $(\hat{\mathbf{u}}_\gamma, \hat{p}_\gamma, g_\gamma)$. Therefore the limits are valid for the whole sequence and not only for the chosen subsequence.

To show strong convergence of $\hat{\mathbf{u}}_{\gamma_n}$ in $H^1(\hat\Omega)^2$ we define $\hat{\mathbf{u}}_n := \hat{\mathbf{u}}_{\gamma_n} - \hat{\mathbf{u}}_\gamma \in H^1_0(\hat\Omega)^2$ which satisfies $(\operatorname{div}\hat{\mathbf{u}}_n, \hat{q})_{\hat\Omega} = 0$ for all $\hat{q} \in L^2_0(\hat\Omega)$. Choosing $\hat{\mathbf{v}} := \hat{\mathbf{u}}_n$ in the first equation of (5.2) for Ω_{γ_n} and Ω_γ and subtracting both equations we get

$$\nu|\hat{\mathbf{u}}_n|^2_{H^1(\hat\Omega)^2} = (\tilde{\mathbf{f}}_{\gamma_n} - \tilde{\mathbf{f}}_\gamma, \hat{\mathbf{u}}_n)_{\hat\Omega} + \langle g_{\gamma_n}, \tau_{\gamma_n}\hat{\mathbf{u}}_n\rangle_{H^*_{\gamma_n},H_{\gamma_n}} - \langle g_\gamma, \tau_\gamma\hat{\mathbf{u}}_n\rangle_{H^*_\gamma,H_\gamma}, \tag{6.1}$$

where $|\hat{\mathbf{u}}|_{H^1(\hat\Omega)^2} := \|\nabla\hat{\mathbf{u}}\|_{L^2(\hat\Omega)^{2\times 2}}$ denotes the H^1 seminorm. We estimate the terms on the right of (6.1) separately: First we split up I into

$$I_+ := \{x \in (0,1) : \gamma_n(x) \geq \gamma(x)\}, \quad I_- := \{x \in (0,1) : \gamma_n(x) < \gamma(x)\}$$

and obtain, using the fact that $\mathbf{f}_\gamma := \mathbf{f}|_{\Omega_\gamma}$ with $\mathbf{f} \in L^\infty(\hat\Omega)^2$ for all $\gamma \in \mathcal{S}$, that

$$(\tilde{\mathbf{f}}_{\gamma_n} - \tilde{\mathbf{f}}_\gamma, \hat{\mathbf{u}}_n)_{\hat\Omega} = \int_{I_+}\int_\gamma^{\gamma_n} \mathbf{f}(x,y)\cdot\hat{\mathbf{u}}_n(x,y)\,dydx - \int_{I_-}\int_{\gamma_n}^{\gamma} \mathbf{f}(x,y)\cdot\hat{\mathbf{u}}_n(x,y)\,dydx.$$

We denote the first integral on the right by A and use $\hat{\mathbf{u}}_n(x,\gamma(x)) = \mathbf{0}$ almost everywhere (a.e.) in I_+:

$$\begin{aligned} |A| &= \left| \int_{I_+}\int_\gamma^{\gamma_n} \mathbf{f}(x,y)\cdot[\hat{\mathbf{u}}_n(x,y) - \hat{\mathbf{u}}_n(x,\gamma)]\;dydx \right| \\ &\leq \|\mathbf{f}\|_{L^\infty(\hat\Omega)^2} \int_{I_+}\int_\gamma^{\gamma_n}\int_\gamma^y \|\hat{\mathbf{u}}_{n,y}(x,\xi)\|_2\,d\xi dydx \\ &\leq \|\mathbf{f}\|_{L^\infty(\hat\Omega)^2} \int_{I_+}\int_\gamma^{\gamma_n} \|\hat{\mathbf{u}}_{n,y}(x,\xi)\|_2(\gamma_n - \xi)\,dx \\ &\leq \frac{1}{3}\|\mathbf{f}\|_{L^\infty(\hat\Omega)^2}\|\hat{\mathbf{u}}_{n,y}\|_{L^2(\hat\Omega)^2}\|\gamma_n - \gamma\|^{3/2}_{L^\infty(I_+)} \;\leq\; C|\hat{\mathbf{u}}_n|_{H^1(\hat\Omega)^2}\|\gamma_n - \gamma\|^{3/2}_{L^\infty(I)} \end{aligned} \tag{6.2}$$

with C independent of γ_n, γ. The second integral can be estimated in a similar way using $\hat{\mathbf{u}}_n(x,\gamma_n(x)) = \mathbf{0}$ a.e. in I_-. Thus there exists L_1 independent of γ_n, γ with

$$|(\tilde{\mathbf{f}}_{\gamma_n} - \tilde{\mathbf{f}}_\gamma, \hat{\mathbf{u}}_n)_{\hat\Omega}| \leq L_1|\hat{\mathbf{u}}_n|_{H^1(\hat\Omega)^2}\|\gamma_n - \gamma\|^{3/2}_{L^\infty(I)}. \tag{6.3}$$

For the second term in (6.1) we obtain using Theorem 5.1:

$$|\langle g_{\gamma_n}, \tau_{\gamma_n}\hat{\mathbf{u}}_n\rangle_{H^*_{\gamma_n},H_{\gamma_n}}| = |(g_{\gamma_n}, \tau_{\gamma_n}\hat{\mathbf{u}}_n)_{\Gamma_{\gamma_n}}| \;\leq\; \|\mathbf{g}_{\gamma_n}\|_{L^2(\Gamma_{\gamma_n})}\|\tau_{\gamma_n}\hat{\mathbf{u}}_n\|_{L^2(\Gamma_{\gamma_n})^2}.$$

The first term on the right-hand side is bounded independently of γ_n by Theorem 5.2. For the last term we use $\hat{\mathbf{u}}_n(x,\gamma_n(x)) = \mathbf{0}$ a.e. in I_- and $\hat{\mathbf{u}}_n(x,\gamma(x)) = \mathbf{0}$ a.e. in I_+:

$$\begin{aligned} \|\tau_{\gamma_n}\hat{\mathbf{u}}_n\|^2_{L^2(\Gamma_{\gamma_n})^2} &= \int_{I_+} \|\hat{\mathbf{u}}_n(x,\gamma_n(x)) - \hat{\mathbf{u}}_n(x,\gamma(x))\|^2_2\sqrt{1+\gamma_n'(x)^2}\,dx \\ &= \int_{I_+} \left\| \int_\gamma^{\gamma_n} \hat{\mathbf{u}}_{n,y}(x,\xi)\,d\xi \right\|^2_2 \sqrt{1+\gamma_n'(x)^2}\,dx \\ &\leq \|\gamma_n - \gamma\|^2_{L^\infty(I)}|\hat{\mathbf{u}}_n|^2_{H^1(\hat\Omega)^2}\|(1 + \|\gamma_n\|^2_{W^{1,\infty}(I)}). \end{aligned}$$

Because $\mathcal{S}$ is bounded in $H^3(I)$ which is continuously embedded in $W^{1,\infty}(I)$ there exists L_2 independent of γ, γ_n such that

$$|\langle g_{\gamma_n}, \tau_{\gamma_n}\hat{\mathbf{u}}_n\rangle_{H^*_{\gamma_n},H_{\gamma_n}}| \leq L_2|\hat{\mathbf{u}}_n|_{H^1(\hat\Omega)^2}\|\gamma_n - \gamma\|_{L^\infty(I)}. \tag{6.4}$$

An analogous computation for the third term on the right-hand side of (6.1) shows the existence of L_3 independent of γ, γ_n with

$$|\langle g_\gamma, \tau_\gamma \hat{\mathbf{u}}_n \rangle_{H_\gamma^*, H_\gamma}| \le L_3 |\hat{\mathbf{u}}_n|_{H^1(\hat\Omega)^2} \|\gamma_n - \gamma\|_{L^\infty(I)}. \tag{6.5}$$

Poincaré's inequality implies $\|\hat{\mathbf{u}}_n\|_{H^1(\hat\Omega)^2} \le c |\hat{\mathbf{u}}_n|_{H^1(\hat\Omega)^2}$ with c independent of γ, γ_n and thus (6.3)–(6.5) give Lipschitz continuity of $\hat{\mathbf{u}}_\gamma$ with respect to γ.

To show strong convergence $p_{\gamma_n} \to p_\gamma$ we note that the operator $(\operatorname{div}, \tau_\gamma)$ maps $H_0^1(\hat\Omega)^2$ onto $L_0^2(\hat\Omega) \times H_\gamma$, see [15, Thm. 2.5]. Moreover, the weak divergence operator is an isomorphism between the orthogonal complement of $\{\hat{\mathbf{v}} \in H_0^1(\hat\Omega)^2 : \operatorname{div} \hat{\mathbf{v}} = 0 \text{ in } \hat\Omega\}$ and $L_0^2(\hat\Omega)$. Thus there exists $\hat{\mathbf{v}}_n \in H_0^1(\hat\Omega)^2$ satisfying $\tau_\gamma \hat{\mathbf{v}}_n = \mathbf{0}, \operatorname{div} \hat{\mathbf{v}}_n = \hat p_{\gamma_n} - \hat p_\gamma$, and $\|\hat{\mathbf{v}}_n\|_{H^1(\hat\Omega)^2} \le c \|\hat p_{\gamma_n} - \hat p_\gamma\|_{L^2(\hat\Omega)}$. We test the first equation of (5.2) with $\hat{\mathbf{v}}_n$ for γ and γ_n, and subtract both equations. With the uniform boundedness of $\{\hat p_{\gamma_n}\}_{n \in \mathbb{N}}$ in $L^2(\hat\Omega)$ we obtain

$$\|\hat p_{\gamma_n} - \hat p_\gamma\|_{0,\hat\Omega}^2 \le c \left(\nu |\hat{\mathbf{u}}_{\gamma_n} - \hat{\mathbf{u}}_\gamma|_{H^1(\hat\Omega)^2} + \|\tilde{\mathbf{f}}_{\gamma_n} - \tilde{\mathbf{f}}_\gamma\|_{L^2(\hat\Omega)^2} \right) + |\langle g_{\gamma_n}, \tau_{\gamma_n} \hat{\mathbf{v}}_n \rangle_{H_{\gamma_n}^*, H_{\gamma_n}}|$$

with c independent of n. Since $\hat{\mathbf{u}}_{\gamma_n} \to \hat{\mathbf{u}}_\gamma$ in $H^1(\Omega_\gamma)^2$ and $\tilde{\mathbf{f}}_{\gamma_n} \to \tilde{\mathbf{f}}_\gamma$ in $L^2(\hat\Omega)^2$ the first term tends to zero. In the second one we may write

$$\langle g_{\gamma_n}, \tau_{\gamma_n} \hat{\mathbf{v}}_n \rangle_{H_{\gamma_n}^*, H_{\gamma_n}} = \langle (\mathcal{I}_{\gamma_n}^{-1})^* g_{\gamma_n}, \mathcal{I}_{\gamma_n} \tau_{\gamma_n} \hat{\mathbf{v}}_n \rangle_{H_I^*, H_I}. \tag{6.6}$$

We already proved that $(\mathcal{I}_{\gamma_n}^{-1})^* g_{\gamma_n} \overset{*}{\rightharpoonup} (\mathcal{I}_\gamma^{-1})^* g_\gamma$. Lemma 6.1 implies $\mathcal{I}_{\gamma_n} \tau_{\gamma_n} \to \mathcal{I}_\gamma \tau_\gamma$ and since $\tau_\gamma \hat{\mathbf{v}}_n = \mathbf{0}$ for all n also (6.6) tends to zero. □

As a consequence of this theorem and the boundedness of $\mathcal{S}$ in $H^3(I)$ we now obtain the following.

COROLLARY 6.3. *The domain optimization problem has at least one solution $\gamma \in \mathcal{S}$.*

Proof. This result can be proved by choosing a minimizing sequence and using the compact embedding of $H^3(I)$ in $C^2(\bar I)$. □

7. Fréchet differentiability and derivative formula. To show differentiability we use the solution of the adjoint system of the domain optimization problem (4.1). We introduce a Lagrangian with two multipliers $\lambda_\gamma, \mu_\gamma$ for the constraints of the momentum and continuity equation, respectively. Then we compute the necessary optimality conditions for a saddle point of this Lagrangian which form the adjoint equations. Since the Stokes equations are linear, the adjoint problem is a Stokes system with a different inhomogeneity and homogeneous boundary conditions: Find $(\lambda_\gamma, \mu_\gamma) \in H_0^1(\Omega_\gamma)^2 \times L_0^2(\Omega_\gamma)$ such that

$$\begin{array}{rcll} \nu(\nabla\lambda_\gamma, \nabla\mathbf{v})_{\Omega_\gamma} - (\mu_\gamma, \operatorname{div}\mathbf{v})_{\Omega_\gamma} & = & -D_u\mathcal{J}(\gamma)\mathbf{v} & \text{for all} \quad \mathbf{v} \in H_0^1(\Omega_\gamma)^2, \\ (\operatorname{div}\lambda_\gamma, q)_{\Omega_\gamma} & = & 0 & \text{for all} \quad q \in L_0^2(\Omega_\gamma), \end{array} \tag{7.1}$$

where $\mathbf{u}_\gamma$ is the velocity component of a solution to (3.1), and $D_u\mathcal{J}(\gamma)\mathbf{v}$ denotes the Fréchet derivative of $\mathcal{J}$ with respect to $\mathbf{u}$ in direction $\mathbf{v}$. The equivalent fictitious domain formulation is the following: Find $(\hat\lambda_\gamma, \hat\mu_\gamma, \chi_\gamma) \in H_0^1(\hat\Omega)^2 \times L_0^2(\hat\Omega) \times H_\gamma^*$ such that

$$\begin{array}{rcll} \nu(\nabla\hat\lambda_\gamma, \nabla\hat{\mathbf{v}})_{\hat\Omega} - (\hat\mu_\gamma, \operatorname{div}\hat{\mathbf{v}})_{\hat\Omega} - \langle \chi_\gamma, \tau_\gamma\hat{\mathbf{v}} \rangle_{H_\gamma^*, H_\gamma} & = & -D_u\mathcal{J}(\gamma)\hat{\mathbf{v}} & \text{for all } \hat{\mathbf{v}} \in H_0^1(\hat\Omega)^2, \\ (\operatorname{div}\hat\lambda_\gamma, \hat q)_{\hat\Omega} & = & 0 & \text{for all } \hat q \in L_0^2(\hat\Omega), \\ \tau_\gamma\hat\lambda_\gamma & = & \mathbf{0}. & \end{array} \tag{7.2}$$

For the solution of (7.2) we get similar results of existence, uniqueness, regularity, and equivalence to (7.1) as for the Stokes equations in Theorems 3.1 and 5.1; see below.

THEOREM 7.1. *Let $\gamma \in \mathcal{S}$ and $(\mathbf{u}_\gamma, p_\gamma)$ be the solution of* (3.1). *Then problem* (7.1) *has a unique solution $(\lambda_\gamma, \mu_\gamma) \in [H^2(\Omega_\gamma)^2 \cap H^1_0(\Omega_\gamma)^2] \times [H^1(\Omega_\gamma) \cap L^2_0(\Omega_\gamma)]$. The regularity is uniform in γ.*

Moreover, $(\hat{\lambda}_\gamma, \hat{\mu}_\gamma, \chi_\gamma) \in H^1(\hat{\Omega})^2 \times L^2_0(\hat{\Omega}) \times H^*_\gamma$ is a solution of (7.2) *if and only if*

- $(\lambda_\gamma, \mu_\gamma) := (\hat{\lambda}_\gamma, \hat{\mu}_\gamma)|_{\Omega_\gamma}$ *is a solution of* (7.1),
- $(\hat{\lambda}_\gamma, \hat{\mu}_\gamma)|_{\Omega^c_\gamma} = (\mathbf{0}, 0)$,
- $\langle \chi_\gamma, \mathbf{h} \rangle_{H^*_\gamma, H_\gamma} = (\nu \frac{\partial \lambda_\gamma}{\partial \mathbf{n}_\gamma} - \mu_\gamma \mathbf{n}_\gamma, \mathbf{h})_{\Gamma_\gamma}$ *for all* $\mathbf{h} \in H_\gamma$

and we have

$$\chi_\gamma = \left(\nu \frac{\partial \lambda_\gamma}{\partial \mathbf{n}_\gamma} - \mu_\gamma \mathbf{n}_\gamma \right) \Big|_{\Gamma_\gamma} \qquad \text{in } L^2(\Gamma_\gamma)^2.$$

We now turn to the differentiability of $\mathcal{J}$ with respect to variations in γ. Since $\mathcal{S}$ is closed we consider $\gamma \in \operatorname{int} \mathcal{S}$, the interior of $\mathcal{S}$, and define the set of admissible directions as

$$\mathcal{S}' := \left\{ \bar{\gamma} \in H^3(I) : \bar{\gamma}|_{[0,\delta] \cup [1-\delta, 1]} = 0 \right\}.$$

For every $\gamma \in \operatorname{int} \mathcal{S}$ and $\bar{\gamma} \in \mathcal{S}'$ there exists $t_0 > 0$ such that $\gamma + t\bar{\gamma} \in \operatorname{int} \mathcal{S}$ for all $t \in [0, t_0)$. Thus we can properly define a directional derivative. We now define $I_+ := \{x \in I : \bar{\gamma}(x) \geq 0\}$, $I_- := \{x \in I : \bar{\gamma}(x) < 0\}$, and present three results that will be used to prove the differentiability of $\mathcal{J}$. Their proofs are given in section 9.

LEMMA 7.2. *Let $\mathbf{f} \in L^\infty(\hat{\Omega})^2, \mathbf{f}_\gamma := \mathbf{f}|_{\Omega_\gamma}, \mathbf{f}_{\gamma+t\bar{\gamma}} := \mathbf{f}|_{\Omega_{\gamma+t\bar{\gamma}}}$, and $\hat{\lambda}_\gamma$ be a solution of problem* (7.2). *Then*

$$\lim_{t \to 0} \frac{1}{t} (\tilde{\mathbf{f}}_\gamma - \tilde{\mathbf{f}}_{\gamma+t\bar{\gamma}}, \hat{\lambda}_\gamma)_{\hat{\Omega}} = 0.$$

LEMMA 7.3. *Let χ_γ be the third component of a solution of* (7.2) *and $\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}$ the first component of the solution of* (5.2) *for $\gamma + t\bar{\gamma}$. Then*

$$\lim_{t \to 0} \frac{1}{t} \langle \chi_\gamma, \tau_\gamma \hat{\mathbf{u}}_{\gamma+t\bar{\gamma}} \rangle_{H^*_\gamma, H_\gamma} = - \int_{I_+} \chi_\gamma(x, \gamma) \cdot \mathbf{u}_{\gamma,y}(x, \gamma) \bar{\gamma} \sqrt{1 + \gamma'^2} \, dx.$$

The integral on the right exists because $\bar{\gamma} \in \mathcal{S}' \subset L^\infty(I)$ and $\chi_\gamma \in L^2(\Gamma_\gamma)^2$. Furthermore $\tau_\gamma \mathbf{u}_{\gamma,y} = \mathbf{u}_{\gamma,y}(x, \gamma)$ and also its restriction on the set $\Gamma^+_\gamma := \{(x, \gamma(x)) : x \in I_+\} \subset \Gamma_\gamma$ are L^2 functions since $\mathbf{u}_\gamma \in H^2(\Omega_\gamma)^2$. Finally we will use the following.

LEMMA 7.4. *Let $g_{\gamma+t\bar{\gamma}}$ be the third component of the solution of* (5.2) *for $\gamma + t\bar{\gamma}$ and $\hat{\lambda}_\gamma$ the first component of the solution of* (7.2). *Then*

$$\lim_{t \to 0} \frac{1}{t} \langle g_{\gamma+t\bar{\gamma}}, \tau_{\gamma+t\bar{\gamma}} \hat{\lambda}_\gamma \rangle_{H^*_{\gamma+t\bar{\gamma}}, H_{\gamma+t\bar{\gamma}}} = \int_{I_-} g_\gamma(x, \gamma) \cdot \lambda_{\gamma,y}(x, \gamma) \bar{\gamma} \sqrt{1 + \gamma'^2} \, dx.$$

By arguments analogous to those for Lemma 7.3 the integral on the right exists.

Now we state the main result of this paper, namely the differentiability of the cost functional with respect to γ and an explicit formula for its derivative.

THEOREM 7.5. *Let $\gamma \in \operatorname{int} \mathcal{S}, \mathbf{f} \in L^\infty(\hat{\Omega})^2$, and $\mathbf{f}_\gamma := \mathbf{f}|_{\Omega_\gamma}$. Then $\mathcal{J}$ is Fréchet differentiable with respect to γ and the derivative in γ in direction $\bar{\gamma} \in \mathcal{S}'$ satisfies*

$$(7.3)\quad D_\gamma \mathcal{J}(\gamma)\bar{\gamma} = \frac{1}{\nu} \int_I \left[g_\gamma\big(x, \gamma(x)\big) \cdot \chi_\gamma\big(x, \gamma(x)\big) - p_\gamma\big(x, \gamma(x)\big) \mu_\gamma\big(x, \gamma(x)\big) \right] \bar{\gamma}(x) \, dx.$$

Proof. A simple computation leads to the identity

$$\begin{aligned}\frac{1}{t}\mathcal{J}(\gamma+t\bar{\gamma})-\mathcal{J}(\gamma) &= \frac{1}{2t}\|\mathcal{A}(\hat{\mathbf{u}}_{\gamma+t\gamma}-\hat{\mathbf{u}}_{\gamma})\|^2_{L^2(\Omega_C)^k}+\frac{1}{t}\big(\mathcal{A}(\hat{\mathbf{u}}_{\gamma+t\gamma}-\hat{\mathbf{u}}_{\gamma}),\mathcal{A}\hat{\mathbf{u}}_{\gamma}-\mathbf{u}_d\big)_{\Omega_C}\\ &= \frac{1}{2t}\|\mathcal{A}(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma})\|^2_{L^2(\Omega_C)^k}+\frac{1}{t}D_u\mathcal{J}(\gamma)(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma}).\end{aligned}$$

The first term on the right can be estimated using the boundedness of $\mathcal{A}$ and the Lipschitz continuity of the velocities proved in Theorem 6.2:

$$\frac{1}{2t}\|\mathcal{A}(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma})\|^2_{L^2(\Omega_C)^k}\le\frac{C}{t}\|\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma}\|^2_{H^1(\hat{\Omega})^2}\ \le\ Ct\|\bar{\gamma}\|^2_{L^\infty(I)},$$

where C is independent of $\gamma,\bar{\gamma}$, and t. Since $\bar{\gamma}\in\mathcal{S}'\subset W^{1,\infty}(I)$, this term tends to zero for $t\to 0$ and we obtain

$$D_\gamma\mathcal{J}(\gamma)\bar{\gamma}=\lim_{t\to 0}\frac{1}{t}D_u\mathcal{J}(\gamma)(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma}).$$

We show that this limit equals the right-hand side of (7.3). The first equation of (7.2) with $\hat{\mathbf{v}}=\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma}\in H^1_0(\hat{\Omega})^2$ as test function gives

$$\begin{aligned}D_u\mathcal{J}(\gamma)(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma}) = -\nu\big(\nabla\hat{\lambda}_\gamma,\nabla(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma})\big)_{\hat{\Omega}}+\big(\hat{\mu}_\gamma,\operatorname{div}(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma})\big)_{\hat{\Omega}}\\ +\langle\chi_\gamma,\tau_\gamma(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma})\rangle_{H^*_\gamma,H_\gamma}.\end{aligned}$$

As solutions to (5.2) for γ and $\gamma+t\bar{\gamma}$, respectively, the functions $\hat{\mathbf{u}}_\gamma$ and $\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}$ are weakly divergence free. Since $\hat{\mu}_\gamma\in L^2_0(\hat{\Omega})$ the second term on the right vanishes. With $\tau_\gamma\hat{\mathbf{u}}_\gamma=\mathbf{0}$ we obtain

$$D_u\mathcal{J}(\gamma)(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma})=\nu(\nabla\hat{\lambda}_\gamma,\nabla\hat{\mathbf{u}}_\gamma)_{\hat{\Omega}}-\nu(\nabla\hat{\lambda}_\gamma,\nabla\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}})_{\hat{\Omega}}+\langle\chi_\gamma,\tau_\gamma\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}\rangle_{H^*_\gamma,H_\gamma}.$$

For the first two terms on the right we use again the first equation in (5.2) for γ and $\gamma+t\bar{\gamma}$, respectively, with $\hat{\mathbf{v}}=\hat{\lambda}_\gamma\in H^1_0(\hat{\Omega})^2$. We get

$$\begin{aligned}D_u\mathcal{J}(\gamma)(\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}-\hat{\mathbf{u}}_{\gamma}) = (\tilde{\mathbf{f}}_\gamma-\tilde{\mathbf{f}}_{\gamma+t\bar{\gamma}},\hat{\lambda}_\gamma)_{\hat{\Omega}}+(\hat{p}_\gamma-\hat{p}_{\gamma+t\bar{\gamma}},\operatorname{div}\hat{\lambda}_\gamma)_{\hat{\Omega}}+\langle g_\gamma,\tau_\gamma\hat{\lambda}_\gamma\rangle_{H^*_\gamma,H_\gamma}\\ -\langle g_{\gamma+t\bar{\gamma}},\tau_{\gamma+t\bar{\gamma}}\hat{\lambda}_\gamma\rangle_{H^*_{\gamma+t\bar{\gamma}},H_{\gamma+t\bar{\gamma}}}+\langle\chi_\gamma,\tau_\gamma\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}\rangle_{H^*_\gamma,H_\gamma}.\end{aligned}$$

The second and third terms on the right both vanish since $\hat{\lambda}_\gamma$ as solution of (7.2) is weakly divergence free, $\hat{p}_\gamma-\hat{p}_{\gamma+t\bar{\gamma}}\in L^2_0(\hat{\Omega})$, and $\tau_\gamma\hat{\lambda}_\gamma=\mathbf{0}$. Thus we obtain

$$\begin{aligned}D_\gamma\mathcal{J}(\gamma)\bar{\gamma}=\lim_{t\to 0}\frac{1}{t}\Big[(\tilde{\mathbf{f}}_\gamma-\tilde{\mathbf{f}}_{\gamma+t\bar{\gamma}},\hat{\lambda}_\gamma)_{\hat{\Omega}}+\langle\chi_\gamma,\tau_\gamma\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}\rangle_{H^*_\gamma,H_\gamma}\\ -\langle g_{\gamma+t\bar{\gamma}},\tau_{\gamma+t\bar{\gamma}}\hat{\lambda}_\gamma\rangle_{H^*_{\gamma+t\bar{\gamma}},H_{\gamma+t\bar{\gamma}}}\Big].\end{aligned}$$

Using Lemmas 7.2 to 7.4, this implies

$$D_\gamma\mathcal{J}(\gamma)\bar{\gamma}=-\int_{I_+}\chi_\gamma(x,\gamma)\cdot\mathbf{u}_{\gamma,y}(x,\gamma)\bar{\gamma}\sqrt{1+\gamma'^2}dx-\int_{I_-}g_\gamma(x,\gamma)\cdot\lambda_{\gamma,y}(x,\gamma)\bar{\gamma}\sqrt{1+\gamma'^2}dx.$$

Because of $\tau_\gamma\mathbf{u}_\gamma=\tau_\gamma\lambda_\gamma=\mathbf{0}$ the partial derivatives times the arc length can be expressed as normal derivatives along Γ_γ. This leads to

$$D_\gamma\mathcal{J}(\gamma)\bar{\gamma}=\int_{I_+}\chi_\gamma(x,\gamma)\cdot\frac{\partial\mathbf{u}_\gamma(x,\gamma)}{\partial\mathbf{n}_\gamma}\,\bar{\gamma}\,dx+\int_{I_-}g_\gamma(x,\gamma)\cdot\frac{\partial\lambda_\gamma(x,\gamma)}{\partial\mathbf{n}_\gamma}\,\bar{\gamma}\,dx.$$

By the representations of g_γ and χ_γ we obtain

$$g_\gamma \cdot \frac{\partial \lambda_\gamma}{\partial \mathbf{n}_\gamma} = \frac{1}{\nu}(g_\gamma \cdot \chi_\gamma - p_\gamma \mu_\gamma) + \mu_\gamma \frac{\partial \mathbf{u}_\gamma}{\partial \mathbf{n}_\gamma} \cdot \mathbf{n}_\gamma \ = \ \frac{1}{\nu}(g_\gamma \cdot \chi_\gamma - p_\gamma \mu_\gamma).$$

The term with the normal derivative vanishes because of $\frac{\partial \hat{\mathbf{u}}_\gamma}{\partial \mathbf{n}_\gamma} \cdot \mathbf{n}_\gamma = \operatorname{div} \hat{\mathbf{u}}_\gamma$ on Γ_γ and since $\hat{\mathbf{u}}_\gamma$ has zero divergence in $\hat{\Omega}$ and thus on Γ_γ. Analogously we get

$$\chi_\gamma \cdot \frac{\partial \mathbf{u}_\gamma}{\partial \mathbf{n}_\gamma} = \frac{1}{\nu}(g_\gamma \cdot \chi_\gamma - p_\gamma \mu_\gamma).$$

Summing up we have shown that $\mathcal{J}$ has a directional derivative for every $\gamma \in \operatorname{int} \mathcal{S}$ in every direction $\bar{\gamma} \in \mathcal{S}'$ given by (7.3).

To show that the linear mapping $D_\gamma \mathcal{J}(\gamma) : \mathcal{S}' \to \mathbb{R}, \bar{\gamma} \mapsto D_\gamma \mathcal{J}(\gamma)\bar{\gamma}$ is a Gâteaux derivative we show that it is bounded on $\mathcal{S}'$. At first we have

$$(7.4)\ |D_\gamma \mathcal{J}(\gamma)\bar{\gamma}| \le \frac{1}{\nu}\left(\int_I |g_\gamma(x,\gamma) \cdot \chi_\gamma(x,\gamma)|\, dx + \int_I |p_\gamma(x,\gamma)\mu_\gamma(x,\gamma)|\, dx\right) \|\bar{\gamma}\|_{L^\infty(I)}.$$

Since $g_\gamma, \chi_\gamma \in L^2(\Gamma_\gamma)^2$ we obtain $\mathcal{I}_\gamma g_\gamma, \mathcal{I}_\gamma \chi_\gamma \in L^2(I)^2$. The first integral in brackets can be estimated by Hölder's inequality as

$$\int_I |g_\gamma(x,\gamma) \cdot \chi_\gamma(x,\gamma)|\, dx \le \|\mathcal{I}_\gamma g_\gamma\|_{L^2(I)} \|\mathcal{I}_\gamma \chi_\gamma\|_{L^2(I)}.$$

Since γ is fixed the terms on the right are constants with respect to $\bar{\gamma}$. Concerning the second integral in (7.4) we have $\tau_\gamma p_\gamma, \tau_\gamma \mu_\gamma \in L^2(\Gamma_\gamma)$ because of $p_\gamma, \mu_\gamma \in H^1(\Omega_\gamma)$. Thus it can be estimated analogously by a constant which is independent of $\bar{\gamma}$:

$$\int_I |p_\gamma(x,\gamma)\mu_\gamma(x,\gamma)| dx \le \int_I |p_\gamma(x,\gamma)\mu_\gamma(x,\gamma)| \sqrt{1+\gamma'^2} dx \le \|\tau_\gamma p_\gamma\|_{L^2(\Gamma_\gamma)} \|\tau_\gamma \mu_\gamma\|_{L^2(\Gamma_\gamma)}.$$

To prove that $D_\gamma \mathcal{J}(\gamma)$ is a Fréchet derivative we show that the mapping $\gamma \mapsto D_\gamma \mathcal{J}(\gamma)$ is continuous from $\operatorname{int} \mathcal{S}$ to $\mathcal{L}(\mathcal{S}', \mathbb{R})$. We write (7.3) as

$$D_\gamma \mathcal{J}(\gamma)\bar{\gamma} = \frac{1}{\nu} \int_I [(\mathcal{I}_\gamma g_\gamma)(x) \cdot (\mathcal{I}_\gamma \chi_\gamma)(x) - (\mathcal{I}_\gamma \tau_\gamma p_\gamma)(x)(\mathcal{I}_\gamma \tau_\gamma \mu_\gamma)(x)]\bar{\gamma}\, dx.$$

For all $\gamma \in \mathcal{S}$ the functional $g_\gamma \in H_\gamma^*$ can be extended onto $L^2(\Gamma_\gamma)^2$; see Theorem 5.1. Therefore $(\mathcal{I}_\gamma^{-1})^* g_\gamma \in H_I^*$ can be extended onto $L^2(I)^2$ by the definition

$$\begin{aligned} \langle (\mathcal{I}_\gamma^{-1})^* g_\gamma, \Psi \rangle_{(L^2(I)^2)^*, L^2(I)^2} &:= \langle g_\gamma, \mathcal{I}_\gamma^{-1}\Psi \rangle_{(L^2(\Gamma_\gamma)^2)^*, L^2(\Gamma_\gamma)^2} \\ &= (g_\gamma, \mathcal{I}_\gamma^{-1}\Psi)_{\Gamma_\gamma} \ = \ ((\mathcal{I}_\gamma^{-1})^\star g_\gamma, \Psi)_I, \qquad \Psi \in L^2(I)^2, \end{aligned}$$

where we introduced the Hilbert space adjoint of $\mathcal{I}_\gamma^{-1}$:

$$(7.5) \qquad (\mathcal{I}_\gamma^{-1})^\star :\ L^2(\Gamma_\gamma)^2 \to L^2(I)^2, \qquad g \mapsto g(\cdot, \gamma(\cdot))\sqrt{1+\gamma'(\cdot)^2}.$$

Since $\mathcal{I}_\gamma$ is an isomorphism between $L^2(I)^2$ and $L^2(\Gamma_\gamma)^2$ and $\mathcal{S}$ is bounded in $H^3(I)$ which is continuously embedded in $W^{1,\infty}(I)$ also $\{\|(\mathcal{I}_\gamma^{-1})^\star\|_{\mathcal{L}(L^2(\Gamma_\gamma)^2, L^2(I)^2)}\}_{\gamma \in \mathcal{S}}$ is bounded. This implies that the family $\{(\mathcal{I}_\gamma^{-1})^\star g_\gamma\}_{\gamma \in \mathcal{S}}$ is uniformly bounded in $L^2(I)^2$. Hence for $\gamma_n \to \gamma$ in $\mathcal{S}$ there exists a weakly convergent subsequence, i.e.,

$$((\mathcal{I}_{\gamma_n}^{-1})^\star g_{\gamma_n}, \Psi)_I \to (g, \Psi)_I \qquad \text{for all } \Psi \in L^2(I)^2$$

with some $g \in L^2(I)^2$. On the other hand, by Theorem 6.2 $(\mathcal{I}_{\gamma_n}^{-1})^* g_{\gamma_n}$ converges weak-$*$ in H_I^* to $(\mathcal{I}_\gamma^{-1})^* g_\gamma$ which means that for $n \to \infty$ and all $\Psi \in H_I$ we have

$$\langle (\mathcal{I}_{\gamma_n})^* g_{\gamma_n}, \Psi\rangle_{H_I^*,H_I} \;=\; ((\mathcal{I}_{\gamma_n}^{-1})^\star g_{\gamma_n}, \Psi)_I \to \langle (\mathcal{I}_\gamma^{-1})^* g_\gamma, \Psi\rangle_{H_I^*,H_I} \;=\; ((\mathcal{I}_\gamma^{-1})^\star g_\gamma, \Psi)_I.$$

Because H_I is dense in $L^2(I)^2$ we have $g = \left(\mathcal{I}_\gamma^{-1}\right)^\star g_\gamma$. By definition of $\left(\mathcal{I}_\gamma^{-1}\right)^\star$ this implies $\mathcal{I}_{\gamma_n} g_{\gamma_n} \to \mathcal{I}_\gamma g_\gamma$ in $L^2(I)^2$. The same arguments hold for χ_γ. Since $\mathcal{I}_\gamma$ is continuous we have shown that the mappings $\gamma \mapsto \mathcal{I}_\gamma g_\gamma$ and $\gamma \mapsto \mathcal{I}_\gamma \chi_\gamma$ are continuous from $\mathcal{S}$ to $L^2(I)^2$ which implies that $\gamma \mapsto (\mathcal{I}_\gamma g_\gamma)\cdot(\mathcal{I}_\gamma \chi_\gamma)$ is continuous from $\mathcal{S}$ to $L^1(I)$.

To show that the mapping $\gamma \mapsto \mathcal{I}_\gamma \tau_\gamma p_\gamma$ is continuous we recall that the family $\{\|p_\gamma\|_{H^1(\Omega_\gamma)}\}_{\gamma\in\mathcal{S}}$ is bounded. Since the solution $\hat{p}_\gamma$ of (5.2) is not even necessarily in $H^1(\hat{\Omega})$ we extend each p_γ to some $\bar{p}_\gamma \in H^1(\hat{\Omega})$ such that the family $\{\bar{p}_\gamma\}_{\gamma\in\mathcal{S}}$ is bounded in $H^1(\hat{\Omega})$: We can easily define a family of bijective transformations $T_\gamma : \Omega_\gamma \to \hat{\Omega}$ which together with their inverse mappings are uniformly Lipschitz continuous in $\mathcal{S}$. The family $\{T_\gamma p_\gamma\}_{\gamma\in\mathcal{S}}$ of transformed functions is uniformly bounded in $H^1(\hat{\Omega})$; see [13, Lem. II.3.2] and the proofs of this lemma and of [13, Lem. II.3.1]. Extending $T_\gamma p_\gamma$ by reflection to a function $\check{p}_\gamma$ defined on $\check{\Omega} := (0,1)\times(-1,1)$ the family $\{\check{p}_\gamma\}_{\gamma\in\mathcal{S}}$ is uniformly bounded in $H^1(\check{\Omega})$; see [14, Lem. IX.2]. Using [13, Lem. II.3.2] again we obtain $T_\gamma^{-1}\check{p}_\gamma \in H^1(T_\gamma^{-1}(\check{\Omega}))$ and the uniform boundedness of $\{\bar{p}_\gamma\}_{\gamma\in\mathcal{S}}$ in $H^1(\hat{\Omega})$ for $\bar{p}_\gamma := (T_\gamma^{-1}\check{p}_\gamma)|_{\hat{\Omega}}$. For every sequence $\gamma_n \to \gamma$ in $\mathcal{S} \subset W^{1,\infty}(I)$ we thus have for a subsequence $\bar{p}_{\gamma_n} \rightharpoonup \bar{p}$ weakly in $H^1(\hat{\Omega})$ and $\bar{p}_{\gamma_n} \to \bar{p}$ strongly in $L^2(\hat{\Omega})$. Since $\hat{p}_{\gamma_n} \to \hat{p}_\gamma$ strongly in $L^2(\hat{\Omega})$ by Theorem 6.2 and $\hat{p}_\gamma|_{\Omega_\gamma} = p_\gamma = \bar{p}_\gamma|_{\Omega_\gamma}$ this implies $\bar{p}|_{\Omega_\gamma} = p_\gamma$ and by construction $\bar{p} = \bar{p}_\gamma$ in $\hat{\Omega}$. Therefore we have $\bar{p}_{\gamma_n} \rightharpoonup \bar{p}_\gamma$ weakly in $H^1(\hat{\Omega})$ for the whole sequence. Because of $\tau_\gamma \bar{p}_\gamma = \tau_\gamma p_\gamma$ for all $\gamma \in \mathcal{S}$ we now obtain

$$\begin{aligned} &\|\mathcal{I}_{\gamma_n}\tau_{\gamma_n}p_{\gamma_n} - \mathcal{I}_\gamma\tau_\gamma p_\gamma\|_{L^2(I)} = \|\mathcal{I}_{\gamma_n}\tau_{\gamma_n}\bar{p}_{\gamma_n} - \mathcal{I}_\gamma\tau_\gamma\bar{p}_\gamma\|_{L^2(I)} \\ &\le \|(\mathcal{I}_{\gamma_n}\tau_{\gamma_n} - \mathcal{I}_\gamma\tau_\gamma)\bar{p}_{\gamma_n}\|_{L^2(I)} + \|\mathcal{I}_\gamma\|_{\mathcal{L}(L^2(\Gamma_\gamma),L^2(I))}\|\tau_\gamma(\bar{p}_{\gamma_n} - \bar{p}_\gamma)\|_{L^2(\Gamma_\gamma)}. \end{aligned}$$

The first term tends to zero because of strong convergence of the transformed trace operators $\mathcal{I}_{\gamma_n}\tau_{\gamma_n}$; see [15, Lem. 2.11]. In the second term the first factor is uniformly bounded for all $\gamma \in \mathcal{S}$, and the second one tends to zero due to the compact embedding of $H^1(\hat{\Omega})$ into $L^2(\Gamma_\gamma)$. Thus we have shown that $\gamma \mapsto \mathcal{I}_\gamma\tau_\gamma p_\gamma$ is continuous from $\mathcal{S}$ to $L^2(I)$. Using an analogous argument for μ_γ we obtain that $\gamma \mapsto (\mathcal{I}_\gamma\tau_\gamma p_\gamma)(\mathcal{I}_\gamma\tau_\gamma\mu_\gamma)$ is continuous from $\mathcal{S}$ to $L^1(I)$. This implies the Fréchet differentiability of $\mathcal{J}$ with respect to γ. □

Let us note here that this derivative formula does not include normal derivatives of the state variables along the domain boundary, but only the Lagrange multipliers introduced by the embedding domain method.

8. Numerical methods. The numerical example presented below was computed using the formula (7.3) for the derivative and an SQP method (see [16]) for the optimization. It requires one gradient and at least one cost functional evaluation in each iteration. Thus the systems (5.2) and (7.2) have to be solved several times for different γ. Both systems were discretized by stabilized linear finite elements (see [17]) for velocity and pressure and piecewise constant elements for g_γ and χ_γ. To satisfy the inf-sup condition for the latter we used a coarser mesh size for their discretization as suggested in [18]. The discretized counterparts of (5.2) and (7.2) then read

$$(8.1) \qquad \begin{pmatrix} A & B^T & D_\gamma^T \\ B & C & 0 \\ D_\gamma & 0 & 0 \end{pmatrix} \begin{pmatrix} U_\gamma \\ P_\gamma \\ G_\gamma \end{pmatrix} = \begin{pmatrix} F_\gamma \\ H \\ 0 \end{pmatrix},$$

where the matrix C and the vector H appear due to the stabilization. Only the entities with subscript γ in (8.1) change during the optimization process. Note that D_γ represents a discrete one-dimensional trace operator and thus is very sparse. Therefore the effort of rebuilding the system when γ changes is negligible. This is a typical advantage of embedding domain techniques.

For the solution of the discrete systems we used an Uzawa algorithm with an outer preconditioned conjugate gradient iteration on the pair (P_γ, G_γ). For the inner system (a discrete Laplacian) we exploited the fact that the matrix A is fixed for all γ. Thus we computed its Cholesky factorization only once using a reverse Cuthill–McKee re-ordering algorithm to obtain a minimal fill-up by the factorization. For every gradient evaluation the inner system of the Uzawa algorithm thus requires only solution of two sparse triangular systems. Once the discrete counterparts of $p_\gamma, g_\gamma, \mu_\gamma, \chi_\gamma$ are computed the evaluation of the gradient via (7.3) can be done fast since it involves only the computation of a one-dimensional integral. Furthermore, this can be evaluated *exactly* since for the used basis functions simple integration rules are exact. Thus no additional discretization error is introduced by the gradient evaluation and no normal derivatives of the velocities have to be computed. Their values are implicitly included in g_γ, χ_γ. This fast and stable algorithm is fundamentally based on the embedding domain technique by which these two Lagrange multipliers were introduced.

9. Numerical examples. We considered a driven cavity flow as test example to show feasibility of the embedding domain technique and the derivative formula (7.3). The effort into the efficiency of the optimization was restricted to the choice of the regularization type and parameter. The computational domain is a square with edge length one. On the top edge a constant positive horizontal velocity is prescribed. The other edges are regarded as walls with homogeneous Dirichlet boundary conditions. A part of the bottom wall was set to be variable. We considered a tracking type cost functional with $\mathcal{A}$ being the identity. The parameter ν was set to $\frac{1}{100}$. We chose the observation region as $\Omega_C := [0,1] \times [0.5,1]$ and the desired state $\mathbf{u}_d := \mathbf{u}_{\gamma_d}$. The curve Γ_{γ_d} for the desired state was the graph of a symmetric cubic spline function γ_d between $y = \gamma_d(x) = 0.0$ near the lateral walls and $y = \gamma_d(0.5) = 0.4$. The function γ had 13 degrees of freedom. The starting curve was a straight line at the level $y = 0.2$. We used the box constraints $\gamma \in [0.1, 0.5]$ and added a regularization term $\varepsilon\|\gamma'\|^2_{L^2(I)}$ with $\varepsilon = 10^{-5}$ to the cost functional $\mathcal{J}(\gamma)$.

As can be seen in Table 9.1 the convergence was very fast: After seven iterations and nine function evaluations the cost functional was reduced by more than 99 percent. After another iteration the error, e.g., at the point $\gamma(0.5)$, is only 1 percent. The difference at the other points is higher. The influence of the point $\gamma(0.5)$ is most important because the flow structure is mostly influenced by the height of the bottom wall and less by its form. Streamlines and pressure distributions for desired state, start configuration, and solution are depicted in Figure 9.1.

A comparison between the derivative computed by formula (7.3) and by finite difference derivatives can be seen in Table 9.2. Here only one point (in the middle) of γ was variable, the others were fixed at $\gamma = \gamma_d$ with γ_d as above. A difference in the magnitude of the derivative computed by (7.3) and the finite difference derivatives can be seen. The finite difference derivatives differ much depending on their step-size. The derivative computed by (7.3) represents the behavior of the cost functional on the whole considered interval, specifically near the optimum $\gamma(0.5) = \gamma_d(0.5) = 0.4$.

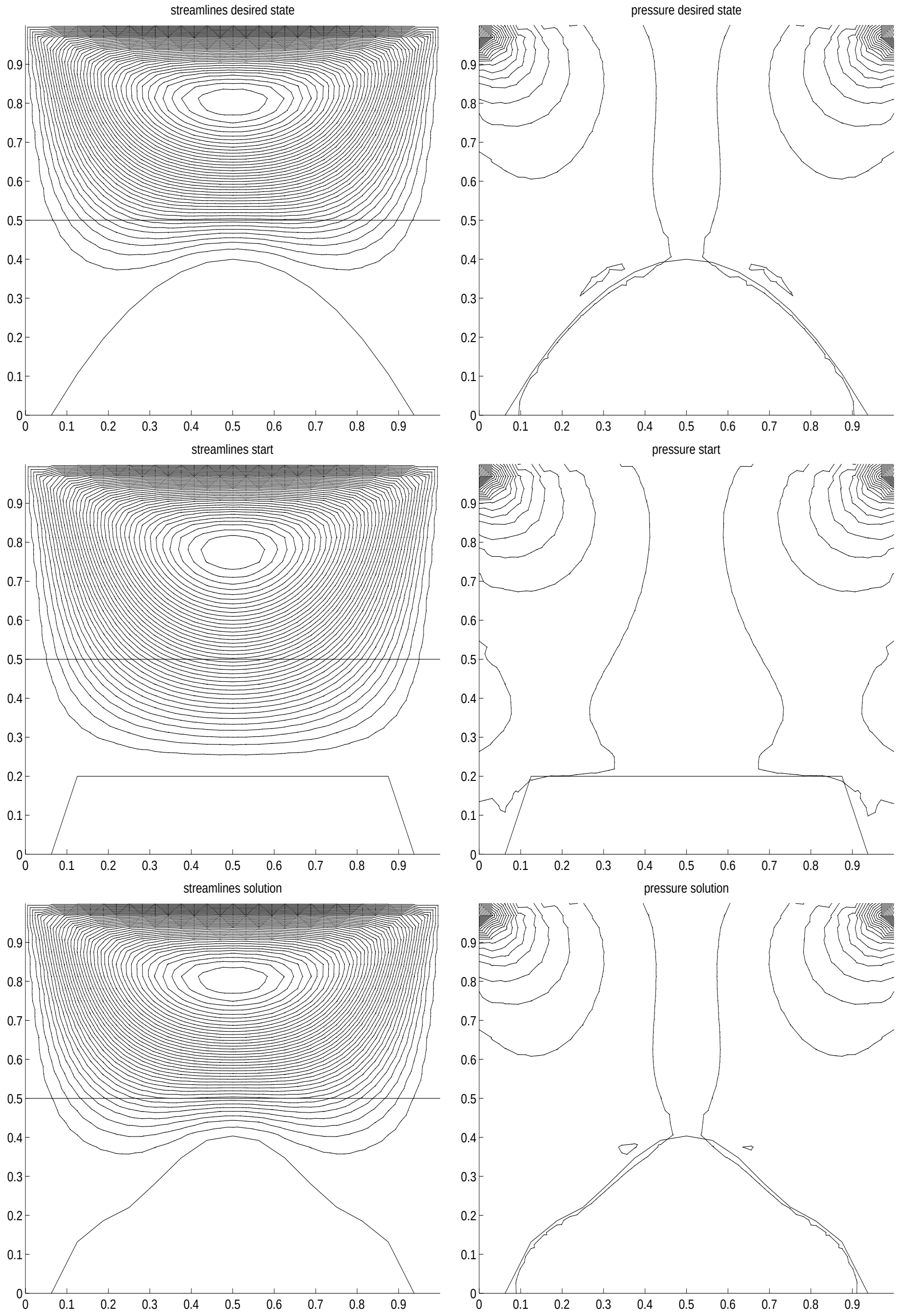


FIG. 9.1. *Streamlines and pressure distribution for the driven cavity test case with* $Re = 100$. *Top: desired state (cubic spline interpolated curve), middle: start curve, bottom: solution. The observation region* Ω_C *is the upper half of the cavity above the straight line.*

Table 9.1

Convergence behavior for the driven cavity test example. Only 7 *of the* 13 *points of* γ *are listed. It./Ev. = number of iterations/function evaluations.*

It.	Ev.	$\mathcal{J}$	$\gamma(0.125)$	$\gamma(0.25)$	$\gamma(0.375)$	$\gamma(0.5)$	$\gamma(0.625)$	$\gamma(0.75)$	$\gamma(0.875)$
	1	6.2999e-04	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000
1	2	6.2079e-04	0.1992	0.2003	0.2016	0.2023	0.2016	0.2003	0.1992
2	3	5.2649e-04	0.1000	0.2300	0.4094	0.5000	0.4097	0.2316	0.1000
3	4	4.831e-04	0.1842	0.2042	0.2311	0.2438	0.2312	0.2044	0.1843
4	5	3.2227e-04	0.1688	0.2084	0.2625	0.2875	0.2627	0.2089	0.1688
5	7	5.2006e-06	0.1344	0.2190	0.3399	0.3937	0.3408	0.2201	0.1344
6	9	2.7258e-06	0.1299	0.2204	0.3499	0.4079	0.3509	0.2216	0.1300
7	10	2.5865e-06	0.1312	0.2200	0.3473	0.4044	0.3482	0.2212	0.1312
8	25	2.5865e-06	0.1312	0.2200	0.3473	0.4044	0.3482	0.2212	0.1312
9	27	2.5722e-06	0.1316	0.2200	0.3466	0.4035	0.3475	0.2211	0.1316
10	42	2.5722e-06	0.1316	0.2200	0.3466	0.4035	0.3475	0.2211	0.1316
11	43	2.5665e-06	0.1317	0.2200	0.3466	0.4035	0.3475	0.2212	0.1318
γ_d			0.1061	0.2694	0.3673	0.4000	0.3673	0.2694	0.1061

Table 9.2

Comparison between derivatives computed by formula (7.3) *and by finite differences with different step-sizes* h. *The only control parameter here was* $\gamma(x = 0.5)$, *the other points of the curve were fixed at* $\gamma(x) = \gamma_d(x), x \neq 0.5$. *Note the changing of the sign of the analytic derivative and the finite difference derivative with small step-size near the minimum at* $\gamma(0.5) = 0.4$.

	Derivative $D_\gamma\mathcal{J}(\gamma)$ computed by					
	Formula	Finite differences with step-size				
$\gamma(0.5)$	(7.3)	$h = 0.2$	$h = 0.1$	$h = 0.05$	$h = 0.02$	$h = 0.01$
0.36	-3.6967e-04	2.9337e-03	2.8154e-04	-6.0385e-05	-9.4314e-05	-1.2357e-04
0.395	-1.2531e-04	4.3284e-03	1.4187e-03	2.0772e-04	-1.4312e-05	-2.7662e-05
0.4	2.6122e-06	4.5469e-03	1.5743e-03	3.1872e-04	4.5478e-05	-4.7811e-06
0.405	7.3862e-05	4.7846e-03	1.6900e-03	4.5924e-04	1.3941e-04	3.7376e-05
0.49	4.1378e-02	9.6293e-03	8.7152e-03	8.6892e-03	7.6160e-03	7.2823e-03

Proof of Lemma 7.2. Since $\hat{\lambda}_\gamma|_{\Omega^c_\gamma} = \mathbf{0}$ and $\tilde{\mathbf{f}}_\gamma = \tilde{\mathbf{f}}_{\gamma+t\bar{\gamma}} = \mathbf{f}$ on $\Omega_\gamma \cap \Omega_{\gamma+t\bar{\gamma}}$ the integral in $(\tilde{\mathbf{f}}_\gamma - \tilde{\mathbf{f}}_{\gamma+t\bar{\gamma}}, \hat{\lambda}_\gamma)_{\hat{\Omega}}$ has only to be taken over the set $\Omega_\gamma \setminus (\Omega_\gamma \cap \Omega_{\gamma+t\bar{\gamma}}) = I_- \times (\gamma + t\bar{\gamma}, \gamma)$, where $I_- := \{x \in I : \bar{\gamma}(x) < 0\}$. With $\tau_\gamma \lambda_\gamma = \mathbf{0}$ we obtain

$$(\tilde{\mathbf{f}}_\gamma - \tilde{\mathbf{f}}_{\gamma+t\bar{\gamma}}, \hat{\lambda}_\gamma)_{\hat{\Omega}} = \int_{I_-} \int_{\gamma+t\bar{\gamma}}^{\gamma} \mathbf{f}(x,y) \cdot [\lambda_\gamma(x,\gamma) - \lambda_\gamma(x,y)]\, dydx.$$

Estimating this term analogously to (6.2) we finally get

$$\frac{1}{t}|(\tilde{\mathbf{f}}_\gamma - \tilde{\mathbf{f}}_{\gamma+t\bar{\gamma}}, \hat{\lambda}_\gamma)_{\hat{\Omega}}| \leq \frac{1}{3}\, \|\mathbf{f}\|_{L^\infty(\hat{\Omega})^2} \|\lambda_\gamma\|_{H^1(\Omega_\gamma)^2} t^{1/2} \|\bar{\gamma}\|^{3/2}_{L^\infty(I)}.$$

Since $\mathbf{f} \in L^\infty(\hat{\Omega})^2, \lambda_\gamma \in H^1(\Omega_\gamma)^2$, and $\bar{\gamma} \in \mathcal{S}' \subset L^\infty(I)$ this tends to zero for $t \to 0$.

Proof of Lemma 7.3. Using Theorem 7.1 we obtain

$$\langle \chi_\gamma, \tau_\gamma \hat{\mathbf{u}}_{\gamma+t\bar{\gamma}} \rangle_{H^*_\gamma, H_\gamma} = \int_I \chi_\gamma(x,\gamma) \cdot \hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}(x,\gamma) \sqrt{1+\gamma'^2}\, dx.$$

By Theorem 5.1 we have $\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}|_{\Omega^c_{\gamma+t\bar{\gamma}}} = \mathbf{0}$ which implies $\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}(x,\gamma) = \mathbf{0}$ for $x \in I_-$ and $\hat{\mathbf{u}}_{\gamma+t\bar{\gamma}}(x,\gamma) = \mathbf{u}_{\gamma+t\bar{\gamma}}(x,\gamma)$ for $x \in I_+ := I \setminus I_-$. With $\tau_{\gamma+t\bar{\gamma}} \mathbf{u}_{\gamma+t\bar{\gamma}} = \mathbf{0}$ we get

$$\langle \chi_\gamma, \tau_\gamma \hat{\mathbf{u}}_{\gamma+t\bar{\gamma}} \rangle_{H^*_\gamma, H_\gamma} = -\int_{I_+} \chi_\gamma(x,\gamma) \cdot (\mathbf{u}_{\gamma+t\bar{\gamma}}(x,\gamma+t\bar{\gamma}) - \mathbf{u}_{\gamma+t\bar{\gamma}}(x,\gamma))\, \sqrt{1+\gamma'^2}\, dx$$

$$= -\int_{I_+} \chi_\gamma(x,\gamma)\cdot\left(\int_\gamma^{\gamma+t\bar\gamma} \mathbf{u}_{\gamma+t\bar\gamma,y}(x,\xi)\,d\xi\right)\sqrt{1+\gamma'^2}\,dx.$$

Note that $\mathbf{u}_{\gamma+t\bar\gamma,y} \in H^1(\Omega_{\gamma+t\bar\gamma})^2$ and thus its trace on the segment $\{(x,\xi) : \xi \in [\gamma(x), \gamma(x)+t\bar\gamma(x)]\}$ (for fixed x) is a L^2 function. With the same argument the restriction of $\tau_\gamma \mathbf{u}_{\gamma+t\bar\gamma,y}$ onto Γ_γ^+ is an L^2 function. We show that

$$C := \frac{1}{t}\langle\chi_\gamma, \tau_\gamma\hat{\mathbf{u}}_{\gamma+t\bar\gamma}\rangle_{H_\gamma^*,H_\gamma} - \int_{I_+}\chi_\gamma(x,\gamma)\cdot\mathbf{u}_{\gamma,y}(x,\gamma)\bar\gamma\sqrt{1+\gamma'^2}\,dx$$

tends to zero. Using $\frac{1}{t}\int_\gamma^{\gamma+t\bar\gamma}\mathbf{u}_{\gamma+t\bar\gamma,y}(x,\gamma)\,d\xi = \mathbf{u}_{\gamma+t\bar\gamma,y}(x,\gamma)\,\bar\gamma$ we obtain

$$\begin{aligned} C = C_1 + C_2 \;:=\; & \int_{I_+}\chi_\gamma(x,\gamma)\cdot[\mathbf{u}_{\gamma+t\bar\gamma,y}(x,\gamma)-\mathbf{u}_{\gamma,y}(x,\gamma)]\,\bar\gamma\sqrt{1+\gamma'^2}\,dx \\ & +\frac{1}{t}\int_{I_+}\chi_\gamma(x,\gamma)\cdot\left(\int_\gamma^{\gamma+t\bar\gamma}[\mathbf{u}_{\gamma+t\bar\gamma,y}(x,\xi)-\mathbf{u}_{\gamma+t\bar\gamma,y}(x,\gamma)]\,d\xi\right)\sqrt{1+\gamma'^2}\,dx. \end{aligned}$$

We show that both terms tend to zero for $t \to 0$: With Hölder's inequality we get

$$|C_1| \le \|\bar\gamma\|_{L^\infty(I)}\|\chi_\gamma\|_{L^2(\Gamma_\gamma)^2}\|\mathbf{u}_{\gamma+t\bar\gamma,y}-\mathbf{u}_{\gamma,y}\|_{L^2(\Gamma_\gamma^+)^2}.$$

The first two terms are bounded and we show that the last one tends to zero: The family $\{\|\mathbf{u}_{\gamma+t\bar\gamma}\|_{H^2(\Omega_{\gamma+t\bar\gamma})^2}\}$ and thus $\{\|\mathbf{u}_{\gamma+t\bar\gamma,y}\|_{H^1(\Omega_{\gamma+t\bar\gamma})^2}\}$ is bounded for $t \in [0,t_0]$ with some t_0 sufficiently small. For $\Omega'_\gamma := \Omega_\gamma \cap \Omega_{\gamma+t_0\bar\gamma}$ which is contained in $\Omega_{\gamma+t\bar\gamma}$ for all $t \in [0,t_0]$ the family $\{\mathbf{u}_{\gamma+t\bar\gamma,y} : t \in [0,t_0]\}$ is bounded in $H^1(\Omega'_\gamma)^2$. This implies that for every sequence $\{t_i\} \subset [0,t_0]$ with $\lim t_i = 0$ there exists a subsequence $\{t_{i'}\}$ such that $\{\mathbf{u}_{\gamma+t_{i'}\bar\gamma,y}\}$ is weakly convergent in $H^1(\Omega'_\gamma)^2$. This implies strong convergence of the sequence $\{\mathbf{u}_{\gamma+t_{i'}\bar\gamma,y}\}$ in $L^2(\Omega'_\gamma)^2$. Theorem 6.2 implies $\hat{\mathbf{u}}_{\gamma+t\bar\gamma,y} \to \hat{\mathbf{u}}_{\gamma,y}$ strongly in $L^2(\hat\Omega)^2$ for $t \to 0$ and thus the weak convergence of $\{\mathbf{u}_{\gamma+t_i\bar\gamma,y}\}$ in $H^1(\Omega'_\gamma)^2$ is valid for the whole sequence. We consider the traces of the functions of any sequence on the set $\Gamma_\gamma^+ = \Gamma_\gamma \cap \partial\Omega'_\gamma = \{(x,\gamma(x)) : x \in I_+\}$. By a classical embedding theorem the weak convergence $\mathbf{u}_{\gamma+t\bar\gamma,y} \rightharpoonup \mathbf{u}_{\gamma,y}$ in $H^1(\Omega'_\gamma)^2$ implies strong convergence $\mathbf{u}_{\gamma+t\bar\gamma,y}|_{\Gamma_\gamma^+} \to \mathbf{u}_{\gamma,y}|_{\Gamma_\gamma^+}$ in $L^2(\Gamma_\gamma^+)^2$. Thus C_1 tends to zero for $t \to 0$. Because $\mathbf{u}_{\gamma+t\bar\gamma,yy}$ is an L^2 function on $\Omega_{\gamma+t\bar\gamma}$ and thus also on $\{(x,y) : x \in I_+, y \in (\gamma(x), \gamma(x)+t\bar\gamma(x))\}$ we write

$$\begin{aligned} C_2 &= \frac{1}{t}\int_{I_+}\chi_\gamma(x,\gamma)\cdot\left(\int_\gamma^{\gamma+t\bar\gamma}\int_\gamma^\xi \mathbf{u}_{\gamma+t\bar\gamma,yy}(x,\eta)\,d\eta d\xi\right)\sqrt{1+\gamma'^2}\,dx \\ &= \frac{1}{t}\int_{I_+}\chi_\gamma(x,\gamma)\cdot\left(\int_\gamma^{\gamma+t\bar\gamma}\mathbf{u}_{\gamma+t\bar\gamma,yy}(x,\eta)(\gamma+t\bar\gamma-\eta)\,d\eta\right)\sqrt{1+\gamma'^2}\,dx. \end{aligned}$$

For a.e. $x \in I_+$ the inner integral exists and can be estimated by Hölder's inequality:

$$\begin{aligned} \left\|\int_\gamma^{\gamma+t\bar\gamma}\mathbf{u}_{\gamma+t\bar\gamma,yy}(x,\eta)(\gamma+t\bar\gamma-\eta)\,d\eta\right\|_2 &\le \max_{\eta\in(\gamma,\gamma+t\bar\gamma)}|\gamma+t\bar\gamma-\eta|\int_\gamma^{\gamma+t\bar\gamma}\|\mathbf{u}_{\gamma+t\bar\gamma,yy}(x,\eta)\|_2\,d\eta \\ &\le t|\bar\gamma|\left(\int_\gamma^{\gamma+t\bar\gamma}d\eta\right)^{1/2}\left(\int_\gamma^{\gamma+t\bar\gamma}\|\mathbf{u}_{\gamma+t\bar\gamma,yy}(x,\eta)\|_2^2\,d\eta\right)^{1/2} \\ &= (t|\bar\gamma|)^{3/2}\left(\int_\gamma^{\gamma+t\bar\gamma}\|\mathbf{u}_{\gamma+t\bar\gamma,yy}(x,\eta)\|_2^2\,d\eta\right)^{1/2}. \end{aligned}$$

Thus we obtain

$$\begin{aligned}|C_2| &\le \frac{1}{t}\int_{I_+}\|\chi_\gamma(x,\gamma)\|_2\left\|\int_\gamma^{\gamma+t\bar{\gamma}}\mathbf{u}_{\gamma+t\bar{\gamma},yy}(x,\eta)(\gamma+t\bar{\gamma}-\eta)\,d\eta\right\|_2\sqrt{1+\gamma'^2}\,dx\\
&\le t^{1/2}\int_{I_+}\|\chi_\gamma(x,\gamma)\|_2\left(\int_\gamma^{\gamma+t\bar{\gamma}}\|\mathbf{u}_{\gamma+t\bar{\gamma},yy}(x,\eta)\|_2^2\,d\eta\right)^{1/2}|\bar{\gamma}|^{3/2}\sqrt{1+\gamma'^2}\,dx\\
&\le t^{1/2}\|\bar{\gamma}\|^{3/2}_{L^\infty(I_+)}\sqrt{1+\|\gamma\|^2_{W^{1,\infty}(I_+)}}\int_{I_+}\|\chi_\gamma(x,\gamma)\|_2\left(\int_\gamma^{\gamma+t\bar{\gamma}}\|\mathbf{u}_{\gamma+t\bar{\gamma},yy}(x,\eta)\|_2^2\,d\eta\right)^{1/2}dx\\
&\le t^{1/2}\|\bar{\gamma}\|^{3/2}_{L^\infty(I)}\sqrt{1+\|\gamma\|^2_{W^{1,\infty}(I)}}\|\chi_\gamma\|_{L^2(\Gamma_\gamma)^2}\,\|\mathbf{u}_{\gamma+t\bar{\gamma}}\|_{H^2(\Omega_{\gamma+t\bar{\gamma}})^2}.\end{aligned}$$

Boundedness of $\{\|\mathbf{u}_\gamma\|_{H^2(\Omega_\gamma)^2}\}_{\gamma\in\mathcal{S}}$, $\mathcal{S}\subset W^{1,\infty}(I)$, and $\mathcal{S}'\subset L^\infty(I)$ now imply $C_2\to 0$.

Proof of Lemma 7.4. Theorem 5.1 gives

$$\langle g_{\gamma+t\bar{\gamma}},\tau_{\gamma+t\bar{\gamma}}\hat{\lambda}_\gamma\rangle_{H^*_{\gamma+t\bar{\gamma}},H_{\gamma+t\bar{\gamma}}}=\int_{I_-}g_{\gamma+t\bar{\gamma}}(x,\gamma+t\bar{\gamma})\cdot\left(\int_\gamma^{\gamma+t\bar{\gamma}}\lambda_{\gamma,y}(x,\xi)\,d\xi\right)\sqrt{1+(\gamma'+t\bar{\gamma}')^2}dx.$$

Adding zero in an appropriate way we get for the difference

$$\begin{aligned}&\frac{1}{t}\langle g_{\gamma+t\bar{\gamma}},\tau_{\gamma+t\bar{\gamma}}\hat{\lambda}_\gamma\rangle_{H^*_{\gamma+t\bar{\gamma}},H_{\gamma+t\bar{\gamma}}}-\int_{I_-}g_\gamma(x,\gamma)\cdot\lambda_{\gamma,y}(x,\gamma)\bar{\gamma}\sqrt{1+\gamma'^2}\,dx\\
&=\int_{I_-}\left(g_{\gamma+t\bar{\gamma}}(x,\gamma+t\bar{\gamma})\sqrt{1+(\gamma'+t\bar{\gamma}')^2}-g_\gamma(x,\gamma)\sqrt{1+\gamma'^2}\right)\cdot\lambda_{\gamma,y}(x,\gamma)\bar{\gamma}\,dx\\
&\quad+\int_{I_-}g_{\gamma+t\bar{\gamma}}(x,\gamma+t\bar{\gamma})\cdot\left(\int_\gamma^{\gamma+t\bar{\gamma}}[\lambda_{\gamma,y}(x,\xi)-\lambda_{\gamma,y}(x,\gamma)]\,d\xi\right)\bar{\gamma}\sqrt{1+(\gamma'+t\bar{\gamma}')^2}\,dx\\
&=:D_1+D_2.\end{aligned}$$

Again we show that both terms tend to zero for $t\to 0$: We define the function

$$\Psi(x):=\begin{cases}\lambda_{\gamma,y}\big(x,\gamma(x)\big)\bar{\gamma}(x), & x\in I_-,\\ \mathbf{0}, & x\in I_+\end{cases}$$

which is in $L^2(I)^2$ because $\lambda_\gamma\in H^2(\Omega_\gamma)^2$ and $\bar{\gamma}\in\mathcal{S}'\subset L^\infty(I)$. Then we have

$$D_1=\int_I\left(g_{\gamma+t\bar{\gamma}}(x,\gamma+t\bar{\gamma})\sqrt{1+(\gamma'+t\bar{\gamma}')^2}-g_\gamma(x,\gamma)\sqrt{1+\gamma'^2}\right)\cdot\Psi(x)\,dx.$$

Using the Hilbert space adjoint of $\mathcal{I}_\gamma^{-1}$ defined in (7.5) we may express D_1 as

$$D_1=\langle((\mathcal{I}_{\gamma+t\bar{\gamma}})^{-1})^*g_{\gamma+t\bar{\gamma}}-(\mathcal{I}_\gamma^{-1})^*g_\gamma,\Psi\rangle_{(L^2(I)^2)^*,L^2(I)^2}.$$

It can be shown that H_I is dense in $L^2(I)^2$. Thus there exists a sequence $\{\psi_k\}_k$ in H_I with $\lim_{k\to\infty}\psi_k=\Psi$ in $L^2(I)^2$. For fixed k we may hence estimate

$$\begin{aligned}|D_1|\le{}&|\langle(\mathcal{I}_{\gamma+t\bar{\gamma}})^*g_{\gamma+t\bar{\gamma}},\Psi-\psi_k\rangle_{(L^2(I)^2)^*,L^2(I)^2}|\\
&+|\langle(\mathcal{I}_{\gamma+t\bar{\gamma}})^*g_{\gamma+t\bar{\gamma}}-(\mathcal{I}_\gamma^{-1})^*g_\gamma,\psi_k\rangle_{H_I^*,H_I}|+|\langle(\mathcal{I}_\gamma^{-1})^*g_\gamma,\psi_k-\Psi\rangle_{(L^2(I)^2)^*,L^2(I)^2}|.\end{aligned}$$

The second term tends to zero since by Theorem 6.2 the Lagrange multipliers are weak-$*$ convergent in H_I^* if $\gamma+t\bar{\gamma}\to\gamma$ in $W^{1,\infty}(I)$ which is the case since $\gamma,\gamma+t\bar{\gamma}\in\mathcal{S}\subset W^{1,\infty}(I)$. For every $k\in\mathbb{N}$ we thus obtain that

$$|D_1|\le|\langle((\mathcal{I}_{\gamma+t\bar{\gamma}})^{-1})^*g_{\gamma+t\bar{\gamma}},\Psi-\psi_k\rangle_{(L^2(I)^2)^*,L^2(I)^2}+\langle(\mathcal{I}_\gamma^{-1})^*g_\gamma,\psi_k-\Psi\rangle_{(L^2(I)^2)^*,L^2(I)^2}|.$$

Both terms on the right tend to zero because $\psi_k \to \Psi$ in $L^2(I)^2$ and the family $\{(\mathcal{I}_\gamma^{-1})^* g_\gamma\}_{\gamma\in\mathcal{S}}$ is uniformly bounded in $(L^2(I)^2)^*$. Thus D_1 tends to zero for $t \to 0$.

For D_2 we use a similar argumentation as for C_2 in the proof of Lemma 7.3. We just replace $\mathbf{u}_{\gamma+t\bar{\gamma}}$ by λ_γ, I_+ by I_-, and $\chi_\gamma(x,\gamma)\sqrt{1+\gamma'^2}$ by $g_{\gamma+t\bar{\gamma}}(x,\gamma)\sqrt{1+(\gamma'+t\bar{\gamma}')^2}$. Because $\{\|\lambda_\gamma\|_{H^2(\Omega_\gamma)^2}\}_{\gamma\in\mathcal{S}}$ is bounded by Theorem 7.1 we get that D_2 tends to zero.

Acknowledgment. The author would like to thank Prof. Karl Kunisch from Karl-Franzens University, Graz, Austria for his support and dedicates this work to the memory of Prof. H. Jeggle.

REFERENCES

[1] C. Börgers, *Domain imbedding methods for the Stokes equations*, Numer. Math., 57 (1990), pp. 435–451.

[2] R. Glowinski, T.-W. Pan, and J. Periaux, *A fictitious domain method for Dirichlet problems and applications*, Comput. Methods Appl. Mech. Engrg., 111 (1994), pp. 283–303.

[3] R. Glowinski, T.-W. Pan, and J. Periaux, *A fictitious Domain method for external incompressible viscous flow modeled by Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 112 (1994), pp. 133–148.

[4] J. Dankova and J. Haslinger, *Numerical realization of a fictitious domain approach used in shape optimization.* I: *Distributed controls*, Appl. Math., 41 (1996), pp. 123–147.

[5] O. Pironneau, *On optimum profiles in Stokes flow*, J. Fluid Mech., 59 (1973), pp. 117–128.

[6] M. D. Gunzburger and H. Kim, *Existence of an optimal solution of a shape control problem for the stationary Navier–Stokes equations*, SIAM J. Control Optim., 36 (1998), pp. 895–909.

[7] J. A. Bello, E. Fernández-Cara, J. Lemoine, and J. Simon, *The differentiability of the drag with respect to the variations of a Lipschitz domain in a Navier–Stokes flow*, SIAM J. Control Optim., 35 (1997), pp. 626–640.

[8] K. Kunisch and G. Peichl, *Shape optimization for mixed boundary value problems based on an embedding domain method*, Dynam. Contin. Discrete Impuls. Systems, 4 (1998), pp. 439–478.

[9] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2, Springer-Verlag, Berlin, 1988.

[10] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations,* Vol. 1*,* Springer-Verlag, New York, 1994.

[11] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer Ser. Comput. Math., Springer-Verlag, Berlin, New York, 1986.

[12] R. B. Kellogg and J. E. Osborn, *A regularity result for the Stokes problem in a convex polygon*, J. Funct. Anal., 21 (1976), pp. 397–431.

[13] J. Necas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, Paris, 1967.

[14] H. Brezis, *Analyse Fonctionelle*, Masson, Paris, 1987.

[15] T. Slawig, *Domain Optimization for the Stationary Stokes and Navier–Stokes Equations by an Embedding Domain Technique*, Ph.D. Thesis, Shaker Verlag, Aachen, 1998.

[16] A. Grace, *Optimization Toolbox—For Use with MATLAB*, The Mathworks Inc., Natick, MA, 1994.

[17] L. P. Franca, T. J. R. Hughes, and R. Stenberg, *Stabilized finite element methods*, in Incompressible Computational Fluid Dynamics—Trends and Advances, R. A. Nicolaides and M. D. Gunzburger, eds., Cambridge University Press, Cambridge, UK, 1993.

[18] V. Girault and R. Glowinski, *Error analysis of a fictitious domain method applied to a Dirichlet problem*, Japan J. Indust. Appl. Math., 12 (1995), 487–514.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 159–178

LINEAR QUADRATIC OPTIMIZATION FOR SYSTEMS IN THE BEHAVIORAL APPROACH*

AUGUSTO FERRANTE† AND SANDRO ZAMPIERI‡

Abstract. In this paper the following formulation of the linear quadratic optimal control problem for dynamical systems in the behavioral setting is proposed: given a linear, time-invariant, and complete system, find the set of trajectories of the system that optimize a quadratic-type cost function and that satisfy some linear static constraints. This formulation provides a unifying framework, where several generalized versions of the classical LQ optimal control can be stated and solved.

The existence of solutions is first discussed. It is shown that a necessary and sufficient condition for the existence of solutions may be obtained as a by-product of a reduction procedure translating the problem into an equivalent one of minimum complexity. Such a procedure is based on the theory of ℓ^2-systems in the behavioral setting. Once the complexity is reduced, a parametrization of the set of optimal solutions is obtained by employing a behavioral realization technique and a two-step optimization procedure.

Key words. linear quadratic optimal control, behavioral approach, ℓ^2-systems, static constraints

AMS subject classifications. 49N05, 49N10, 93C05

PII. 0363012999352431

1. Introduction.

In the behavioral setting, a *system* is defined as a triple

$$\Sigma = (T, W, \mathcal{B}), \tag{1.1}$$

where T is a *time set*, W is the *alphabet* of the system, and the *behavior* $\mathcal{B}$ is the set of trajectories of the system. This definition of system has been introduced by Jan C. Willems in the latter half of the 1980s [22, 23] and, starting from these classical papers, a relevant amount of work has been produced in this direction.

In the above definition, the behavior $\mathcal{B}$ is simply the set of trajectories compatible with the (equations of the) system: the classical distinction of the signal of the system in *inputs*, *states*, and *outputs* is no longer present. Also, $\mathcal{B}$ is a completely arbitrary set of trajectories which may be described by equations (or inequalities) given in *implicit* form, and no causality assumptions are made. For these reasons the behavioral set-up has been revealed to be the most suitable framework to treat modeling and identification for many physical and economic systems where the above-mentioned features are crucial.

The contribution of the present paper is to formulate and solve the *linear quadratic* (LQ) optimal control problem in the behavioral setting: as it will be clarified in section 2, this formulation encompasses as particular cases the classical LQ optimal control problem, singular and cheap optimal control problems, and optimal control for descriptor systems, and allows us to deal with very general dynamic equations, static constraints, and cost functions. This formulation has the advantage, of a methodological and practical nature, of providing a unifying and elegant framework

*Received by the editors February 24, 1999; accepted for publication (in revised form) April 11, 2000; published electronically August 17, 2000.
http://www.siam.org/journals/sicon/39-1/35243.html

†Dipartimento di Elettronica e Informazione, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy (ferrante@elet.polimi.it).

‡Dipartimento di Elettronica ed Informatica, Università di Padova, via Gradenigo 6/A, 35131 Padova, Italy (zampi@dei.unipd.it).

where a large variety of physical and economic problems can be directly stated and solved with no need of preliminary manipulations. In particular, in economical systems, dynamic and algebraic equations are often mixed together, yielding an implicit dynamic, and nonstandard static constraints are often imposed. Moreover, in such systems, the distinction of signals in inputs, states, and outputs is sometimes artificial and not built-in in the physical model, so that it results naturally to avoid an a priori distinction of such a kind in the mathematical model too.

In recent years some works on the optimal control in the behavioral setting have been produced [25, 16]. In [25] the behavioral approach to LQ optimal control was first introduced in the following form: given a system and a quadratic cost function, find the subsystem whose behavior is constituted by all the trajectories w^* whose cost cannot be decreased by a perturbation of finite support. The subbehavior $\mathcal{B}^*$ of optimal trajectories is obtained by interconnecting the original system with a suitable "optimal controller." We recall that in the behavioral setting the interconnection of systems corresponds to the classical feedback control [24].

In the present paper we propose a different formulation of the LQ optimal control problem. Such a formulation, in our opinion, is more in the spirit of the classical optimal control problems, where the optimal solution is not required a priori to be of feedback type and boundary conditions are assigned. For example, the classical LQ optimal control problem for systems described by state equations can be expressed in the following way. Find the set of controls $u^*(\cdot)$ such that the corresponding trajectories $(x^*(\cdot), u^*(\cdot))$ (with fixed initial condition $x(0) = x_0$) satisfy

$$x(t+1) = Ax(t) + Bu(t) \tag{1.2}$$

and minimize the cost function

$$J(x,u) = \sum_{t=0}^{\infty} y^T(t)y(t), \tag{1.3}$$

where

$$y(t) = Cx(t) + Du(t). \tag{1.4}$$

The search for the optimal solution is performed in the space of all possible controls and the fact that the optimal solution happens to be of closed-loop type has no relation with the formulation of the problem and it is only due to the particular choice of boundary conditions. Note that, if the initial condition $x(0)$ is not fixed, one of the optimal controls is identically zero, the optimal cost is zero, and the problem is trivial. The fact that makes the problem meaningful is that $x(0)$ is fixed a priori.

In this paper we address the following optimization problem which appears to be the direct extension to the behavioral framework of the classical LQ optimal control problem just described. Given a linear time-invariant and complete (see section 1.1 for the precise definition) system with behavior $\mathcal{B}$ and a quadratic type cost function $J(\cdot)$, find the set $\mathcal{B}^* \subseteq \mathcal{B}$ of trajectories of the behavior which minimize the cost function J and satisfy a static constraint: for example, we may require that some components of the trajectory take some fixed values at particular time instants. In the classical LQ optimal control setting, this static constraint corresponds to the position $x(0) = x_0$, which fixes the initial state to a certain value x_0.

The first issue that we address is the problem solvability, which is shown to be equivalent to the existence of trajectories of the system satisfying the static constraints and giving rise to a finite cost. We then establish a reduction procedure which allows us to substitute the original system with a system with minimal complexity (i.e., a linear time-invariant and complete system having the smallest possible set of trajectories), so that the search of optimal trajectories can be performed among a minimal set. These issues are shown to be strictly connected to the theory of ℓ^2-systems, to which the first part of the paper is dedicated. In particular, some results on ℓ^2-systems, which appear to be of independent interest, are derived in section 3.

The paper is organized as follows. In subsection 1.1 we fix the notation and briefly recall some well-known results on behavioral theory of linear systems. In section 2 we give a precise mathematical formulation of our LQ optimization problem for autoregressive (AR) systems. In section 3 we provide a very easy necessary and sufficient condition for the problem solvability, and we present a procedure to reduce the complexity of the problem by eliminating from the behavior the trajectories corresponding to infinite cost. In section 4 we finally present the solution of our problem. In section 5 we briefly draw some conclusions.

1.1. Preliminaries. The complete theory of behavioral approach to dynamical systems is really ponderous, and it is beyond the scope of this paper to illustrate such a theory; for an illustration of the theory, we refer the reader to [23]. However, for the ease of the reader and to fix the notation we now detail the class of systems that will be considered in what follows and some of the very basic properties of the systems in this class.

We deal with systems of the form (1.1) with the set of integers $\mathbb{Z}$ as the time set (discrete-time systems), with the finite dimensional vector space $\mathbb{R}^q$ as the signal alphabet, and with a behavior that coincides with the set of solutions of a linear difference equation with (matrix valued) constant coefficients.

More precisely, if σ is the usual backward shift operator, let $\mathbb{R}[\sigma,\sigma^{-1}]^{l\times q}$ be that set of matrices whose entries are Laurent polynomials in σ (i.e., polynomials with both positive and negative powers of the indeterminate). A polynomial matrix in $\mathbb{R}[\sigma,\sigma^{-1}]^{l\times q}$ induces a linear shift-invariant operator from $(\mathbb{R}^q)^{\mathbb{Z}}$ to $(\mathbb{R}^l)^{\mathbb{Z}}$ in an obvious way. This operator is called matrix shift operator. A matrix shift operator $R(\sigma)$ in $\mathbb{R}[\sigma,\sigma^{-1}]^{l\times q}$ can be represented in the following way:

$$R(\sigma) = \sum_{i=k}^{K} R_i\sigma^i,$$

where $k \le K$ and $R_i \in \mathbb{R}^{l\times q}$. The kernel of this matrix shift operator has therefore the following structure:

$$\ker R(\sigma) = \left\{ w \in (\mathbb{R}^q)^{\mathbb{Z}} : \sum_{i=k}^{K} R_i w(t+i) = 0 \quad \forall t \in \mathbb{Z} \right\}. \tag{1.5}$$

In this paper we consider dynamical systems whose behaviors coincide with kernels of matrix shift operators. We recall that the class of AR systems coincides with the class of systems whose behaviors are linear, shift-invariant, and complete, where a behavior $\mathcal{B}$ is said to be complete if

$$w \in \mathcal{B} \quad \Leftrightarrow \quad w_{|[t_1,t_2]} \in \mathcal{B}_{|[t_1,t_2]} \quad \forall t_1, t_2 \in \mathbb{Z},$$

and where the symbol $w_{|[t_1,t_2]}$ means the restriction of the trajectory w to the interval $[t_1, t_2]$.

Summarizing, we consider the class of dynamical systems $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \ker R(\sigma))$, where $R(\sigma)$ is a matrix shift operator in $\mathbb{R}[\sigma, \sigma^{-1}]^{l \times q}$. These systems are called *AR systems* and the representations of the form (1.5) are called *AR representations*. The importance and the properties of AR systems and AR representations are widely investigated, especially in [23].

2. Problem formulation. In this section we give a precise mathematical formulation of our problem. Let $\mathcal{B} = \ker R(\sigma)$, $R(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{l \times q}$, be the behavior of an AR system $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathcal{B})$. Let w be a trajectory of $\mathcal{B}$: we consider a quadratic cost function J defined by

$$J(w) := ||w_{2_{|[0,+\infty)}}||_2^2 = \sum_{t=0}^{\infty} w_2^T(t) w_2(t), \tag{2.1}$$

where we have defined $w_2 := R_2(\sigma)w$ with $R_2(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{r \times q}$ being a polynomial matrix. Observe that, since w_2 is uniquely specified by w, $J(w)$ is a well-defined quadratic cost function of w.

Clearly we have $J(w) \geq 0$ for all the trajectories w and then the set $\mathcal{B}_0$ of trajectories of $\mathcal{B}$ which minimize J is trivially $\mathcal{B}_0 = \ker R(\sigma) \cap \ker R_2(\sigma) = \ker \left[\begin{smallmatrix} R(\sigma) \\ R_2(\sigma) \end{smallmatrix} \right]$.

The above minimization may be viewed as the behavioral counterpart of the classical LQ optimal control problem in the case when the initial state is unconstrained. As we have seen in the introduction, a static constraint renders the problem meaningful. This happens in the behavioral setting too. A very general linear constraint is the following:

$$r_h w(h) + r_{h+1} w(h+1) + \cdots + r_H w(H) = b, \tag{2.2}$$

where $r_i \in \mathbb{R}^{s \times q}$ and $b \in \mathbb{R}^s$. Defining the polynomial matrix $R_1(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{s \times q}$ as $R_1(\sigma) := \sum_{i=h}^{H} r_i \sigma^i$, and setting $w_1 := R_1(\sigma)w$, the constraint (2.2) may be rewritten in the more compact form

$$w_1(0) = (R_1(\sigma)w)(0) = b. \tag{2.3}$$

The main issue of this paper is the analysis and the solution of the following optimization problem. Find the set $\mathcal{T}(b)$ of all trajectories w such that

$$\left\{ \begin{array}{l} R(\sigma)w = 0, \\ w_1(0) = b, \\ J(w) = ||w_{2|[0,+\infty)}||_2^2 \text{ is minimal,} \end{array} \right. \tag{2.4}$$

where $R(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{l \times q}$, $R_1(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{s \times q}$, $R_2(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{r \times q}$, $w_1 := R_1(\sigma)w$, and $w_2 := R_2(\sigma)w$.

Problem (2.4) presents the following features which are compared with those of the classical LQ optimal control problem.

The dynamics. The first equation of (2.4) represents a very general linear dynamics. It encompasses, as a special case, the *descriptor systems* dynamics

$$Ex(t+1) = Ax(t) + Bu(t). \tag{2.5}$$

In fact, (2.5) may be rewritten in the form $R(\sigma)w = 0$, just by setting

$$w = \left[\begin{array}{c} x \\ u \end{array} \right] \tag{2.6}$$

and $R(\sigma) := R_0 + \sigma R_1$, where $R_0 := [-A \mid -B]$ and $R_1 := [E \mid 0]$. Moreover, linear difference equations (without control variables), which are typical of calculus of variations problems, are included in the dynamics of (2.4) too.

The static constraint. Instead of just fixing the initial condition, a constraint of the form (2.3) fixes some linear combinations of the values of (some component of) the trajectory at arbitrary times. As a particular case, (2.3) can force the trajectory to assume fixed values at fixed times. This possibility is very interesting in many applications (see, e.g., [4, 3] and references therein); a further example is the *landing problem* [20]: a control tower passes to an airplane a collection of way points that are vectors of position and velocity in $\mathbb{R}^3$ together with times of arrival. This is fitted very well by the type of constraints (2.3).

The cost function. The cost index (2.1) is remarkably general because $R_2(\sigma)$ is a Laurent polynomial matrix and this allows us to weigh together values of the trajectory $w(t)$ at different times and to consider the values of the trajectory $w(t)$ for $t \geq T$ with $T \in \mathbb{Z}$ arbitrarily fixed. Singular optimal control problems may be naturally formulated into this framework. With reference to the dynamics (2.5), such problems consist in minimizing a cost function (1.3) with y being given by (1.4) and $R := D^T D$ being singular. It is clear that such a cost function is expressible in the form (2.1) just by setting w as in (2.6) and $R_2(\sigma) = [C \mid D]$. Another interesting observation is that frequency shaping techniques in classical optimal control theory (see [1]) can be expressed directly in the time domain, using the generalized cost function (2.1). Finally, notice that the cost function (2.1) can be seen to be the discrete-time version of a *quadratic differential form*, as defined in [28].

The above observations show that formulation (2.4) is pretty general and covers a variety of optimal control problems. It is particularly useful in problems with implicit dynamics which are frequent in applications (see, e.g., [5, 13, 2] and references therein). Moreover, formulation (2.4) suits very well control problems of economic systems (one of the fields in which the behavioral framework seems more natural) in which the dynamic is often implicit and it is required that at some time instants $T_1, T_2, \ldots, T_r$, some reference variables match exactly their target values. One of these cases is the Leontief model for which many types of optimization problems have been stated and studied (see, e.g., [15] and references therein).

A different version of the problem (2.4) is the case of *finite time horizon,* i.e., the case when the cost function is

$$J(w) := \sum_{t=0}^{T} w_2^T(t) w_2(t). \tag{2.7}$$

As it will be clear in what follows, the solution of the problem with cost function (2.7) may be regarded as a subproblem of (2.4). For this reason we will not address the cost (2.7).

We have defined the dynamic equations of the system over $\mathbb{Z}$, even though the cost is determined by the values of $w_2(t)$ only for $t \geq 0$. This choice is based on the fact that most of the classical literature on behavioral approach to discrete-time linear systems deals with the biinfinite time axis case, even though the theory for discrete-time linear systems over the time axis $\mathbb{N}$ is quite analogous [22]. On the other hand, the choice of cost functions depending on the restriction of the signal on the nonnegative time axis is quite natural in most of the applications.

In some cases, however, it may be natural to consider the cost function

$$J(w) := \sum_{t=-\infty}^{\infty} w_2^T(t) w_2(t), \tag{2.8}$$

depending on the whole time axis $\mathbb{Z}$. This choice of considering the whole $\mathbb{Z}$ as the time axis permits us to treat the cost function (2.8) in the same framework. The corresponding optimization can be solved employing essentially the same techniques which we propose for problem (2.4).

3. Problem solvability and reduction of complexity. In this section we provide a procedure to test the solvability of the problem and to reduce its complexity. To this aim we will derive some interesting results on ℓ^2-systems [24, 27, 8].

Define $(\ell_+^2)^r$ as the set of all $v \in (\mathbb{R}^r)^{\mathbb{Z}}$ such that

$$||v_{|[0,+\infty)}||_2^2 := \sum_{t=0}^{\infty} v^T(t) v(t) < \infty.$$

Notice that $(\ell_+^2)^r$ is not a Hilbert space, since any trajectory which is zero on the nonnegative time axis have zero norm. In order to obtain a Hilbert space structure it is enough to consider equivalent two trajectories if they coincide on the nonnegative time axis [6, p. 7].

Given the problem (2.4), let

$$\begin{aligned} \mathcal{T}(b) := \{w \in (\mathbb{R}^q)^{\mathbb{Z}} \ : \ & R(\sigma)w = 0,\ w_1(0) = b, \\ & \text{and } J(w) \text{ is finite and minimal}\} \end{aligned} \tag{3.1}$$

be the set of all optimal trajectories. An important preliminary question is to understand when the previous problem admits a solution. In other words, we want to determine the vector space

$$B = \{b \in \mathbb{R}^s : \mathcal{T}(b) \neq \emptyset\}, \tag{3.2}$$

that is the set of vectors b for which an optimal solution exists.

It is clear that only trajectories $w \in \ker R(\sigma)$ such that $R_2(\sigma)w \in (\ell_+^2)^r$ are relevant in problem (2.4). In fact, all the other trajectories give rise to an infinite cost. Then, define

$$\mathcal{B}_f := \{w \in \ker R(\sigma) : R_2(\sigma)w \in (\ell_+^2)^r\} \tag{3.3}$$

to be the set of trajectories in $\mathcal{B}$ for which the corresponding cost is finite.

This set is linear and shift-invariant. However, it is not complete and thus it cannot be described by an AR representation.

This can be overcome by introducing the concept of *completion* of a behavior. Let $\mathcal{B} \subseteq (\mathbb{R}^q)^{\mathbb{Z}}$ be any behavior. The completion of $\mathcal{B}$ is given by the following set of trajectories:

$$CP(\mathcal{B}) := \{w \in (\mathbb{R}^q)^{\mathbb{Z}} : w_{|[t_1,t_2]} \in \mathcal{B}_{|[t_1,t_2]}\ \forall\ t_1, t_2 \in \mathbb{Z}\},$$

where $\mathcal{B}_{|[t_1,t_2]}$ is defined in section 1.1. Notice that, if we consider in $(\mathbb{R}^q)^{\mathbb{Z}}$ the pointwise convergence topology [23], then the concept of completion just defined coincides with closure with respect to this topology.

It is clear that $CP(\mathcal{B})$ is always complete and that, if $\mathcal{B}$ is linear and shift-invariant, then $CP(\mathcal{B})$ is linear, shift-invariant, and complete, and so it can be characterized by an AR representation. In particular, this is the case for the behavior

$$\mathcal{B}_r := CP(\mathcal{B}_f), \tag{3.4}$$

whose importance is clarified by the following proposition.

PROPOSITION 3.1. *Let B and $\mathcal{B}_r$ be defined in* (3.2) *and in* (3.4), *respectively. Then we have*

$$B = \{b \in \mathbb{R}^s : \ \exists w_1 \in R_1(\sigma)\mathcal{B}_r \text{ such that } w_1(0) = b\}. \tag{3.5}$$

Proof. One way is easy. Suppose, conversely, that there exists $v \in \mathcal{B}_r$ such that $(R_1(\sigma)v)(0) = b$. Then for all $n \in \mathbb{N}$ there exists $w \in \ker R(\sigma)$ such that $R_2(\sigma)w \in (\ell^2_+)^r$ and $w_{|[-n,n]} = v_{|[-n,n]}$. Choosing n big enough, we have that $(R_1(\sigma)w)(0) = (R_1(\sigma)v)(0) = b$. The existence of a trajectory satisfying the restrictions and providing a finite cost implies that $\mathcal{T}(b) \neq \emptyset$. □

The previous proposition has the following straightforward corollary which connects the problem solvability with a system-theoretic property: the trimness. We recall that an AR system $\Sigma = (T, W, \mathcal{B})$ is called *trim* if for all $\alpha \in W$ there exists $w \in \mathcal{B}$ such that $w(0) = \alpha$, [23, p. 188].

COROLLARY 3.1. *Problem* (2.4) *is solvable for all $b \in \mathbb{R}^s$ if and only if the behavior $\mathcal{B}_1 := R_1(\sigma)\mathcal{B}_r$ is trim.*

An interesting consequence of Proposition 3.1 is that if we substitute $\mathcal{B}$ with $\mathcal{B}_r$, the set of optimal trajectories $\mathcal{T}(b)$ does not change. Hence it may be convenient to perform the search for the optimal trajectories in the set $\mathcal{B}_r$ since $\mathcal{B}_r \subseteq \mathcal{B}$, and so this reduction will cause a simplification of the optimization problem. For these reasons it becomes interesting to find a procedure which provides an AR representation of $\mathcal{B}_r$ starting from $R(\sigma)$ and $R_2(\sigma)$. We will call the behavior $\mathcal{B}_r$ the *reduction* of $\mathcal{B} = \ker R(\sigma)$ with respect to $R_2(\sigma)$. Moreover, we will say that $\mathcal{B}$ is *reduced* with respect to $R_2(\sigma)$ if it coincides with its reduction with respect to $R_2(\sigma)$. Notice that $\mathcal{B}$ is reduced with respect to $R_2(\sigma)$ if and only if $\mathcal{B}_1 := R_2(\sigma)\mathcal{B}$ is reduced with respect to the identity or, equivalently, if and only if

$$\mathcal{B}_1 = CP(\mathcal{B}_1 \cap (\ell^2_+)^q).$$

In this case, we will say that the behavior $\mathcal{B}_1$ is *reduced.*

It can be shown that the problem of constructing AR representations of reduced behaviors is connected with the theory of ℓ^2-systems as presented in [24, 27, 12, 8, 11, 17, 21]. In the next subsection we will present some definitions and results of this theory.

3.1. ℓ^2-systems and their properties.

A linear shift-invariant system $\tilde{\Sigma} = (\mathbb{Z}, \mathbb{R}^q, \tilde{\mathcal{B}})$ is called an ℓ^2*-system* if $\tilde{\mathcal{B}}$ is a linear shift-invariant and closed (with respect to the ℓ^2 topology) subspace of

$$(\ell^2)^q := \left\{ w \in (\mathbb{R}^q)^{\mathbb{Z}} : ||w||_2^2 := \sum_{t=-\infty}^{+\infty} w^T(t)w(t) < \ \infty \right\}.$$

A particularly important class of ℓ^2-systems is the class of so-called *finite dimensional* ℓ^2-systems (see [24, p. 280]) which is the set of ℓ^2-systems whose behavior $\tilde{\mathcal{B}}$ satisfies the condition

$$\tilde{\mathcal{B}} = CP(\tilde{\mathcal{B}}) \cap (\ell^2)^q.$$

It can be shown that in this case $CP(\tilde{\mathcal{B}})$ is always controllable [24]. Moreover, if $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathcal{B})$ is a linear shift-invariant complete system, then $\tilde{\mathcal{B}} := \mathcal{B} \cap (\ell^2)^q$ is the behavior of a finite dimensional ℓ^2-system, and in this case $CP(\tilde{\mathcal{B}})$ is the behavior of the *controllable subsystem* of Σ (see [24, p. 266]), that is, the system $\Sigma_c = (\mathbb{Z}, \mathbb{R}^q, \mathcal{B}_c)$ such that

$$\mathcal{B}_c = CP(\{w \in \mathcal{B} : w \text{ has finite support}\}),$$

and which is the largest controllable subsystem of Σ.

Analogous considerations can be done if, instead of taking ℓ^2-systems, we consider ℓ^2_+-systems. In this case finite dimensional ℓ^2_+-systems are systems $\tilde{\Sigma} = (\mathbb{Z}, \mathbb{R}^q, \tilde{\mathcal{B}})$ such that

$$\tilde{\mathcal{B}} = CP(\tilde{\mathcal{B}}) \cap (\ell^2_+)^q.$$

Also in this case, if $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathcal{B})$ is a linear shift-invariant complete system, then $\tilde{\mathcal{B}} := \mathcal{B} \cap (\ell^2_+)^q$ is the behavior of a finite dimensional ℓ^2_+-system. However, now the behavior is not controllable, but only *stabilizable* (see [26]). More precisely, $CP(\tilde{\mathcal{B}})$ is the behavior of the stabilizable subsystem of Σ namely the system $\Sigma_s = (\mathbb{Z}, \mathbb{R}^q, \mathcal{B}_s)$ such that

$$\mathcal{B}_s = CP\left(\left\{w \in \mathcal{B} : \lim_{t\to+\infty} w(t) = 0\right\}\right)$$

which is the largest stabilizable subsystem of Σ. This is a consequence of the following proposition.

PROPOSITION 3.2. *Let $R(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{p\times q}$ be full row rank and $\mathcal{B} = \ker R(\sigma)$. Moreover, let $F_s(\sigma), F_i(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{p\times p}$, and $R'(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{p\times q}$ be polynomial matrices such that $\det F_s(\sigma)$ has zeros in $\mathbb{C}_< := \{z \in \mathbb{C} : \ |z| < 1\}$, $\det F_i(\sigma)$ has zeros in $\mathbb{C}_\geq := \{z \in \mathbb{C} : \ |z| \geq 1\}$, $R'(\sigma)$ is left prime, and*

$$R(\sigma) = F_i(\sigma)F_s(\sigma)R'(\sigma). \tag{3.6}$$

Then

$$CP(\mathcal{B} \cap (\ell^2_+)^q) = \ker F_s(\sigma)R'(\sigma).$$

Proof. We show first that

$$\ker R(\sigma) \cap (\ell^2_+)^q \subseteq \ker F_s(\sigma)R'(\sigma),$$

which implies immediately that

$$CP(\mathcal{B} \cap (\ell^2_+)^q) \subseteq \ker F_s(\sigma)R'(\sigma).$$

Let $w \in \ker R(\sigma)$ and $w \in (\ell^2_+)^q$. Since $R'(\sigma)$ is left prime, then there exists a polynomial matrix $X(\sigma)$ such that $R'(\sigma)X(\sigma) = I$, where I is the identity matrix. Then define $w_1 := X(\sigma)R'(\sigma)w$ and $w_2 := w - w_1$. First notice that $w_2 \in \ker R'(\sigma)$ and so $w_2 \in \ker F_s(\sigma)R'(\sigma)$. Notice, moreover, that $w_1 \in X(\sigma)\ker F_i(\sigma)F_s(\sigma)$. Since $w \in (\ell^2_+)^q$, then it is clear that

$$\lim_{t\to+\infty} w(t) = 0.$$

We can argue that $\lim_{t\to+\infty} w_1(t) = 0$. Let $v \in \ker F_i(\sigma)F_s(\sigma)$ such $w_1 = X(\sigma)v$. Then $v = R'(\sigma)w_1$ and so $\lim_{t\to+\infty} v(t) = 0$. Therefore we have $F_s(\sigma)v \in \ker F_i(\sigma)$ and that $\lim_{t\to+\infty} F_s(\sigma)v(t) = 0$. It is easy to see that this can happen if and only if $F_s(\sigma)v = 0$. This implies that $w_1 \in \ker F_s(\sigma)R'(\sigma)$ and so $w \in \ker F_s(\sigma)R'(\sigma)$.

Suppose, conversely, that $w \in \ker F_s(\sigma)R'(\sigma)$. We want to show that for all $n \in \mathbb{N}$ there exists $w' \in \ker R(\sigma) \cap (\ell^2_+)^q$ such that $w_{|[-n,n]} = w'_{|[-n,n]}$. This implies that

$$\ker F_s(\sigma)R'(\sigma) = CP(\ker F_s(\sigma)R'(\sigma)) \subseteq \mathcal{B}.$$

Fix $n \in \mathbb{N}$ and let $w_1 := X(\sigma)R'(\sigma)w$ and $w_2 := w - w_1$. Since $R'(\sigma)w \in \ker F_s(\sigma)$, then $w_1 \in (\ell^2_+)^q$. On the other hand, since $w_2 \in R'(\sigma)$ and since $\ker R'(\sigma)$ is controllable, then there exists $w'_2 \in R'(\sigma)$ such that $w_{2|[-n,n]} = w'_{2|[-n,n]}$ and such that $w'_{|[N,+\infty)} = 0$ for $N \in \mathbb{N}$ big enough. Define $w' := w_1 + w'_2$. Then it is clear that $w' \in \ker R(\sigma) \cap (\ell^2_+)^q$ and that $w_{|[-n,n]} = w'_{|[-n,n]}$. □

As a corollary of the previous proposition, we obtain an effective characterization of stabilizable behaviors which has been already proposed in [26]. Its proof easily follows from the previous proposition and from proposition 4.3 in [23].

COROLLARY 3.2. *Let $R(\sigma) \in \mathbb{R}[\sigma,\sigma^{-1}]^{p\times q}$ be a rank r polynomial matrix. Then $\mathcal{B} := \ker R(\sigma)$ is stabilizable if and only if*

$$\operatorname{rank} R(\lambda) = r$$

for all $\lambda \in \mathbb{C}_{\geq}$.

3.2. Construction of reduced behaviors. Consider again the problem of constructing AR representations of reduced behaviors. In the particular case when $R_2(\sigma) = I$, the reduction may be obtained by employing Proposition 3.2. In order to extend such a procedure to the general case, we need the following technical lemma.

LEMMA 3.1. *Let $\mathcal{B} = \ker R(\sigma)$, where $R(\sigma) \in \mathbb{R}[\sigma,\sigma^{-1}]^{l\times q}$, and let $R_2(\sigma) \in \mathbb{R}[\sigma,\sigma^{-1}]^{r\times q}$. If $\{v_n\}_{n=0}^{\infty}$ is a sequence in $R_2(\sigma)\mathcal{B}$, converging in the pointwise convergence topology, then there exists a sequence $\{w_n\}_{n=0}^{\infty}$ in $\mathcal{B}$ converging in the pointwise convergence topology, such that $v_n = R_2(\sigma)w_n$.*

Proof. We first suppose that $\mathcal{B} = (\mathbb{R}^q)^{\mathbb{Z}}$ and consider the scalar case $p = q = 1$. Suppose that

$$R_2(\sigma) = \sum_{i=l}^{L} R_i\sigma^i,$$

where R_l, R_L are nonzero reals. Let $\{v_n\}_{n=0}^{\infty}$ be a sequence in $R_2(\sigma)\mathcal{B}$, and let $\{w_n\}_{n=0}^{\infty}$ be a sequence of trajectories satisfying

(i) $w_n(t) = 0$ for all n and for all $t \in [l, L-1]$, and

(ii) $v_n = R_2(\sigma)w_n$.

It is clear that (i) and (ii) fix uniquely w_n. We want to show that if $\{v_n\}_{n=0}^{\infty}$ converges in the pointwise convergence topology, then the sequence $\{w_n\}_{n=0}^{\infty}$ converges in the same topology, i.e., for all $\bar{t} \in \mathbb{Z}$ the sequence of real numbers $\{w_n(\bar{t})\}_{n=0}^{\infty}$ converges. This is clearly true for $\bar{t} \in [l, L-1]$. Suppose, by induction, that $\{w_n(\bar{t})\}_{n=0}^{\infty}$ for all $\bar{t}$ in an interval $[h, H]$. Then, since we have

$$v_n(H+1-L) = R_l w_n(H+1+l-L) + \cdots + R_{L-1}w_n(H) + R_L w_n(H+1)$$

and since $v_n(H+1-L), w_n(H+1-L+l), \ldots, w_n(H)$ all converge as sequences in n, then also $w_n(H+1)$ must converge. In the same way we can show that $w_n(h-1)$ must converge.

Consider now the vector case and consider the Smith form [9] of $R_2(\sigma)$,

$$U(\sigma)R_2(\sigma)V(\sigma) = \begin{bmatrix} \Lambda(\sigma) & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Lambda(\sigma)$ is a diagonal matrix and $U(\sigma), V(\sigma)$ are unimodular. Take $\bar{v}_n := U^{-1}(\sigma)v_n$. Since the matrix shift operators are continuous, then $\{\bar{v}_n\}_{n=0}^{\infty}$ converges. Moreover, since $V(\sigma)$ is onto, we have that

$$\bar{v}_n \in \operatorname{im} \begin{bmatrix} \Lambda(\sigma) & 0 \\ 0 & 0 \end{bmatrix}.$$

Using the result obtained in the scalar case, we may argue that there exists a converging sequence $\{\bar{w}_n\}_{n=0}^{\infty}$ such that

$$\bar{v}_n = \begin{bmatrix} \Lambda(\sigma) & 0 \\ 0 & 0 \end{bmatrix} \bar{w}_n.$$

Finally letting $w_n = V^{-1}(\sigma)\bar{w}_n$, we have that $\{w_n\}_{n=0}^{\infty}$ converges and $v_n = R_2(\sigma)w_n$.

Suppose, finally, that $v_n \in R_2(\sigma)\mathcal{B}$ and that $\{v_n\}_{n=0}^{\infty}$ converges. Then there exists $w_n' \in \mathcal{B}$ such that $v_n = R(\sigma)w_n'$. Letting

$$\bar{R}(\sigma) := \begin{bmatrix} R(\sigma) \\ R_2(\sigma) \end{bmatrix},$$

we have that

$$\bar{R}(\sigma)w_n' = \begin{bmatrix} R(\sigma) \\ R_2(\sigma) \end{bmatrix} w_n' = \begin{bmatrix} 0 \\ v_n \end{bmatrix}$$

converges and hence, by the previous arguments, there exists a convergent sequence $\{w_n\}_{n=0}^{\infty}$ such that

$$\bar{R}(\sigma)w_n = \begin{bmatrix} R(\sigma) \\ R_2(\sigma) \end{bmatrix} w_n = \begin{bmatrix} 0 \\ v_n \end{bmatrix}.$$

This implies that $R(\sigma)w = 0$ and that $v_n = R_2(\sigma)w_n$. □

Remark. The previous lemma is equivalent to the fact that the linear map

$$\begin{array}{rcccl} R_2(\sigma)_{|\mathcal{B}} & : & \mathcal{B} & \longrightarrow & R_2(\sigma)\mathcal{B}, \\ & & w & \mapsto & R_2(\sigma)w \end{array}$$

is *open* [7, p. 221]. In fact, it can be proved that Lemma 3.1 is a particular case of the *open mapping theorem* [18].

We are now in position to prove the next proposition.

PROPOSITION 3.3. *Let $\mathcal{B} = \ker R(\sigma)$, where $R(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{l\times q}$, and let $R_2(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{r\times q}$. Then the reduced behavior $\mathcal{B}_r$ (defined in* (3.4)*) is given by the expression*

$$\mathcal{B}_r = \ker \begin{bmatrix} R(\sigma) \\ M(\sigma)R_2(\sigma) \end{bmatrix}, \tag{3.7}$$

where $M(\sigma) \in \mathbb{R}[\sigma, \sigma^{-1}]^{g\times r}$ is such that $\ker M(\sigma) = CP(R_2(\sigma)\mathcal{B} \cap (\ell_+^2)^r))$.

Proof. We first verify that

$$\mathcal{B}_r \subseteq \mathcal{B} \cap \ker M(\sigma)R_2(\sigma).$$

Clearly $\mathcal{B}_r \subseteq \mathcal{B}$. Moreover, if $w \in \mathcal{B}_r$, then $R(\sigma)w = 0$ and $R_2(\sigma)w \in (\ell^2_+)^r$, so that $R_2(\sigma)w \in R_2(\sigma)\mathcal{B} \cap (\ell^2_+)^r \subseteq \ker M(\sigma)$ or, equivalently, $w \in \ker[M(\sigma)R_2(\sigma)]$.

Suppose, conversely, that $w \in \mathcal{B} \cap \ker M(\sigma)R_2(\sigma)$. Let $\mathcal{B}_f$ be the behavior defined in (3.3). We have to show that

$$w \in CP(\mathcal{B}_f).$$

Since $w \in \ker M(\sigma)R_2(\sigma)$, then $v := R_2(\sigma)w \in \ker M(\sigma) = CP(R_2(\sigma)\mathcal{B} \cap (\ell^2_+)^r)$, and so there exists a sequence $\{v_n\}_{n=0}^\infty$ in $R_2(\sigma)\mathcal{B} \cap (\ell^2_+)^r$ converging to v in the pointwise convergence topology. By the previous lemma we can argue that there exists a sequence $\{w_n\}_{n=0}^\infty$ in $\mathcal{B}$ that converges in the pointwise convergence topology to $w' \in \mathcal{B}$ and such that $v_n = R_2(\sigma)w_n$ and hence, by the continuity of $R_2(\sigma)$, we have $R_2(\sigma)w = R_2(\sigma)w'$. Consider the new sequence $\{w_n - w' + w\}_{n=0}^\infty$ and observe that its elements are still in $\mathcal{B}$. Moreover, observe that $R_2(\sigma)w_n = v_n \in (\ell^2_+)^r$ so that $w_n \in \mathcal{B}_f$. Consequently, w is in the closure of the previous set, just as we need to show. □

By the previous proposition, an essential step in determining $\mathcal{B}_r$ is the computation of an AR representation of $CP(R_2(\sigma)\mathcal{B} \cap (\ell^2_+)^r)$. To this aim we first look for an AR representation of $R_2(\sigma)\mathcal{B}$. Since $w_2 \in R_2(\sigma)\mathcal{B}$ if and only if

$$\begin{bmatrix} I \\ 0 \end{bmatrix} w_2 = \begin{bmatrix} R_2(\sigma) \\ R(\sigma) \end{bmatrix} w \tag{3.8}$$

for some $w \in (\mathbb{R}^q)^{\mathbb{Z}}$, an AR representation of $R_2(\sigma)\mathcal{B}$,

$$R_2(\sigma)\mathcal{B} = \ker N(\sigma),$$

can be obtained by eliminating the latent variable w in (3.8) as suggested in [24, p. 265]. After this elimination we are in the range of application of Proposition 3.2. Therefore, it is possible to obtain an AR representation for $CP(R_2(\sigma)\mathcal{B} \cap (\ell^2_+)^r)$:

$$CP(R_2(\sigma)\mathcal{B} \cap (\ell^2_+)^r) = \ker M(\sigma).$$

Notice that the previous arguments prove the following result.

Corollary 3.3. *The behavior $\mathcal{B} = \ker R(\sigma)$ is reduced with respect to $R_2(\sigma)$ if and only if $R_2(\sigma)\mathcal{B}$ is stabilizable.*

Therefore, if we want to verify whether $\mathcal{B}$ is reduced with respect to $R_2(\sigma)$, we can simply apply Corollary 3.2 to any AR representation of $R_2(\sigma)\mathcal{B}$.

In view of the previous arguments, from now on, we can assume without loss of generality that in problem (2.4) the behavior $\mathcal{B} = \ker R(\sigma)$ is reduced with respect to $R_2(\sigma)$.

4. Optimal trajectories in AR systems. In this section we show how the problem of finding the set of the optimal trajectories of a system Σ can be translated into an LQ optimization problem for a system in state form. This will be done employing a driving variable state space representation of Σ. This second optimization problem is not the classical one, because it has a linear constraint on the initial state and on the initial input. However, we will give a closed form solution using a two-step optimization.

In the next subsection we derive a state space reformulation of problem (4.1). In subsection 4.2 we describe the state space counterpart of the results of section 3. More precisely, we derive a necessary and sufficient condition, based on the state space representation, for the existence of solutions. We also describe a procedure to perform the reduction described in subsection 3.2 in terms of the state space representation. These results allow us to derive in subsection 4.3 a parametrization of the set of optimal trajectories.

4.1. From AR to state space representation. We recall from [23] that given an AR system $\Sigma = (\mathbb{Z}, \mathbb{R}^q, \mathcal{B})$, there exist $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{q\times n}, D \in \mathbb{R}^{q\times m}$ such that $w \in \mathcal{B}$ if and only if there exist $x \in (\mathbb{R}^n)^{\mathbb{Z}}$ and $u \in (\mathbb{R}^m)^{\mathbb{Z}}$ such that

$$\left\{\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \\ w(t) &= Cx(t) + Du(t) \end{aligned}\right. \qquad \forall t \in \mathbb{Z}. \tag{4.1}$$

The signal x is called *state variable* while the signal u is called *driving variable.* The representation (4.1) of the original system is called *driving variable state representation.* This state space representation has been introduced in [23], where the properties of minimal representations are also analyzed. For our purposes it is sufficient to recall that minimal driving variable representations exist, they may be computed by employing linear algebra techniques, and they have the *state trimness* property. The realization (4.1) is said to be *state trim* [23] if and only if

$$\begin{gathered} \{x_0 \in \mathbb{R}^n : \exists x \in (\mathbb{R}^n)^{\mathbb{Z}}, u \in (\mathbb{R}^m)^{\mathbb{Z}}, w \in \mathcal{B} \text{ satisfying (4.1)} \\ \text{such that } x(0) = x_0\} = \mathbb{R}^n. \end{gathered} \tag{4.2}$$

The procedure determining the set of all optimal trajectories $\mathcal{T}(b)$ proposed in this paper is based on the driving variable representation. Suppose that we are dealing with problem (2.4), where we can suppose that $\mathcal{B} = \ker R(\sigma)$ is reduced with respect to $R_2(\sigma)$. Consider the AR system $\bar{\Sigma} = (\mathbb{Z}, \mathbb{R}^{q+s+r}, \bar{\mathcal{B}})$, where

$$\begin{aligned} \bar{\mathcal{B}} &= \ker \begin{bmatrix} R(\sigma) & 0 & 0 \\ R_1(\sigma) & -I & 0 \\ R_2(\sigma) & 0 & -I \end{bmatrix} \\ &= \left\{ \begin{bmatrix} w \\ w_1 \\ w_2 \end{bmatrix} \in (\mathbb{R}^{q+s+r})^{\mathbb{Z}} : R(\sigma)w = 0,\ w_1 = R_1(\sigma)w,\ w_2 = R_2(\sigma)w \right\}, \end{aligned}$$

and consider a driving variable state representation of $\bar{\Sigma}$:

$$\left\{\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \\ w(t) &= Cx(t) + Du(t), \\ w_1(t) &= C_1x(t) + D_1u(t), \\ w_2(t) &= C_2x(t) + D_2u(t), \end{aligned}\right. \qquad t \in \mathbb{Z}, \tag{4.3}$$

where $x \in (\mathbb{R}^n)^{\mathbb{Z}}$ is the state variable and $u \in (\mathbb{R}^m)^{\mathbb{Z}}$ is the driving variable. Since (4.3) is a state representation of the system $\bar{\Sigma}$, then $(w, w_1, w_2) \in \bar{\mathcal{B}}$ if and only if there exists a state trajectory x and a driving variable trajectory u such that (4.3) is satisfied for all $t \in \mathbb{Z}$. The complexity of the solution of our problem will depend on the dimension n of the state space of the representation (4.3). Hence, it is convenient to consider a minimal state representation of $\bar{\mathcal{B}}$. For this reason, from now on, without loss of generality, we assume that (4.3) is a minimal representation of $\bar{\mathcal{B}}$ and, in particular, that state trim property (4.2) holds true.

Consider now the following optimization problem. Find the set of trajectories $(x, u) \in (\mathbb{R}^{n+m})^{\mathbb{N}}$ such that

$$(4.4) \quad \begin{cases} x(t+1) = Ax(t) + Bu(t), \qquad t \in \mathbb{N}, \\ C_1 x(0) + D_1 u(0) = b, \\ J(x,u) := \sum_{t=0}^{\infty} [x^T(t) u^T(t)] \begin{bmatrix} C_2^T C_2 & C_2^T D_2 \\ D_2^T C_2 & D_2^T D_2 \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \text{ is minimal.} \end{cases}$$

The following proposition shows that problem (4.4) is equivalent to problem (2.4).

PROPOSITION 4.1. *Let $\mathcal{T}(b)$ be the set of optimal trajectories defined in* (3.1). *We have that $w \in \mathcal{T}(b)$ if and only if there exists $(x, u) \in (\mathbb{R}^{n+m})^{\mathbb{N}}$ satisfying the requirements of problem* (4.4) *and such that*

$$w(t) = Cx(t) + Du(t) \qquad \forall t = 0, 1, 2, \ldots .$$

Proof. Suppose that $w \in \mathcal{T}(b)$. Then there exists $(x, u) \in (\mathbb{R}^{n+m})^{\mathbb{Z}}$ satisfying (4.3) for all $t \in \mathbb{Z}$. Let $\bar{x} := x_{|[0,+\infty)}$ and $\bar{u} := u_{|[0,+\infty)}$. Then $(\bar{x}, \bar{u})$ satisfies problem (4.4). Actually, the first two equations of (4.4) are clearly satisfied. It remains to show the minimality of $J(\bar{x}, \bar{u})$. Let $(\bar{x}', \bar{u}') \in (\mathbb{R}^{n+m})^{\mathbb{N}}$ such that

$$\begin{cases} \bar{x}'(t+1) = A\bar{x}'(t) + B\bar{u}'(t), \\ C_1 \bar{x}'(0) + D_1 \bar{u}'(0) = b. \end{cases}$$

By trimness of the state representation (4.3), there exists $(x', u') \in (\mathbb{R}^{n+m})^{\mathbb{Z}}$ such that $x'(t+1) = Ax'(t) + Bu'(t)$ for all $t \in \mathbb{Z}$ and $\bar{x}' = x'_{|[0,+\infty)}$ and $\bar{u}' = u'_{|[0,+\infty)}$. Define $w' := Cx' + Du'$, $w_1' := C_1 x' + D_1 u'$, and $w_2' := C_2 x' + D_2 u'$. Then it is clear that $R(\sigma) w' = 0$, $w_1' = R_1(\sigma) w'$, and $w_2' = R_2(\sigma) w'$, and so

$$J(\bar{x}, \bar{u}) = ||w_{2|[0,+\infty)}||_2^2 \le ||w'_{2|[0,+\infty)}||_2^2 = J(\bar{x}', \bar{u}'),$$

which shows the minimality of $J(\bar{x}, \bar{u})$.

Suppose, conversely, that $(\bar{x}, \bar{u}) \in (\mathbb{R}^{n+m})^{\mathbb{N}}$ satisfies the requirements of (4.4) and that $w(t) = Cx(t) + Du(t)$ for all $t = 0, 1, 2, \ldots$. Then, by trimness of the state representation (4.3), there exists $(x, u) \in (\mathbb{R}^{n+m})^{\mathbb{Z}}$ such that $x(t+1) = Ax(t) + Bu(t)$ for all $t \in \mathbb{Z}$ and $\bar{x} = x_{|[0,+\infty)}$ and $\bar{u} = u_{|[0,+\infty)}$. Define $w := Cx + Du$, $w_1 := C_1 x + D_1 u$, and $w_2 := C_2 x + D_2 u$. Then it is clear that $R(\sigma) w = 0$, $w_1 = R_1(\sigma) w$, $w_2 = R_2(\sigma) w$, and $w_1(0) = b$. It remains to show the minimality of $||w_{2|[0,+\infty)}||_2^2$. Actually, suppose that w' is such that $R(\sigma) w' = 0$ and let $w_1' = R_1(\sigma) w'$ and $w_2' = R_2(\sigma) w'$. Suppose, moreover, that $w_1(0) = b$. Then there exists $(x', u') \in (\mathbb{R}^{n+m})^{\mathbb{Z}}$ satisfying (4.3) for all $t \in \mathbb{Z}$. Consequently, if we define $\bar{x}' := x'_{|[0,+\infty)}$ and $\bar{u}' := u'_{|[0,+\infty)}$, then

$$||w_{2|[0,+\infty)}||_2^2 = J(\bar{x}, \bar{u}) \le J(\bar{x}', \bar{u}') = ||w'_{2|[0,+\infty)}||_2^2,$$

and this implies the minimality of $||w_{2|[0,+\infty)}||_2^2$. □

By the previous proposition we can argue that (2.4) is solvable, i.e., $\mathcal{T}(b) \neq \emptyset$ for all $b \in \mathbb{R}^s$ if and only if for any b there exists a trajectory (x, u) compatible with the equations

$$(4.5) \qquad \begin{cases} x(t+1) = Ax(t) + Bu(t), \qquad t \in \mathbb{N}, \\ C_1 x(0) + D_1 u(0) = b, \end{cases}$$

and such that $J(x, u) < \infty$.

4.2. Complexity reduction in state space representation. It is possible to find a very easy necessary and sufficient condition for the solvability of problem (2.4) based on the representation (4.3). To derive this condition we need some definitions and a technical lemma.

Given the system

$$\begin{cases} x(t+1) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases} \tag{4.6}$$

we will denote by $\mathcal{V}_+$ the space of states $x_0 \in \mathbb{R}^n$ for which there exists a trajectory u, x, y compatible with equations of system (4.6) and such that $y(t) = 0$ for all $t \geq 0$ and $x(0) = x_0$. We will call $\mathcal{V}_+$ the *weakly unobservable space of* (A, B, C, D). Similarly, we will denote by $\mathcal{V}_-$ the space of states $x_0 \in \mathbb{R}^n$ for which there exists a trajectory u, x, y compatible with equations of system (4.6) and such that $y(t) = 0$ for all $t \leq 0$ and $x(0) = x_0$. $\mathcal{V}_-$ will be called the *backward weakly unobservable space of* (A, B, C, D). The space

$$\mathcal{X}_c := \mathcal{X}_s(A) + \langle A|\text{im } B\rangle + \mathcal{V}_+, \tag{4.7}$$

where $\mathcal{X}_s(A)$ is the stability space of A, $\langle A|\text{im } B\rangle$ is the controllability space of (A, B), and $\mathcal{V}_+$ is the weakly unobservable space of (A, B, C, D), is called the *output stabilizability space of* (A, B, C, D). We recall from [10] the following result.

Lemma 4.1. *The space $\mathcal{X}_c$ of (A, B, C_2, D_2) coincides with the set of initial condition $x(0)$ such that there exists u yielding to a finite cost $J(x, u)$ in* (4.4).

We now prove a technical result concerning the space $\mathcal{X}_c$ that will be useful below.

Lemma 4.2. *Given the system* (4.6), *let $\mathcal{R} = \langle A|\text{im } B\rangle$ be its controllability space, $\mathcal{V}_+$ be its weakly unobservable space, and $\mathcal{V}_-$ be its backward weakly unobservable space. Then we have*

$$\mathcal{V}_- \subseteq \mathcal{V}_+ + \mathcal{R}. \tag{4.8}$$

Proof. Let n be the dimension of the matrix A, and let $\bar{x}$ be a point of $\mathcal{V}_-$. By definition there exists a trajectory u, x, y compatible with equations of system (4.6) and such that $y(t) = 0$ for all $t \leq 0$ and $x(0) = \bar{x}$. This clearly implies that $x(-k)$ is an element of the set $\mathcal{V}_+^k$ of states of system (4.6) which is weakly unobservable in k steps, i.e., the set of points $x_0 \in \mathbb{R}^n$ for which there exists a trajectory u, x, y compatible with equations of system (4.6) and such that $y(t) = 0$ for all $0 \leq t \leq k$ and $x(0) = x_0$. Since the sequence of sets $\{\mathcal{V}_+^t\}_{t=1,2,\ldots}$ is decreasing and for $t \geq n$ it becomes stationary [23], we have that for k sufficiently large $x(-k) \in \mathcal{V}_+$. Hence there exists a trajectory u_+, x_+, y_+ compatible with equations of system (4.6) and such that $x_+(t) = x(t)$ for $t \leq -k$ and $y_+(t) = 0$ for $t \geq -k$. Then, $x_+(0) \in \mathcal{V}_+$. Moreover, $x_+(0) - \bar{x} \in \mathcal{R}$, since they are both reachable starting from the state $x(-k)$. Therefore, $\bar{x} \in \mathcal{R} + \mathcal{V}_+$, and this concludes the proof. □

The following result provides a link between the reduced behaviors setting and system theoretic properties of the corresponding driving variable representations.

Proposition 4.2. *Consider problem* (2.4) *and the driving variable representation* (4.3). *Moreover, let $\mathcal{X}_c$ be the output stabilizability space of (A, B, C_2, D_2). Then, $\mathcal{B} = \ker R(\sigma)$ is reduced with respect to $R_2(\sigma)$ if and only if $\mathcal{X}_c = \mathbb{R}^n$.*

Proof. Suppose that $\mathcal{X}_c = \mathbb{R}^n$. Then for any trajectory $w_2 \in R_2(\sigma)\mathcal{B}$ there exists a state trajectory x and an input trajectory u such that

$$\begin{cases} x(t+1) = Ax(t) + Bu(t), \\ \quad w_2(t) = C_2x(t) + D_2u(t), \end{cases} \qquad t \in \mathbb{Z}. \tag{4.9}$$

Then in view of Lemma 4.1 and taking into account the state separation property, it is clear that there exist $\bar{w}_2 \in R_2(\sigma)\mathcal{B}$, $\bar{x}$, and $\bar{u}$ satisfying (4.9) and such that $w_2(t) = \bar{w}_2(t)$, $x(t) = \bar{x}(t)$, $t \leq 0$, and $\lim_{t\to\infty} \bar{w}_2(t) = 0$. This is means, by definition, that $R_2(\sigma)\mathcal{B}$ is stabilizable, or, in view of Corollary 3.3, that $\mathcal{B} = \ker R(\sigma)$ is reduced with respect to $R_2(\sigma)$.

Suppose, conversely, that $\mathcal{B} = \ker R(\sigma)$ is reduced with respect to $R_2(\sigma)$, or, equivalently, that $R_2(\sigma)\mathcal{B}$ is reduced, and let $R_2(\sigma)\mathcal{B} = \ker M(\sigma)$ with $M(\sigma)$ full row rank. Then, by Proposition 3.2, $M(\sigma) = F(\sigma)M'(\sigma)$ with $F(\sigma)$ square with stable determinant and $M'(\sigma)$ left prime. Let $X(\sigma)$ be such that $M'(\sigma)X(\sigma) = I$. Suppose that $x_0 \in \mathbb{R}^n$. Then, by trimness, there exist x, w_2 satisfying the state equations in (4.3) and such that $x(0) = x_0$. Obviously, $w_2 \in R_2(\sigma)\mathcal{B}$. As in the proof of Proposition 3.2 define $w_2' := X(\sigma)M'(\sigma)w_2$ and $w_2'' := w_2 - w_2'$. Since $M'(\sigma)w_2 \in \ker F(\sigma)$, then $w_2' \in (\ell_+^2)^r$. On the other hand, since $w_2'' \in \ker M'(\sigma)$ and since $\ker M'(\sigma)$ is controllable, there exists $\bar{w}_2'' \in \ker M'(\sigma)$ such that $\bar{w}''_{2|(-\infty,0]} = w''_{2|(-\infty,0]}$ and such that $\bar{w}''_{2|[N,+\infty)} = 0$ for $N \in \mathbb{N}$ big enough. Define $\bar{w}_2 := w_2' + \bar{w}_2''$. Then it is clear that $\bar{w}_2 \in R_2(\sigma)\mathcal{B} \cap (\ell_+^2)^r$ and that $\bar{w}_{2|(-\infty,0]} = w_{2|(-\infty,0]}$. Consider the state trajectory $\bar{x}$ such that $\bar{x}, \bar{w}_2$ satisfy the state equations in (4.3). Then it is clear that $\bar{x}(0) \in \mathcal{X}_c$. From Lemma 4.2 it follows that $\bar{x}(0) - x_0 \in \mathcal{X}_c$, and hence, since $\mathcal{X}_c$ is a linear space, that $x_0 \in \mathcal{X}_c$. □

We are now ready to prove the following corollary which gives an easy necessary and sufficient condition for the solvability of our problem.

COROLLARY 4.1. *Consider problem* (2.4) *and the driving variable representation* (4.3). *Let* $\mathcal{T}(b)$ *be the set of optimal trajectories defined in* (3.1), *and let* $\mathcal{X}_c$ *be the output stabilizability space of* (A, B, C_2, D_2). *Then* $\mathcal{T}(b) \neq \emptyset$ *if and only if*

$$b = C_1 x + D_1 u, \tag{4.10}$$

where x *and* u *are such that* $Ax + Bu \in \mathcal{X}_c$.

If $\mathcal{B} = \ker R(\sigma)$ *is reduced with respect to* $R_2(\sigma)$, *then* $\mathcal{T}(b) \neq \emptyset$ *if and only if*

$$b \in \text{im } [C_1 \; D_1]. \tag{4.11}$$

Proof. The first part follows from Lemma 4.1. For the second part, notice that if $\mathcal{B} = \ker R(\sigma)$ is reduced with respect to $R_2(\sigma)$, (4.10) reduces to (4.11) since, in view of Proposition 4.2, $\mathcal{X}_c = \mathbb{R}^n$. □

Before presenting the solution of the problem, we make a brief remark which is suggested by the previous proposition. Assume that $\mathcal{B} = \ker R(\sigma)$ is not reduced with respect to $R_2(\sigma)$, and let (4.3) be a driving variable representation of $\bar{\Sigma}$. It is clear that its output stabilizability space $\mathcal{X}_c$ is A-invariant and it contains im B. Then by a suitable change of basis we can transform the representation (4.3) into the following:

$$\left\{ \begin{aligned} x(t+1) &= \bar{A}x(t) + \bar{B}u(t), \\ w(t) &= \bar{C}x(t) + \bar{D}u(t), \\ w_1(t) &= \bar{C}_1 x(t) + \bar{D}_1 u(t), \\ w_2(t) &= \bar{C}_2 x(t) + \bar{D}_2 u(t), \end{aligned} \right. \tag{4.12}$$

where

$$\bar{A} = \begin{bmatrix} A^{11} & A^{12} \\ 0 & A^{22} \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} B^1 \\ 0 \end{bmatrix},$$

$\bar{C} = [C^1\ C^2]$, $\bar{C}_1 = [C_1^1\ C_1^2]$, $\bar{C}_2 = [C_2^1\ C_2^2]$, $\bar{D} = D$, $\bar{D}_1 = D_1$, $\bar{D}_2 = D_2$, and where the output stabilizability space of the reduced driving variable representation

$$\left\{ \begin{aligned} x^1(t+1) &= A^{11}x^1(t) + B^1u(t), \\ w^1(t) &= C^1x^1(t) + Du(t), \\ w_1^1(t) &= C_1^1x^1(t) + D_1u(t), \\ w_2^1(t) &= C_2^1x^1(t) + D_2u(t), \end{aligned} \right. \tag{4.13}$$

is the whole state space $\mathbb{R}^{n_1}$. It is clear that the set of trajectories w^1 compatible with system (4.13) is a subset of $\mathcal{B}$. Actually it is not difficult to show that this set is exactly the reduction $\mathcal{B}_r$ of $\mathcal{B}$ with respect to $R_2(\sigma)$:

$$\mathcal{B}_r = \{w^1 \in (\mathbb{R}^q)^{\mathbb{Z}} : \exists x^1 \in (\mathbb{R}^{n_1})^{\mathbb{Z}}, u \in (\mathbb{R}^m)^{\mathbb{Z}} \text{ such that (4.13) is satisfied } \forall t \in \mathbb{Z}\}.$$

This observation furnishes a procedure to perform the reduction of $\mathcal{B}$ with respect to $R_2(\sigma)$ in terms of the state space representation.

4.3. Computation of optimal solutions. Next we furnish a parametrization of the set of optimal trajectories $\mathcal{T}(b)$, or, equivalently (Proposition 4.1), of the set of solutions of (4.4). Set $Q := C_2^TC_2$, $S := C_2^TD_2$, and $R := D_2^TD_2$. Problem (4.4) is thus equivalent to find

$$\begin{aligned} J^* = \min_{C_1x(0)+D_1u(0)=b} &\left\{ [x^T(0)u^T(0)] \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x(0) \\ u(0) \end{bmatrix} \right. \\ &\left. + \min_{u_{|[1,+\infty)}} \sum_{t=1}^{\infty} [x^T(t)u^T(t)] \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \right\}. \end{aligned} \tag{4.14}$$

As we have shown in section 3 or in subsection 4.2, we can assume that $\mathcal{B}$ is reduced with respect to $R_2(\sigma)$. This assumption, in view of Proposition 4.1, implies that, in the representation (4.3), the output stabilizability space of (A, B, C_2, D_2) is the whole space $\mathbb{R}^n$. In turn, this implies that the optimization problem

$$J_1^* = \min_{u_{|[1,+\infty)}} \sum_{t=1}^{\infty} [x^T(t)u^T(t)] \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \tag{4.15}$$

subject to

$$\left\{ \begin{aligned} x(t+1) &= Ax(t) + Bu(t), \\ x(1) &= x_1, \end{aligned} \right. \tag{4.16}$$

admits solutions u which render the cost function J_1 finite for all $x_1 \in \mathbb{R}^n$ [10]. Hence [10] the ARE[1]

$$M = Q + A^TMA - (S + A^TMB)(R + B^TMB)^\sharp(S^T + B^TMA) \tag{4.17}$$

admits a minimum positive semidefinite solution M_∞ which can be computed iterating the corresponding difference Riccati equation starting from $M(0) = 0$ (see [19, Chapter 4]). Moreover, minimization (4.15) can be performed in closed form yielding

[1] Given a matrix Δ, $\Delta^\sharp$ will denote the Moore–Penrose pseudoinverse of Δ.

$J_1 = x(1)^T M_\infty x(1)$, and the set of optimal state and input trajectories are given by the solution of the following linear system:

$$\left\{\begin{array}{l} x(t+1) = Fx(t) + Jv(t), \\ u(t) = Kx(t) + Gv(t), \\ x(1) = x_1, \end{array}\right. \tag{4.18}$$

where

$$K := -(R + B^T M_\infty B)^\sharp (S^T + B^T M_\infty A), \tag{4.19a}$$

$$G := I - (R + B^T M_\infty B)^\sharp (R + B^T M_\infty B), \tag{4.19b}$$

$$F := A + BK, \tag{4.19c}$$

$$J := BG, \tag{4.19d}$$

and v is an arbitrary trajectory parametrizing the set of optimal solutions.

We can now perform the first minimization of (4.14). This reduces (4.14) to the following form:

$$\begin{aligned} &\min_{C_1 x(0) + D_1 u(0) = b} [x^T(0) u^T(0)] \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x(0) \\ u(0) \end{bmatrix} + x(1)^T M_\infty x(1) \\ &= \min_{C_1 x(0) + D_1 u(0) = b} [x^T(0) u^T(0)] \begin{bmatrix} Q + A^T M_\infty A & S + A^T M_\infty B \\ S^T + B^T M_\infty A & R + B^T M_\infty B \end{bmatrix} \begin{bmatrix} x(0) \\ u(0) \end{bmatrix}. \end{aligned} \tag{4.20}$$

The latter is a static optimization problem which admits solutions if and only if $b \in \text{im } [C_1 \ D_1]$. In this case the set of solutions is given by [14, p. 235]

$$\begin{bmatrix} x(0)^* \\ u(0)^* \end{bmatrix} = \Xi b + Z\xi, \tag{4.21}$$

where ξ is an arbitrary vector and the matrices Ξ and Z are given by

$$\begin{aligned} \Xi &= (\Delta + H^T H)^\sharp H^T [H(\Delta + H^T H)^\sharp H^T]^\sharp, \\ Z &= I - (\Delta + H^T H)^\sharp (\Delta + H^T H), \end{aligned} \tag{4.22}$$

where $H := [C_1 \ D_1]$ and $\Delta := \begin{bmatrix} Q + A^T M_\infty A & S + A^T M_\infty B \\ S^T + B^T M_\infty A & R + B^T M_\infty B \end{bmatrix}$.

This static optimization furnishes the set of vectors $x(0)^*$ and $u(0)^*$, which minimize the cost function. From them, using the state update equation, we find the optimal $x(1)$. Initializing system (4.18) with this $x(1)$, we have a parametrization of all the optimal trajectories x, u. Using (4.3), we immediately get a parametrization of all the optimal trajectories w of $\mathcal{B}$.

In the case when $b \notin \text{im } [C_1 \ D_1]$, all the trajectories of the behavior $\mathcal{B}$ which satisfy the static constraint give rise to an infinite cost, or, equivalently, the problem is not solvable.

5. Concluding remarks. Since in the classical LQ optimization problem, the optimal control can be expressed as a linear state feedback, the question that naturally arises is whether or not this form of the solution remains valid in our setup. In the behavioral approach, feedback control is interpreted as an interconnection of systems which corresponds to an intersection of behaviors [24]. The previous issue reduces then to the following question. Does there exist an AR system (controller) $\Sigma_{ct} =$

$(\mathbb{Z}, \mathbb{R}^q, \mathcal{B}_{ct})$ such that $\mathcal{B}_{opt} = \mathcal{B} \cap \mathcal{B}_{ct}$, where $\mathcal{B}_{opt} = \{\mathcal{T}(b) : \ b \in \mathbb{R}^s\}$? This can occur only if the set $\mathcal{B}_{opt}$ is linear shift-invariant and complete. The following examples show that this is not true in general.

Example 1. Let $\Sigma = (\mathbb{Z}, \mathbb{R}, \mathcal{B})$ and $\mathcal{B} = \mathbb{R}^{\mathbb{Z}}$, and let $w_2 = w$ and $w_1 = (1+\sigma)w$. In this case the optimization problem is

$$J(w) = \min_{w(0)+w(1)=b} \sum_{t=0}^{\infty} w^2(t) = \min_{w(0)+w(1)=b} (w(0)^2 + w(1)^2). \tag{5.1}$$

The solution is easily $w(0) = w(1) = b/2$ and $J_{opt} = b^2/2$. The set of optimal trajectories, as b spans all the real axis $\mathbb{R}$, is

$$\mathcal{B}_{opt} = \{w \in \mathbb{R}^{\mathbb{Z}} : \ w(1) = w(0)\}. \tag{5.2}$$

Clearly this set is not shift-invariant.

One may suspect that the solution may be expressed as linear feedback at least in particular cases, for example, when the behavior $\mathcal{B}$ is autonomous or in correspondence of a classical LQ problem for descriptor systems. This is not the case as the following examples show.

Example 2. Let $\mathcal{B}$ be described by the following state model:

$$\left\{ \begin{array}{c} x(t+1) = Ax(t), \\ w(t) = x(t), \end{array} \right. \qquad t \in \mathbb{Z}, \tag{5.3}$$

where $A = \operatorname{diag}(1/2, 1/4)$. Let $w_1(t) = C_1 w(t)$ and $w_2(t) = C_2 w(t)$, where $C_1 = [1 \ 1]$ and $C_2 = \operatorname{diag}(\sqrt{3}/2, \sqrt{15}/4)$. It is easy to see that, as b spans all the real axis $\mathbb{R}$, the set of optimal trajectories is

$$\mathcal{B}_{opt} = \left\{ w(t) = A^t \begin{bmatrix} a \\ a \end{bmatrix} : \ a \in \mathbb{R} \right\}. \tag{5.4}$$

Again, it is immediate to check that this set is not shift-invariant.

Example 3. Let $\mathcal{B}$ be described by the following state model:

$$\left\{ \begin{array}{c} 0 = x(t) + u(t), \\ w(t) = x(t), \end{array} \right. \qquad t \in \mathbb{Z}. \tag{5.5}$$

Let $w_1(t) = w_2(t) = w(t)$. It is easy to see that, as b spans all the real axis $\mathbb{R}$, the set of optimal trajectories is

$$\mathcal{B}_{opt} = \{b\delta(t) : \ b \in \mathbb{R}\}. \tag{5.6}$$

Clearly, this set is not shift-invariant.

Observe, moreover, that in the formulation of a practical problem in the form (2.4), the dimension of the trajectories vector w may be very large. However, the reduction described in section 3.2 and the fact that the realization (4.3) is assumed to be state trim insure that the complexity of the resulting Riccati equation is in any case minimal. In other words, the complexity of the presented method is minimal.

The last observation we want to make concerns our technique for the solution of the optimization problem. One of the basic philosophies of the behavioral approach is to express the solution of a problem directly in terms of the data in which the problem is formulated. In our case the data of the problem are the polynomial matrices $R(\sigma)$,

$R_1(\sigma)$, $R_2(\sigma)$, and the vector b, and it would be desirable to describe the optimal trajectories in terms of these data. In this paper we have not been able to achieve this goal, since the solution we propose is obtained by passing through intermediate state space representations. The research of a solution realizing this goal remains a subject of our present investigation.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Control Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[2] D. J. Bender and A. J. Laub, *The linear-quadratic optimal regulator for descriptor systems: Discrete-time case*, Automatica J. IFAC, 23 (1987), pp. 71–85.

[3] J. Chen, *Acausal Stochastic Process Models: Structural Properties and Signal Processing*, Ph.D. thesis, Johns Hopkins University, Baltimore, MD, 1991.

[4] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, *Modeling stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[5] D. Cobb, *Descriptor variable systems and optimal state regulation*, IEEE Trans. Automat. Control, 28 (1983), pp. 601–611.

[6] J. B. Conway, *A Course in Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1990.

[7] J. Dugundji, *Topology*, Allyn and Bacon, Boston, MA, 1966.

[8] F. Fagnani and J. C. Willems, *Controllability of l^2-systems*, SIAM J. Control Optim., 30 (1992), pp. 1101–1125.

[9] F. R. Gantmacher, *Matrix Theory*, Chelsea, New York, 1959.

[10] T. Geerts, *The Algebraic Riccati Equation and Singular Optimal Control: The Discrete Time Case*, in Systems and Networks: Mathemathical Theory and Applications II, U. Helmke, R. Mennicken, and J. Saurer, eds., Akademie-Verlag, Berlin, 1994, pp. 129–134.

[11] T. T. Georgiou and M. C. Smith, *Graphs, causality, and stabilizability: Linear shift-invariant systems on $L_2[0,\infty)$*, Math. Control Signals Systems, 6 (1993), pp. 195–223.

[12] C. Heij, *Deterministic Identification of Dynamical Systems*, Lecture Notes in Control and Inform. Sci. 127, Springer-Verlag, Berlin, 1989.

[13] F. L. Lewis, *A survey of linear singular systems*, Circuits Systems Signal Process., 5 (1986), pp. 3–36.

[14] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Economics*, John Wiley and Sons, New York, 1988.

[15] Y. Murata, *Optimal Control Methods for Linear Discrete-Time Economic Systems*, Springer-Verlag, New York, 1982.

[16] J. W. Nieuwenhius, *Another look at linear-quadratic optimization problems over time*, Systems Control Lett., 25 (1995), pp. 89–97.

[17] B. Roorda and C. Heij, *Global total least square modeling of multivariable time series*, IEEE Trans. Automat. Control, 40 (1995), pp. 50–63.

[18] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[19] L. M. Silverman, *Discrete Riccati equations: Alternative algorithms, asymptotic properties, and system theory interpretations*, In Control and Dynamic Systems, C. T. Leondes, ed., Academic Press, New York, 1976, pp. 313–386.

[20] G. C. Walsh, R. Montgomery, and S. Sastry, *Optimal path planning on matrix Lie groups*, in Proceedings of the IEEE Conference on Decision and Control, Lake Buena Vista, FL, 1994, pp. 1258–1263.

[21] S. Weiland and A. Stoorvogel, *Rational representation of behaviors: Interconnectability and stabilizability*, Math. Control Signals Systems, 10 (1997), pp. 125–164.

[22] J. C. Willems, *From time series to linear systems, part* I: *Finite dimensional linear time invariant systems*, Automatica J. IFAC, 22 (1986), pp. 561–580.

[23] J. C. Willems, *Models for dynamics*, in Dynam. Report. Ser. Dynam. Syst. Appl. 2, Wiley, Chichester, UK, 1989, pp. 171–269.

[24] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[25] J. C. Willems, *LQ-control: A behavioral approach*, in Proceedings of the IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 3664–3668.

[26] J. C. Willems, *On interconnections, control, and feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 326–339.

[27] J. C. Willems and C. Heij, *l^2-systems and their scattering representation*, in Operator Theory and Systems, H. Bart, I. Gohberg, and M. A. Kaashock, eds., Birkhäuser Verlag, Basel, Boston, MA, 1986, pp. 443–448.

[28] J. C. Willems and H. L. Trentelman, *On quadratic differential forms*, SIAM J. Control Optim., 36 (1998), pp. 1703–1749.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 179–195

FROM THE MOSCO AND SLICE CONVERGENCES FOR CONVEX NORMAL INTEGRANDS TO THAT OF INTEGRAL FUNCTIONALS ON L^p SPACES*

JÉRÔME COUVREUX†

Abstract. Given a sequence of integrands $f_n : T \times X \to R$ $(n \geq 1)$ which converges in the sense of the slice-topology to an integrand f, $(T, \mathcal{A}, \mu)$ being a complete probability space and X a nonreflexive Banach space with separable dual, we show that the sequence of integral functionals $I(f_n) : u \to \int f_n(t, u(t))d\mu$ $(n \geq 1)$ associated to the (f_n) converges to $I(f) : u \to \int f(t, u(t))d\mu$ in the sense of the slice-topology on $L^p(X)$ and that the sequence of integral functionals associated to the conjugate integrands (f_n^*) converges to $I(f^*) : u \to \int f^*(t, u(t))d\mu$ on $L^q(X^*)$ (with $1 \leq p < +\infty$ and $p^{-1} + q^{-1} = 1$). This is an extension of some results which were shown to hold by Joly and de Thélin for Painlevé–Kuratowski convergence when X is finite dimensional and by Salvadori for Mosco convergence when X is reflexive. We also need to provide some criteria for functional convergence in the slice-topology, using the strong epigraphical upper limit.

Key words. integrands, integral functionals, measurable multifunctions, epi-convergence, slice-topology

AMS subject classifications. 49J45, 46N10, 28A20

PII. S0363012997322224

1. Introduction.

Epi-convergence was introduced by several authors for the study of minimization problems. It is known that the important fact about epi-convergence of a sequence of functions $f_n : X \to R$ $(n \geq 1)$ is that, in the presence of appropriate compactness assumptions (see [At]) this type of convergence entails nice properties of the infimal values $\inf\{f_n(x)/x \in X\}$ and of the set of minimizers *Argmin* f_n. Recall that, moreover, epi-convergence of such a sequence of functions is a special case of set convergence: it means Painlevé–Kuratowski convergence (see [Ku, At, Be1]) of the sequence of their epigraphs regarded as subsets of the product space $X \times R$. We shall be led to use natural extensions of Painlevé–Kuratowski convergence to infinite dimensional spaces, namely the Mosco type convergences. These convergences consist in the Painlevé–Kuratowski convergence of sequences of subsets with respect to, at the same time, two different topologies on a Banach space X: for example, the weak and strong topologies, which was the case originally considered by Mosco [Mo1]. Mosco convergence of epigraphs, induces a functional convergence which is often called Mosco convergence. When X is finite dimensional, Mosco and Painlevé–Kuratowski convergences coincide on $\mathcal{F}(X)$ (the space of all nonempty closed subsets of X). Further, a topology was introduced on $\mathcal{F}(X)$ by Beer [Be1], namely, the Mosco topology, which is compatible with the original Mosco convergence on $\mathcal{F}_c(X)$ (the space of all nonempty closed convex subsets of X). The most important properties of Mosco topology hold on $\mathcal{F}_c(X)$ when the Banach space X is reflexive and separable. Another topology was then introduced on $\mathcal{F}_c(X)$ (or on $\mathcal{F}(X)$), which entails a convergence stronger than Mosco's convergence and whose properties are interesting even without reflexivity: it is the slice-topology. This topology was first

*Received by the editors July 22, 1997; accepted for publication (in revised form) May 27, 1999; published electronically August 17, 2000.

http://www.siam.org/journals/sicon/39-1/32222.html

†Groupe Viabilité, Jeux, Contrôle, Université Paris IX Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France (couvreux@2bo.com).

considered by Joly and more recently by Sonntag and Zalinescu in a survey article (see [SZ]). The slice-topology was in fact intensively studied by Beer [Be2, Be3, Be4, Be6], who has shown its nice properties. Notice also that the term "slice" is due to Beer. As for Painlevé–Kuratowski and Mosco convergences, the slice-topology induces a useful functional convergence. The Painlevé–Kuratowski convergence, as well as the slice-topology, have interesting variational properties. More precisely, the polarity, regarded as a map from $\mathcal{F}_c(X)$ into the set of w*-closed convex subsets of X^*, is bicontinuous with respect to the slice-topology. This continuity property has already been shown to hold for the Painlevé–Kuratowski convergence by Wijsman [Wi] when X is finite dimensional and by Mosco when it is reflexive. In turn, for convex functions identified with epigraphs, this implies the continuity of conjugacy with respect to epi-convergence, which is most useful for approximating the solutions of optimization problems (see [At, Be6, Mo1]).

The purpose of this paper is to present some results about the Mosco convergence and the convergence in the slice-topology for integral functionals. Given an integrand f, that is an extended real-valued function defined on the product $T \times X$, where here $(T, \mathcal{A}, \mu)$ and X are, respectively, a complete probability space and a Banach space, we consider the integral functional defined on $L^p(X) 1 \leq p \leq +\infty$ (the space of all—equivalence classes—of Bochner integrable functions) by $I(f) : u \rightarrow \int f(t, u(t)) d\mu$. These integral functionals were introduced by Rockafellar in [Ro1, Ro2] and were studied in many papers (see, for instance, [Ro5, CV, JT, HU, Sa, CH, Cou]). The epi-convergence of integral functionals associated to a sequence of integrands $f_n : T \times X \rightarrow R$ $(n \geq 1)$ was first studied by Joly and de Thélin [JT] when X is finite dimensional. This result was extended for an infinite dimensional reflexive and separable Banach space X by Salvadori: this author shows (see [Sa, Theorem 3.1]) that when dealing with a sequence (f_n) which Mosco converges to an integrand f, the sequence $(I(f_n))$ of the integral functionals associated to the (f_n) Mosco converges to $I(f)$ on $L^1(X)$ and the sequence $(I(f_n^*))$, where f_n^* denotes the integrand conjugate of f_n, Mosco converges to $I(f^*)$ on $L^\infty(X^*)$. Our present objective is to provide a similar result where the Mosco functional convergence is replaced by the one deduced from the slice-topology. Considering a sequence of integrands $f_n : T \times X \rightarrow R$ $(n \geq 1)$ which converges in the sense of the slice-topology to an integrand f, where X is then a nonreflexive Banach space with separable dual, we show that the sequence of integral functionals associated to the (f_n) converges to $I(f)$ in the sense of the slice-topology on $L^p(X)$ and that the sequence of integral functionals associated with the conjugate integrands (f_n^*) converges to $I(f^*)$ on $L^q(X^*)$ (with $1 \leq p < +\infty$ and $p^{-1} + q^{-1} = 1$).

The paper is organized as follows. In section 2, we give the needed notations and preliminaries. In section 3 some criteria for functional convergence in the sense of the slice-topology are presented. The main results are expressed in section 4. Section 5 deals with some useful properties of integrands and integral functionals, and finally, section 6 contains the proofs of the main results.

2. Definitions and notations. Throughout this section X will be a separable Banach space whose norm is $\|\,.\,\|$. We denote by $\mathcal{P}(X)$ the space of all subsets of X, by $\mathcal{F}(X)$ (resp., $\mathcal{F}_c(X)$, $\mathcal{F}_{cb}(X)$) the family of all nonempty closed (resp., nonempty closed convex, nonempty closed convex and bounded) subsets of X for strong topology, and by $\mathcal{F}_{c^*}(X^*)$ (resp., $\mathcal{F}_{c^*b}(X^*)$) the family of nonempty closed convex (resp., nonempty closed convex and bounded) subsets of X^* for weak star topology. The indicator function associated to $A \in \mathcal{P}(X)$ is the function $\delta(.\,, A)$ such that $\delta(x, A) = 0$ if $x \in A$ and $\delta(x, A) = +\infty$ if $x \notin A$. We write also $\mathcal{B}(X)$ for the Borel tribe of X. The closed ball with center x and radius r is denoted $B(x, r)$.

The *distance function* of $F \in \mathcal{P}(X)$ and its *support function* are defined, respectively, by

$$d(x,F) = \inf\{\|x-y\|/y \in F\},\ x \in X \text{ and } s(x^*,F) = \sup\{\langle x^*,x\rangle/x \in F\},\ x^* \in X^*.$$

The *gap* between two subsets F and G of X is defined by

$$D(F,G) = \inf\{\|x-y\|/x \in F, y \in G\}.$$

Let T be an abstract space. A *multifunction* F is a map from T to $\mathcal{P}(X)$. The *domain* of F is the following subset of T:

$$dom\ F = \{t \in T/F(t) \neq \emptyset\}.$$

A *selection* of such a multifunction F is a function $s : T \to X$ such that for all $t \in dom\ F$, $s(t) \in F(t)$.

Suppose now that $(T,\mathcal{A})$ is a measurable space. A multifunction $F : T \to \mathcal{P}(X)$ is said to be *Effros-measurable* or *measurable* (see [Be2, CV, He2, Him]) if for every open set U in X

$$F^-U \in \mathcal{A}, \quad \text{where} \quad F^-U = \{t \in T/F(t) \cap U \neq \emptyset\}.$$

The *Effros sigma algebra* is denoted by $\mathcal{E}(\mathcal{F}(X))$. It is the smallest sigma algebra of subsets of $\mathcal{F}(X)$ containing all sets of the form $\{A \in \mathcal{F}(X)/A \cap V \neq \emptyset\}$, where V runs over the open subsets of X. Notice that when $\mathcal{F}(X)$ is equipped with some topology $\mathcal{T}$ (see also section 3), if the Effros tribe coincides with the Borel tribe associated with $\mathcal{T}$, then a multifunction $F : T \to \mathcal{F}(X)$ is Effros measurable if and only if it is Borel measurable.

A *measurable multifunction* defined on a probability space may be also called a *random set.* A *measurable selection* of a measurable multifunction F is a selection of F that is $(\mathcal{A},\mathcal{B}(X))$-measurable, and a *Castaing representation* of F is a sequence $s_n : T \to X$ $(n \geq 1)$ of measurable selections of F satisfying for all $t \in dom\ F$, $F(t) = cl\{s_n(t)/n \geq 1\}$. Let $f : X \to [-\infty,+\infty]$ be a function. Its *epigraph* is the following subset of $X \times R$:

$$\acute{e}pi\ f = \{(x,\alpha) \in X \times R/f(x) \leq \alpha\}.$$

Its *domain* is $dom\ f = \{x \in X/f(x) < +\infty\}$. The function f is said to be *proper* if for all $x \in X$, $f(x) > -\infty$ and if for some $x \in X$, $f(x) < +\infty$. The *conjugate* of the function f is defined by

$$f^*(x^*) = \sup\{\langle x^*,x\rangle - f(x)/x \in X\}, \quad x^* \in X^*.$$

Let T be again an abstract space. Every map $f : T \times X \to [-\infty,+\infty]$ is called an *integrand.* The *integrand conjugate* of f is $f^* : T \times X^* \to [-\infty,+\infty]$ defined by

$$f^*(t,x^*) = \sup\{\langle x^*,x\rangle - f(t,x)/x \in X\}, \quad (t,x^*) \in T \times X^*.$$

The *epigraphical multifunction* associated to an integrand f is $F : t \to epi\ f(t,.)$. Let $(T,\mathcal{A})$ be still a measurable space. If the epigraphical multifunction $F(t) = epi\ f(t,.)$ is measurable, f is called a *normal integrand.* Assume that a probability measure μ is given on $(T,\mathcal{A})$. If for almost every $t \in T$ the function $f(t,.)$ satisfies any property $(\mathcal{P})$ (for instance $f(t,.)$ is lower semicontinuous, convex, etc.), the integrand f is said to satisfy $(\mathcal{P})$. Let $(T, \mathcal{A}, \mu)$ be a complete probability space. A lower semicontinuous (lsc) integrand $f : T \times X \to [-\infty, +\infty]$ is *normal* if and only if it is $\mathcal{A} \otimes \mathcal{B}(X)$-measurable.

We write $L^\circ(T, \mathcal{A}, X)$, or $L^\circ(X)$ for the space of (equivalence classes of) $(\mathcal{A}, \mathcal{B}(X))$-measurable functions that are defined on T with values in X. We write also $L^p(T, \mathcal{A}, \mu, X)$ or $L^p(X)$ $(1 \leq p \leq +\infty)$ for the space of (classes of) functions f such that $t \to \|f(t)\|_X$ belongs to $L^p(T, \mathcal{A}, \mu, R)$. For every multifunction $F : T \to \mathcal{P}(X)$ and each p with $1 \leq p \leq +\infty$, let us set

$$S^p(F, \mathcal{A}) = \{f \in L^p(T, \mathcal{A}, \mu, X)/f(t) \in F(t) \text{ almost surely (a.s.)}.$$

The *integral function* associated to a normal and lsc integrand $f : T \times X \to [-\infty, +\infty]$ is the functional defined on $L^p(X)(1 \leq p \leq +\infty)$ by

$$I(f) : L^p(T, \mathcal{A}, \mu, X) \to [-\infty, +\infty],$$

$$u \to \int f(t, u(t))\, d\mu.$$

Let us consider now two functions f, $g : X \to [-\infty, +\infty]$. The *epi-sum* (or *infimal convolution*) of f and g is the function, denoted by $f +_e g$, defined by

$$(f +_e g)(x) = \inf\{f(w) + g(x - w)/w \in X\}, \quad x \in X.$$

An important tool in convex analysis is the method of *regularization* in which a given function f is approximated by the epi-sum of f with members of a parametrized family of smoothing kernels, such that when the parameter approaches zero from above, the epi-sums “converge” to the initial function. Regularization by kernels of the form $k\|\,.\,\|$ $(k > 0)$ is often called *Lipschitz* or *Baire–Wijsman regularization*, and regularization by kernels of the form $(k/2)\|\,.\,\|^2 (k > 0)$ is known as *Moreau–Yosida regularization* (see [Be6]). In what follows we need to use regularization by kernels of the form $\{k\|\,.\,\|^p/k > 0\}$ with $1 \leq p < +\infty$. So we set

$$f^{k,p}(x) = (f +_e k\|\,.\,\|^p)(x) = \inf\{f(u) + k\|u - x\|^p/u \in X\}.$$

Proposition 2.1. *Let $f : T \times X \to\ ]-\infty, +\infty]$ be a normal and lsc integrand. The integrand defined for each $t \in T$ by $f^{k,p}(t, .)$ is a normal and continuous integrand.*

As may be well known by the reader, the Lipschitz regularization with parameter k of some function f is the largest Lipschitz continuous function with constant k that f majorizes (see [Be6]). Considering the regularization by kernels of the form $\{k\|\,.\,\|^p/k > 0\}$ $1 \leq p < +\infty$, we have the following result.

Proposition 2.2. *Let $f : X \to\ ]-\infty, +\infty]$ be a proper function, and let k, p be such that $0 < k < +\infty$ and $1 \leq p < +\infty$. Then for each x and w in X, we have $f^{k,p}(x) \leq f^{2^{p-1}k,p}(w) + 2^{p-1}k\|x - w\|^p$.*

Proof. Let us fix x and w in X, α arbitrary such that $f^{2^{p-1}k,p}(w) < \alpha$ and choose $z \in X$ with $f(z) + 2^{p-1}k\|w - z\|^p < \alpha$. We compute

$$f^{k,p}(x) \leq f(z)+k\|x-z\|^p \leq f(z)+2^{p-1}k\|x-w\|^p+2^{p-1}k\|w-z\|^p \leq \alpha+2^{p-1}k\|x-w\|^p.$$

Since α was arbitrary, the wanted result holds. □

Throughout this paper $(T, \mathcal{A}, \mu)$ will be a complete probability space.

3. The topological concept: Definitions and criteria. Let (Y, ρ) be an abstract topological space and $f_n : Y \to [-\infty, +\infty]$ $(n \geq 1)$ a sequence of functions. We denote by $\mathcal{V}(x)$ the family of neighborhoods of $x \in Y$ relatively to topology ρ. The following functions are said to be, respectively, the *ρ-epigraphical* (*or ρ-epi*) *lower limit* and the *ρ-epigraphical* (*or ρ-epi*) *upper limit* of the sequence (f_n):

$$(\rho\text{-}li_e f_n)(x) = \sup\nolimits_{V \in \mathcal{V}(x)} \liminf\nolimits_{n \geq 1} \inf\nolimits_{u \in V} f_n(u),$$
$$(\rho\text{-}ls_e f_n)(x) = \sup\nolimits_{V \in \mathcal{V}(x)} \limsup\nolimits_{n \geq 1} \inf\nolimits_{u \in V} f_n(u).$$

When these two functions are equal, the common value is called the *ρ-epigraphical* (*or ρ-epi*) *limit* of (f_n), it is denoted by $\rho\text{-}\lim_e f_n$, and the sequence (f_n) is said to be *epi-convergent.* Moreover, let us notice (see [At, Theorem 1.13]) that if Y is a metric space whose strong topology is denoted by s, we have

$$(s\text{-}ls_e f_n)(x) = \min \{\limsup\nolimits_{n \geq 1} f_n(x_n)/(x_n) \ \textit{such that}\ x = s\text{-}\lim\, x_n\}. \tag{3.1}$$

The reader interested in studying more precisely epi-convergence and its nice properties should refer to [At].

If (C_n) is a sequence in $\mathcal{F}(Y)$, we put

$$\rho\text{-}li\ C_n = \{x \in Y/x = \rho\text{-}\lim x_n, x_n \in C_n, n \geq 1\},$$
$$\rho\text{-}ls\ C_n = \{x \in Y/x = \rho\text{-}\lim x_k, x_k \in C_{n(k)}, k \geq 1\}.$$

ρ-*li* C_n and ρ-*ls* C_n are, respectively, the *ρ-lower limit* and the *ρ-upper limit* of (C_n). We say that (C_n) *converges in the sense of Painlevé–Kuratowski* to C relatively to the topology ρ, which is denoted by $C = \rho\text{-}\lim C_n$, if the two following equalities are satisfied:

$$C = \rho\text{-}li\ C_n = \rho\text{-}ls\ C_n.$$

Let σ be another topology on Y. The sequence (C_n) is said to be *Mosco convergent* to C with respect to ρ and σ, and we write $C = M(\rho, \sigma)\text{-}\lim C_n$ if

$$C = \rho\text{-}\lim C_n = \sigma\text{-}\lim C_n.$$

In the special case where σ is finer than ρ, these last equalities are equivalent to the following inclusions: $\rho\text{-}ls\ C_n \subset C \subset \sigma\text{-}li\ C_n$. Let us now denote, respectively, by w and s the weak and strong topologies of a normed space X. Following Mosco [Mo1], a subset C is said to be the *Mosco limit* of a sequence (C_n) in $\mathcal{F}(X)$ if $C = M(w, s)\text{-}\lim C_n$, which may be denoted by $C = M\text{-}\lim C_n$. For more about Mosco convergence, the reader can refer to [Mo1] or [Be1, Be3]. A sequence of extended real functions (f_n) defined on Y is said to be *Mosco convergent* to a function f with respect to the topologies ρ and σ, which is denoted by $f = (M(\rho, \sigma), Y)\text{-}\lim f_n$ if *epi* $f = M(\rho, \sigma)\text{-}\lim\, epi\ f_n$. Moreover (see [Mo1] or [Be1]), we know that $f = (M(\rho, \sigma), Y)\text{-}\lim f_n$ (with $\rho \leq \sigma$) if and only if conditions (3.2) and (3.3) below are satisfied:

for every $x \in Y$, there exists a sequence (x_n) in Y with $x = \sigma\text{-}\lim x_n$ such that

$$\limsup f_n(x_n) \leq f(x), \tag{3.2}$$

for any subsequence (f_{n_k}) of (f_n), if $x \in Y$ is such that $x = \rho\text{-}\lim x_k$, then

$$\liminf f_{n_k}(x_k) \geq f(x). \tag{3.3}$$

Let us consider again a separable Banach space X. A *slice of a ball* is the intersection of a closed ball and a closed half space passing through the interior of the ball. The *slice-topology* on $\mathcal{F}(X)$ is the weak (or initial) topology determined by the family $\{D(B,.)/B$ is a nonempty slice of a ball$\}$. It is denoted by $\mathcal{T}_s$. About the properties of this topology the reader can refer to the intensive works of Beer [Be2, Be3, Be4 and Be6]. Further, from Beer [Be3, Theorem 5.2] we know that the slice-topology restricted to $\mathcal{F}_c(X)$ is the weak topology determined by the family $\{D(B,.)/B \in \mathcal{F}_{cb}(X)\}$. The *dual slice-topology* that is denoted by $\mathcal{T}_s^*$ is the topology on $\mathcal{F}_{c^*}(X^*)$ generated by the family $\{D(B,.)/B \in F_{c^*b}(X^*)\}$. A sequence of extended real functions (f_n) defined on X is said to be convergent to a function f in the sense of the slice-topology, which is denoted by $f = (\mathcal{T}_s, X)\text{-}\lim f_n$ if *epi* $f = \mathcal{T}_s\text{-}\lim$ *epi* f_n.

Wijsman's topology on $\mathcal{F}(X)$, which is denoted by $\mathcal{T}_w$, is the topology of pointwise convergence of distance functions. It was introduced in [Wi] when X is finite dimensional. A sequence (C_n) of closed sets is said to converge to C in the Wijsman topology if, for every $x \in X$, one has $d(x, C) = \lim d(x, C_n)$. For more about the nice properties of this topology, see [Wi, Be2, Be4, He2]. A sequence of extended real functions (f_n) defined on X is said to be convergent to a function f in the sense of the convergence deduced from Wijsman's topology, which is denoted by $f = (\mathcal{T}_w, X)\text{-}\lim f_n$ if *epi* $f = \mathcal{T}_w\text{-}\lim$ *epi* f_n. Notice that the slice-topology is the supremum of all Wijsman's topologies varying the norm on equivalent norms (see [Be4]).

X being still a Banach space, let us consider the space $\mathcal{F}(X)$ equipped with Wijsman's topology $\mathcal{T}_w$. Following Hess [He1, Theorem 3.1.1, P1.6], we know that the Effros tribe $\mathcal{E}(\mathcal{F}(X))$ coincides with the Borel tribe associated with Wijsman's topology $\mathcal{T}_w$ on $\mathcal{F}(X)$. That means that a multifunction $F : T \to \mathcal{F}(X)$ is Effros measurable if and only if it is Borel measurable. Moreover, Beer has shown that if X has a strongly separable dual X^*, then the Borel tribe associated with the slice-topology $\mathcal{T}_s$ on $\mathcal{F}(X)$ still coincides with the Effros tribe (see, for instance, Theorem 5.8 in [Be2]).

We ought to establish now some criteria for functional convergence in the slice-topology. They are given in Propositions 3.4 and 3.6, below. Proposition 3.4 is a result which is easily deduced from the fundamental statement recalled in Proposition 3.1 (Proposition 4.1 of [Be3]). Notice that a more precise result than Proposition 3.4 has been obtained by Attouch and Beer: it is Theorem 3.1 in [AB], where the given conditions need only to hold at points x in *dom* ∂f and x^* in Range ∂f. But for the convenience of the reader, in what follows, we choose to build easily Proposition 3.4 from Proposition 3.1.

In the following results, X will be a Banach space.

Proposition 3.1. *Let f, $f_n : X \to\ ]-\infty, +\infty]$ $(n \geq 1)$ be proper lsc and convex functions. Then, the following properties are equivalent.*

(a) $f = (\mathcal{T}_s, X)\text{-}\lim f_n$.

(b) *For every open subset W of X and each $\alpha \in R$, the condition epi $f \cap \{W \times]-\infty, \alpha[\} \neq \emptyset$ implies that epi $f_n \cap \{W \times]-\infty, \alpha[\} \neq \emptyset$ eventually, and for every open subset V of X^* and each $\alpha' \in R$, the condition epi $f^* \cap \{V \times]-\infty, \alpha'[\} \neq \emptyset$ implies that epi $f_n^* \cap \{V \times]-\infty, \alpha'[\} \neq \emptyset$ eventually.*

Lemma 3.2 below will be useful for proving Proposition 3.3.

Lemma 3.2. *Let (C_n) be a sequence of $\mathcal{F}(X)$. The following statements are equivalent:*

(a) *$C \subset$ s-li C_n;*

(b) *for each open subset V of X satisfying $C \cap V \neq \emptyset$, then $C_n \cap V \neq \emptyset$ eventually.*

Proof. The proof of (a) $\Longrightarrow$ (b) follows from the definition of the strong lower limit. For proving (b) $\Longrightarrow$ (a) let us consider one $x \in C$. Because of assumption (b) one has for each $k \geq 1$, $C_n \cap B(x, 1/k) \neq \emptyset$ eventually. Then, one can build a sequence $x_n \in C_n$ $(n \geq 1)$, which converges strongly to x; that is, $x \in$ *s-li* C_n. □

PROPOSITION 3.3. *Let f, $f_n : X \to \ ]-\infty, +\infty]$ $(n \geq 1)$ be proper lsc and convex functions. Then the following properties are equivalent.*

(a) *For each $x \in dom\ f$, $f(x) \geq (s\text{-}ls_e f_n)(x)$;*

(b) *For every open subset W of X and each $\alpha \in R$, the condition epi $f \cap \{W \times]-\infty, \alpha[\} \neq \emptyset$ implies that epi $f_n \cap \{W \times]-\infty, \alpha[\} \neq \emptyset$ eventually.*

Proof. First see that statement (b) is equivalent to the following: for every open subset W of $X \times R$ the condition *epi* $f \cap W \neq \emptyset$ implies that *epi* $f_n \cap W \neq \emptyset$ eventually. Thanks to Lemma 3.2, that means *epi* $f \subset$ *s-li epi* fn. As $s-li$ *epi* $f_n =$ *epi* $(s\text{-}ls_e f_n)$ (see [At, Theorem 1.36]), that provides the result. □

PROPOSITION 3.4. *Let f, $f_n : X \to \ ]-\infty, +\infty]$ $(n \geq 1)$ be proper lsc and convex functions. Then the following statements are equivalent:*

(a) *$f = (\mathcal{T}_s, X)$-lim f_n;*

(b) *for each $x \in dom\ f$, $f(x) \geq (s\text{-}ls_e f_n)(x)$, and for each $x^* \in dom\ f^*$, $f^*(x^*) \geq (s\text{-}ls_e f_n^*)(x^*)$.*

In Proposition 3.5 we ought to state some inequality between the strong epigraphical upper limit of some functions and their regularization by kernels of the form $\{k\|\,.\,\|^p/k > 0\}$ $1 \leq p < +\infty$, the result of which is vital for proving Theorem 4.1. A similar result was established in [He4] for Lipschitz regularization $(p = 1)$.

PROPOSITION 3.5. *Let $f_n : X \to \ ]-\infty, +\infty]$ $(n \geq 1)$ be a sequence of proper lsc and convex functions, and let p be such that $1 \leq p < +\infty$. Then the following properties hold.*

(1) *For every $x \in X$, $(s\text{-}ls_e f_n)(x) \geq \sup_{k>0} \limsup_{n\geq 1} f_n^{k,p}(x)$.*

(2) *If there exists $u_0 \in X$ and $(a, b) \in R^{+*} \times R$ such that for every $x \in X$, $f_n(x) \geq -a\|x - u_0\| - b$, then $(s\text{-}ls_e f_n)(x) \leq \sup_{k>0} \limsup_{n\geq 1} f_n^{k,p}(x)$.*

Proof. (1) For each $p \geq 1$, $k > 0$, $n \geq 1$, and every $x \in X$, one has $f_n(x) \geq f_n^{2^{p-1}k,p}(x)$ so that we can write, using the definition of $s\text{-}ls_e f_n$,

$$\begin{aligned}(s\text{-}ls_e f_n)(x) &= \sup_{\mu\geq 1} \limsup_{n\geq 1} \inf\{f_n(v)/v \in B(x, 1/\mu)\}\\ &\geq \sup_{\mu\geq 1} \limsup_{n\geq 1} \inf\{f_n^{2^{p-1}k,p}(v)/v \in B(x, 1/\mu)\}.\end{aligned}$$

Using Proposition 2.2, it follows that

$$\begin{aligned}&\sup_{\mu\geq 1} \limsup_{n\geq 1} \inf\{f_n(v)/v \in B(x, 1/\mu)\}\\ &\geq \ \sup_{\mu\geq 1} \limsup_{n\geq 1} \inf\{f_n^{k,p}(x) - 2^{p-1}k\|x - v\|^p/v \in B(x, 1/\mu)\}\\ &\geq \sup_{\mu\geq 1}\{\limsup_{n\geq 1}(f_n^{k,p}(x) - 2^{p-1}k/\mu^p)\} = \limsup_{n\geq 1} f_n^{k,p}(x).\end{aligned} \tag{3.4}$$

Since (3.4) is true for each $k > 0$, the wanted inequality holds.

(2) We ought to show that for every $x \in X$,

$$\sup_{\mu\geq 1} \limsup_{n\geq 1} \inf\{f_n(v)/v \in B(x, 1/\mu)\} \leq \sup_{k>0} \limsup_{n\geq 1} f_n^{k,p}(x).$$

Let us define for each $x \in X : w(x) = \sup_{k\geq 1} \limsup_{n\geq 1} f_n^{k,p}(x)$. So, we wish to prove that for all $x \in X$ $(s\text{-}ls_e f_n)(x) \leq w(x)$, the inequality of which is true if $w(x) = +\infty$. Else, let us fix $p \geq 1$ and $\alpha \in]0, 1[$. Consider for each $n \geq 1$, $k \geq 1$, and $p \geq 1$ the following statements:

$$f_n(v) + k\|x - v\|^p \geq \inf\{f_n(u) + k\|x - u\|^p/u \in X\} + \alpha \quad \text{for all } v \notin B(x, 1/\mu) \tag{3.5}$$

and

$$f_n^{k,p}(x) = \inf\{f_n(v) + k\|x - v\|^p/v \in B(x, 1/\mu)\}. \tag{3.6}$$

Clearly (3.5) implies (3.6). Further, fix $k \geq 1$. From the definition of w there exists $n_0(k) \geq 1$ such that for all $n \geq n_0(k)$, $w(x)+\alpha \geq f_n^{k,p}(x)$ so that obviously $w(x)+2\alpha \geq f_n^{k,p}(x)+\alpha$. Our goal is to obtain (3.5) for each $n \geq n_0(k)$ and every $k \geq k_0$, k_0 being chosen later in the proof. It is readily seen that this objective is reached if

$$f_n(v) + k\|x - v\|^p \geq w(x) + 2\alpha. \tag{3.7}$$

But we suppose that for all $v \in X$ and for all $n \geq 1$, $f_n(v) \geq -a\|v-u_0\|-b$. Then, (3.7) holds if $-a\|v-u_0\|-b+k\|x-v\|^p \geq w(x)+2\alpha \Leftrightarrow -a\|v-u_0\|+k\|x-v\|^p \geq w(x)+2\alpha+b$. But as $v \notin B(x, 1/\mu)$, we have $\|x - v\| \geq 1/\mu$ and $-\|u_0 - v\| \geq -\|x - v\| - \|x - u_0\|$. Therefore, (3.4) is implied by

$$k\|x - v\|^p - a\|x - v\| \geq w(x) + 2\alpha + b + a\|x - u_0\|. \tag{3.8}$$

Consequently, (3.8) is true if $k \geq k_0$, where

$$k_0 = \textit{Integer part of } \{\max\{\mu^{p-1}(\mu(w(x) + 2\alpha + b + a\|x - u_0\|) + a); a\mu^{p-1}\}\} + 1$$

Thus, if $k \geq k_0$, (3.5) is obtained (for each $n \geq n_0(k)$, and every $v \notin B(x, 1/\mu)$), which, as described above, entails (3.6):

$$f_n^{k,p}(x) = \inf\{f_n(v) + k\|x - v\|^p/v \in B(x, 1/\mu)\}.$$

Hence, it follows that for every $k \geq k_0$

$$w(x) + \alpha \geq \limsup_{n\geq 1} f_n^{k,p}(x) = \limsup_{n\geq 1} \inf\{f_n(v) + k\|x - v\|^p/v \in B(x, 1/\mu)\}.$$

This inequality being true for any $\mu \geq 1$ and any $\alpha \in]0, 1[$, then $s\text{-}ls_e f_n(x) \leq w(x)$, and then $s\text{-}ls_e f_n(x) \leq \sup_{k>0} \limsup_{n\geq 1} f_n^{k,p}(x)$, which finishes the proof. □

REMARK. *Similar inequalities could be stated with the strong epigraphical lower limit, but the proof is left to the reader.*

Proposition 3.4 together with Proposition 3.5 entails Proposition 3.6 below.

PROPOSITION 3.6. *Let f, $f_n : X \to\ ]-\infty, +\infty]$ $(n \geq 1)$ be proper lsc and convex functions, and let p be such that $1 \leq p < +\infty$.*

1. *If $f = (\mathcal{T}_s, X)\text{-}\lim f_n$, then, for each $x \in$ dom f, $f(x) = \sup_{k>0} \limsup_{n\geq 1} f_n^{k,p}(x)$ and for each $x^* \in$ dom f^*, $f^*(x^*) = \sup_{k>0} \limsup_{n\geq 1} f_n^{*k,p}(x^*)$.*
2. *Suppose that there exists $(a, b) \in R^{+*} \times R$ and $(a', b') \in R^{+*} \times R$ satisfying for each $n \geq 1$, each $x \in X$, and each $x^* \in X^*$, $f_n(x) \geq -a\|x\| - b$ and $f_n^*(x^*) \geq -a'\|x^*\| - b'$. If for each $x \in$ dom f, $f(x) \geq \sup_{k>0} \limsup_{n\geq 1} f_n^{k,p}(x)$ and for each $x^* \in$ dom f^*, $f^*(x^*) \geq \sup_{k>0} \limsup_{n\geq 1} f_n^{*k,p}(x^*)$, then $f = (\mathcal{T}_s, X)\text{-}\lim f_n$.*

Proof. As $f = (\mathcal{T}_s, X)\text{-}\lim f_n, f^* = (\mathcal{T}_x^*, X^*)\text{-}\lim f_n^*$ because of the continuity for the slice-topology of the Young–Fenchel transform (Theorem 4.2 of [Be3]), and we have $f = s\text{-}\lim_e f_n$ and $f^* = s\text{-}\lim_e f_n^*$. But epi-convergence of a sequence of functions at some points of the effective domain of the limit implies uniform linear minorization for the entire sequence (Proposition 3.7 of [Be7]). Then invoking Proposition 3.5 for f in X and for f^* in X^*, the wanted result is understood.

(2) is nothing else than Proposition 3.4 together with (2) of Proposition 3.5. □

4. Main results. This section provides the main results of this paper. From previous section, we know that a sequence of functions (f_n) converges to a function f in the sense of the slice-topology if $f \geq s\text{-}ls_e f_n$ and $f^* \geq s\text{-}ls_e f_n^*$. The purpose of Theorem 4.1 is to tell us that if such an inequality between a sequence of integrands and its strong epigraphical upper limit is almost surely satisfied, then it still holds for the integral functionals associated to them.

Theorem 4.1. *Let X be a separable Banach space, $f_n : T \times X \to \,]-\infty, +\infty\,]$ $(n \geq 1)$ a sequence of normal proper and lsc integrands, $f : T \times X \to \,]-\infty, +\infty]$ a proper integrand, p, q with $1 \leq p < +\infty$, and $p^{-1} + q^{-1} = 1$, satisfying the following assumptions:*

(a) *for almost every $t \in T$ and each $x \in dom\ f(t, .)$, $f(t, x) \geq s\text{-}ls_e f_n(t, x)$;*

(b) *there exists a sequence (u_n) in $L^p(X)$ and functions k and k_0 in $L^p(R)$, such that for each $n \geq 1$, $\|u_n(t)\| \leq k(t)$ and $f_n(t, u_n(t)) \leq k_0(t)$ a.s.;*

(c) *for almost every $t \in T$, each $x \in X$ and each $n \geq 1$, $f_n(t, x) \geq -h(t)\|x\| - h_0(t)$, where h and h_0 belong to $L^q(R)$ with $h(t) > 0$ a.s.*

Then for every function u in $L^p(X)$, $I(f)(u) \geq s\text{-}ls_e I(f_n)(u)$.

In the three following theorems, the Banach space X will have a separable dual X^*. In Theorem 4.2 we state that if a sequence of integrands converges in the sense of the slice-topology on X, then the sequence of associated integral functionals converges in the sense of the slice-topology on $L^p(X)$ $1 < p < +\infty$ and the sequence of conjugate integral functionals converges in the sense of the slice-topology on $L^q(X^*)$ $(p^{-1} + q^{-1} = 1)$.

Theorem 4.2. *Let $f_n : T \times X \to \,]-\infty, +\infty]$ $(n \geq 1)$ be a sequence of normal proper lsc and convex integrands, $f : T \times X \to \,]-\infty, +\infty]$ a proper integrand, and p, q such that $1 < p < +\infty$ and $p^{-1} + q^{-1} = 1$, satisfying the following assumptions:*

(a) *for almost every $t \in T$, $f(t, .) = (\mathcal{T}_s, X)\text{-}\lim f_n(t, .)$;*

(b) *there exists a sequence (u_n) in $L^p(X)$ and functions k and k_0 in $L^P(R)$, such that for each $n \geq 1$, $\|u_n(t)\| \leq k(t)$ and $f_n(t, u_n(t)) \leq k_0(t)$ a.s.;*

(c) *there exists a sequence (v_n) in $L^q(X^*)$ and functions h and h_0 in $L^q(R)$, such that for each $n \geq 1$, $\|v_n(t)\| \leq h(t)$ and $f_n^*(t, v_n(t)) \leq h_0(t)$ a.s.*

Then, $I(f) = (\mathcal{T}_s, L^p(X))\text{-}\lim I(f_n)$ and $I(f^*) = (\mathcal{T}_s^*, L^q(X^*))\text{-}\lim I(f_n^*)$.

The following theorem tells us about the special case of convergence for integral functionals defined on $L^1(X)$. Salvadori has shown (see [Sa, Theorem 3.1]) that if a sequence of integrands $f_n : T \times X \to \,]-\infty, +\infty]$ $(n \geq 1)$ Mosco converges to an integrand f, X being reflexive, then the sequence $(I(f_n))$ Mosco converges to $I(f)$ on $L^1(X)$ with respect to the weak and strong topologies, and the sequence $(I^*(f_n))$ Mosco converges to $I^*(f)$ on $L^\infty(X^*)$ with respect to the Mackey and weak star topologies. Theorem 4.3 below is an extension of Salvadori's property in the nonreflexive case; instead of the Mackey topology, the result deals with a stronger topology, namely, the topology of the uniform, convergence on the uniformly integrable and bounded subsets of $L^1(X)$ (for which we write $\rho(L^\infty(X^*), L^1(X))$ or ρ). Notice that because of Dunford's theorems (see [DU]), the Mackey topology and the topology $\rho(L^\infty(X^*), L^1(X))$ coincide when X is reflexive. Moreover, assuming an additional condition on integrands f and (f_n), this result of convergence holds with respect to the slice-topology.

Theorem 4.3. *Let $f_n : T \times X \to \,]-\infty, +\infty]$ $(n \geq 1)$ be a sequence of normal proper lsc and convex integrands, and let $f : T \times X \to \,]-\infty, +\infty]$ be a proper integrand, such that*

(a) *for almost every $t \in T$, $f(t, .) = (\mathcal{T}_s, X)\text{-}\lim f_n(t, .)$;*

(b) *there exists a sequence* (u_n) *in* $L^1(X)$ *and functions* k *and* k_0 *in* $L^1(R)$, *such that for each* $n \geq 1$, $\|u_n(t)\| \leq k(t)$ *and* $f_n(t, u_n(t)) \leq k_0(t)$ *a.s.;*

(c) *there exists a sequence* (v_n) *in* $L^\infty(X^*)$ *and functions* h *and* h_0 *in* $L^\infty(R)$, *such that for each* $n \geq 1$, $\|v_n(t)\| \leq h(t)$ *and* $f^*_n(t, v_n(t)) \leq h_0(t)$ *a.s.*

(1) *Then,* $I(f) = (M(w,s), L^1(X))\text{-}\lim I(f_n)$ *and* $I(f^*) = (M(w^*,\rho),\ L^\infty(X^*))\text{-}\lim I(f^*_n)$.

(2) *Moreover, if for almost every* $t \in T$, *each* $x \in X$ *and each* $n \geq 1$, $f_n(t,x) \geq f(t,x)$, *then* $I(f) = (\mathcal{T}_s, L^1(X))\text{-}\lim I(f_n)$ *and* $I(f^*) = (\mathcal{T}^*_s, L^\infty(X^*))\text{-}\lim I(f^*_n)$.

Theorem 4.4 is an application of previous theorems that gives some results about the slice-convergence of sets of integrable selections of randoms sets.

Theorem 4.4. *Let* F, $F_n : T \to F_c(X)$ $(n \geq 1)$ *be multifunctions and let* p *be with* $1 \leq p < +\infty$, *such that*

(a) *for almost every* $t \in T$, $F(t) = \mathcal{T}_s\text{-}\lim F_n(t)$;

(b) *the function* $t \to \sup\{d(0, F_n(t))/n \geq 1\}$ *belongs to* $L^p(R)$;

(1) $p = 1$. *If for almost every* $t \in T$, $F_n(t) \subset F(t)$ $(n \geq 1)$, *one has* $S^1(F, \mathcal{A}) = \mathcal{T}_s\text{-}\lim S^1(F_n, \mathcal{A})$;

(2) $1 < p < +\infty$. *One has* $S^p(F, \mathcal{A}) = \mathcal{T}_s\text{-}\lim S^p(F_n, \mathcal{A})$.

5. Some tools about integrands and integral functions. This section is devoted to some results concerning integrands and integral functionals that will be useful in the next section for proving the main results. First we wish some properties of integral functionals to be recalled. For this, the fundamental Proposition 5.1 below, which is due to Hiai and Umegaki [HU, Theorem 2.2], is useful as an important tool. It was used in particular in [Cou] for studying easily conjugacy for integral functionals; that is, if $f : T \times X \to\ ]-\infty, +\infty]$ is a normal and lsc integrand, where $(T, \mathcal{A}, \mu)$ is a complete probability space and X is a separable Banach space, then $I^*(f) = I(f^*)$. We shall use this result in what follows (it is an extension of some results of [Ro1, Ro4] where reflexivity was needed). Here we consider a Banach space with separable dual.

Proposition 5.1. *Let* $f : T \times X \to\ ]-\infty, +\infty]$ *be a normal and lsc integrand, let* $F : T \to \mathcal{F}(X)$ *be a measurable multifunction, and let* p *be such that* $1 \leq p < +\infty$. *If there exists* u_0 *in* $S^p(F)$ *satisfying* $I(f)(u_0) < +\infty$, *then* $\inf\{I(f)(u)/u \in S^p(F)\} = \int \inf\{f(t,x)/x \in F(t)\}\, d\mu$.

The preceding property entails Proposition 5.2 below, which tells us about regularization of integral functionals by kernels of the form $\{k\| \,.\, \|^p/k > 0\}$ $1 \leq p < +\infty$.

Proposition 5.2. *Let* $f : T \times X \to\ ]-\infty, +\infty]$ *be a normal and lsc integrand, and let* k, p *be such that* $k > 0$ *and* $1 \leq p < +\infty$. *If there exists* u_0 *in* $L^p(X)$ *satisfying* $I(f)(u_0) < +\infty$, *then* $I^{k,p}(f) = I(f^{k,p})$.

Proof. $I(f)$ being the integral functional associated to f, for each $u \in L^p(X)$ we compute

$$\begin{aligned} I^{k,p}(f)(u) &= \inf\{I(f)(v) + k(\|u - v\|_{L^p(X)})^p / v \in L^p(X)\} \\ &= \inf\left\{ \int (f(t, v(t)) + k(\|u(t) - v(t)\|_X)^p)\, d\mu / v \in L^p(X) \right\}. \end{aligned}$$

As there exists $u_0 \in L^p(X)$ satisfying $I(f)(u_0) < +\infty$, invoking Proposition 5.1, one has

$$I^{k,p}(f)(u) = \int \inf\{f(t,x) + k(\|u(t) - x\|_X)^p / x \in X\}\, d\mu = I(f^{k,p})(u). \qquad \square$$

In the main results, we suppose the integrands to satisfy some assumptions which were used in [JT] and [Sa]. But as in [Cou], it seems convenient to see them in Lemma 5.3 below as integrability conditions.

LEMMA 5.3. *Let f, $f_n : T \times X \to \ ]-\infty, +\infty]$ $(n \geq 1)$ be proper normal and lsc integrands satisfying for almost every $t \in T$, $f(t, .) = (\mathcal{T}_s, X)\text{-}\lim f_n(t, .)$, and let p be such that $1 \leq p < +\infty$. Let us consider the following statements.*

(a) *There exists a sequence (u_n) in $L^p(X)$ and functions k and k_0 in $L^p(R)$ such that for each $n \geq 1$, $\|u_n(t)\| \leq k(t)$ and $f_n(t, u_n(t)) \leq k_0(t)$ a.s.*

(b) *The function $t \to \sup\{d((0,0);\ epi\ f_n(t, .))/n \geq 1\}$ belongs to $L^p(R)$.*

(c) *There exists a function u_0 in $L^p(X)$ with $I(f)(u_0) < +\infty$.*

Then we have (a) $\Leftrightarrow$ (b) $\Rightarrow$ (c).

Proof. Let us define the measurable functions $t \to r_n(t) = d((0,0),\ epi\ f_n(t, .))$ $(n \geq 1)$ and $t \to r(t) = d((0,0),\ epi\ f(t, .))$.

(a) $\Rightarrow$ (b) As $r_n(t) = \inf\{\|x\| + f_n^+(t, x)/x \in X\}$ (see for instance [Cou]), we have

$$r_n(t) \leq \|u_n(t)\| + f_n^+(t, u_n(t)) \leq k(t) + k_0(t) \text{ a.s.} \quad n \geq 1.$$

(b) $\Rightarrow$ (a) The measurable selection theorem applied to the multifunctions $t \to epi\ f_n(t, .) \cap B((0,0); r_n(t) + 1)$ $(n \geq 1)$ shows the existence of a sequence $(u_n, \alpha_n) : T \to X \times R$ $(n \geq 1)$ such that

$$f_n(t, u_n(t)) \leq \alpha_n(t) \quad \text{and} \quad \|u_n(t)\| + |\alpha_n(t)| \leq r_n(t) + 1 \leq \sup_{n \geq 1} r_n(t) + 1 \text{ a.s.} \quad n \geq 1.$$

The wanted result can then be easily deduced.

(b) $\Rightarrow$ (c) Since the slice-topology is stronger than Wijsman's topology, we have $f(t, .) = (\mathcal{T}_w, X)\text{-}\lim f_n(t, .)$ a.s.; that is $r(t) = \lim_n r_n(t)$. Thus $r(t) \leq \sup_{n \geq 1} r_n(t)$ a.s. and $r \in L^p(R)$. Therefore, as in the proof of (b) $\Rightarrow$ (a), one can easily show the existence of $u_0 \in L^p(X)$ such that $f^+(\,.\,, u_0(.)) \in L^p(R)$. □

6. Proofs of the main results. It remains to provide the proofs of the main results, which is now possible according to previous sections.

Proof of Theorem 4.1. First, let us recall that on $\mathcal{F}_c(X)$ $\mathcal{T}_s$ is stronger than $\mathcal{T}_w$ and that the convergence induced by Wijsman's topology implies Painlevé–Kuratowski convergence. Therefore, see that because of assumptions on the f_n $(n \geq 1)$, f is a proper and lsc integrand. Second, Proposition 5.2 tells us that $I^{k,p}(f)(u) = I(f^{k,p})(u)$ and $I^{k,p}(f_n)(u) = I(f_n^{k,p})(u)$ for all $u \in L^p(X)$, $k > 0$, $n \geq 1$, $1 \leq p < +\infty$.

Using assumption (a) and Proposition 3.5, we have for almost every $t \in T$ and each $x \in X$

$$f(t, x) \geq s\text{-}ls_e f_n(t, x) \geq \sup_{k > 0} \limsup_{n \geq 1} f_n^{k,p}(t, x). \tag{6.1}$$

Then from (6.1) one has $f(t, u(t)) \geq s\text{-}ls_e f_n(t, u(t)) \geq \limsup_{n \geq 1} f_n^{k,p}(t, u(t))$ a.s. for all $u \in L^p(X)$ and for all $k > 0$. Further, we have $f_n^{k,p}(t, u(t)) \leq f_n(t, u_n(t)) + 2^{p-1}k\{\|u(t)\|^p + \|u_n(t)\|^p\} \leq k_0(t) + 2^{p-1}k\{\|u(t)\|^p + (k_0(t))^p\}$, and, invoking Fatou's lemma, we can write

$$\int f(t, u(t))\, d\mu \geq \int \limsup f_n^{k,p}(t, u(t))\, d\mu \geq \limsup \int f_n^{k,p}(t, u(t))\, d\mu. \tag{6.2}$$

Equation (6.2) being true for every $k > 0$, we have

$$I(f)(u) \geq \sup_{k > 0} \limsup_{n \geq 1} I^{k,p}(f_n)(u). \tag{6.3}$$

In order to apply the second property of Proposition 3.5 to the sequence $(I(f_n))$, we wish (6.4) to hold for every $n \geq 1$ and each $u \in L^p(X)$:

$$I(f_n)(u) \geq -h\|u\|_{L^p(X)} - h_0, \quad \text{where } (h, h_0) \in R^{+*} \times R. \tag{6.4}$$

Equation (6.4) can be deduced from assumption (c) and Holder's inequality. Hence, this provides the wanted result; that is, for each $u \in L^p(X)$

$$I(f)(u) \geq s\text{-}ls_e I(f_n)(u). \qquad \square$$

Proof of Theorem 4.2. On one hand, notice that assumption (c) implies that

$$f_n(t,x) \geq \langle x, u_n(t)\rangle - f_n^*(t, v_n(t)) \geq -\|x\|h(t) - h_0(t), \quad (t,x) \in T \times X, \quad n \geq 1,$$

and on the other hand that, similarly, assumption (b) implies that

$$f_n^*(t,x^*) \geq -\|x^*\|k(t) - k_0(t), \quad (t,x^*) \in T \times X^*, \quad n \geq 1.$$

Thus, Theorem 4.1 applied with the sequence $f_n : T \times X \to\ ]-\infty, +\infty]$ $(n \geq 1)$ tells us that for every $u \in L^p(X)$

$$I(f)(u) \geq s\text{-}ls_e I(f_n)(u), \tag{6.5}$$

and, since X^* is also a separable Banach space, that for every function $v \in L^q(X^*)$

$$I(f^*)(v) \geq s\text{-}ls_e I(f_n^*)(v). \tag{6.6}$$

Because of properties of conjugacy for integral functionals, we have $I^*(f) = I(f^*)$ and $I^*(f_n) = I(f_n^*)$ $(n \geq 1)$, so that (6.6) is $I^*(f)(v) \geq s\text{-}ls_e I^*(f_n)(v)$. The wanted result is then provided by (6.5), (6.6), and Proposition 3.4. $\square$

Let us now give the long proof of Theorem 4.3. It can be found in [Cou]. As in Theorem 3.1 in [Sa], we shall use for proving part (1) the well-known criteria for functional Mosco convergence recalled in (3.2) and (3.3); and for proving part (2), we shall use the criteria for functional convergence in the slice-topology expressed in section 3 (see also [Cou]). Moreover, we need to deal with Lemma 6.1, but for the convenience of the reader, we choose to express it after the following proof. Lemma 6.1 is a characterization of topology ρ in terms of measure convergence and may also be seen as an extension of a similar result for the Mackey topology due to Castaing and Grothendieck (see [Ca] and [Gr]).

Proof of Theorem 4.3. 1) Following criteria (3.2) and (3.3) for functional Mosco convergence, we know that $I(f) = (M(w,s),\, L^1(X))\text{-}\lim I(f_n)$ if and only if (6.7) and (6.8) below hold and that $I(f^*) = (M(w^*,\rho),\, L^\infty(X^*))\text{-}\lim\, I(f_n^*)$ if we have (6.9) and (6.10), where

$$\begin{gathered}\text{for each function } u \text{ in } L^1(X), \text{ there exists a sequence } (u_n) \text{ in } L^1(X)\\ \text{with } u = s\text{-}\lim u_n \text{ such that } \limsup I(f_n)(u_n) \leq I(f)(u);\end{gathered} \tag{6.7}$$

$$\begin{gathered}\text{for every sequence } (u_n) \text{ in } L^1(X) \text{ such that } u = \sigma(L^1(X), L^\infty(X^*))\text{-}\lim u_n,\\ \text{then } I(f)(u) \leq \liminf I(f_n)(u_n);\end{gathered} \tag{6.8}$$

$$\begin{gathered}\text{for each function } v \text{ in } L^\infty(X^*), \text{ there exists a sequence } (w_n) \text{ in } L^\infty(X^*) \text{ with}\\ v = \rho(L^\infty(X^*), L^1(X))\text{-}\lim w_n \text{ such that } \limsup I(f_n^*)(w_n) \leq I(f^*)(v);\end{gathered} \tag{6.9}$$

$$\begin{gathered}\text{for every sequence } (v_n) \text{ in } L^\infty(X^*) \text{ such that } v = \sigma(L^\infty(X^*), L^1(X))\text{-}\lim v_n,\\ \text{then } I(f^*)(v) \leq \liminf I(f_n^*)(v_n).\end{gathered} \tag{6.10}$$

Because of Theorem 4.1 and (3.1), (6.7) holds.

We ought to show now that (6.9) holds too. For this, let us consider $v \in L^\infty(X^*)$. If $I(f^*)(v) = +\infty$, then (6.9) is satisfied. Else, let us define the sequence of proper normal and lsc integrands (h_n) where

$$h_n : (t, x^*) \to [f_n^*(t, x^*) - f^*(t, v(t))]^+ \quad (n \geq 1) \tag{6.11}$$

and the sequence of nonempty closed convex and measurable multifunctions F_n with

$$\begin{aligned} F_n : t \to \{x^* \in B^*(t)/\|v(t) - x^*\| + h_n(t, x^*) \\ \leq \ \inf[\|v(t) - u^*\| + h_n(t, u^*)/u^* \in B^*(t)] + 1/n\} \quad (n \geq 1). \end{aligned} \tag{6.12}$$

B^* is the multifunction defined by $B^* : t \to \{x^* \in X^*/\|v(t) - x^*\| \leq R\}$, where $R = Ess\sup\{\|v(t)\| + |h(t)|/t \in T\}$ $(R > 0)$. For each $n \geq 1$, the multifunction F_n admits a measurable selection $w_n : T \to X^*$. The sequence (w_n) is bounded in $L^\infty(X^*)$ by construction. Moreover, $f(t, .) = (\mathcal{T}_s, X)\text{-}\lim f_n(t, .)$ a.s., and thanks to Proposition 3.4 and to (3.1), there exists for almost every $t \in T$ a sequence $(x_n^*(t))$ in X such that $v(t) = s\text{-}\lim x_n^*(t)$ and $f^*(t, v(t)) \geq \limsup f_n^*(t, x_n^*(t))$. Thus

$$\lim h_n(t, x_n^*(t)) = 0. \tag{6.13}$$

And for almost every $t \in T$, there exists $n_0(t)$ such that for every $n \geq n_0(t)$, $x_n^*(t) \in B^*(t)$. So that for each $n \geq n_0(t)$ we have

$$\|v(t) - w_n(t)\| + h_n(t, w_n(t)) \leq \|v(t) - x_n^*(t)\| + h_n(t, x_n^*(t)) + 1/n. \tag{6.14}$$

From (6.13) and (6.14), we can write for almost every $t \in T$

$$\lim[\|v(t) - w_n(t)\| + h_n(t, w_n(t))] = 0. \tag{6.15}$$

Because of definition of F_n and invoking the fact that $v_n(t) \in B^*(t)$ a.s., one can write $\|v(t) - w_n(t)\| + h_n(t, w_n(t)) \leq \|v(t) - v_n(t)\| + h_n(t, v_n(t)) + 1/n \leq R + [h_0(t) - f^*(t, v(t))]^+ + 1/n \leq R + h_0(t) + \|v(t)\|k(t) + k_0(t) + 1$. This last function is integrable. With (6.15), this implies that

$$\lim \int h_n(t, w_n(t))\, d\mu = 0. \tag{6.16}$$

Hence

$$\begin{aligned} &\lim \left[\int f_n^*(t, w_n(t))\, d\mu - \int f^*(t, v(t))\, d\mu\right]^+ = 0 \\ &\Rightarrow \limsup \int f_n^*(t, w_n(t))\, d\mu \leq \int f^*(t, v(t))\, d\mu, \end{aligned}$$

and that provides

$$\limsup I(f_n^*)(w_n) \leq I(f^*)(v). \tag{6.17}$$

Since the sequence (w_n) is bounded in $L^\infty(X^*)$ by construction and since $\lim \|v(t) - w_n(t)\| = 0$ a.s., which implies that (w_n) converges in measure to v, by virtue of Lemma 6.1 below, (w_n) converges to v with respect to topology ρ. This last result, together with (6.17), is nothing else than (6.9).

Now we wish to establish (6.10). So let us consider first some functions v and (v_n) in $L^\infty(X^*)$ $(n \geq 1)$ such that $v = \sigma(L^\infty(X^*),\, L^1(X))\text{-}\lim v_n$. Second, let us consider, fixed $u \in L^1(X)$, a sequence (u_n) in $L^1(X)$ obtained as in (6.7). These functions are then satisfying $I(f)(u) \geq \limsup I(f_n)(u_n)$. From the definition of the conjugate of $I(f_n)$, we can write for each $n \geq 1$

$$I^*(f_n)(v_n) = I(f_n^*)(v_n) \geq \langle v_n, u_n\rangle - I(f_n)(u_n). \tag{6.18}$$

As $v = \sigma(L^\infty(X^*),\, L^1(X))\text{-}\lim v_n$ and $u = s\text{-}\lim u_n$, we have $\lim\langle v_n, u_n\rangle = \langle v, u\rangle$. Hence

$$\begin{aligned}\liminf I(f_n^*)(v_n) &\geq \langle v, u\rangle - \limsup I(f_n)(u_n)\\ &\geq \langle v, u\rangle - I(f)(u),\end{aligned} \tag{6.19}$$

which proves that

$$\liminf I(f_n^*)(v_n) \geq I(f^*)(v). \tag{6.20}$$

Finally, we wish to prove (6.8). For this, consider first some functions u and (u_n) in $L^1(X)$ $(n \geq 1)$ such that $u = \sigma(L^1(X),\, L^\infty(X^*))\text{-}\lim u_n$. Second, fixed v in $L^\infty(X^*)$ as in (6.9), consider a bounded sequence (w_n) in $L^\infty(X^*)$ converging to v for topology ρ and such that $I(f^*)(v) \geq \limsup I(f_n^*)(w_n)$. By conjugacy it follows that

$$I(f_n)(u_n) \geq \langle u_n, w_n\rangle - I(f_n^*)(w_n). \tag{6.21}$$

It is not hard to see that $\lim\langle u_n, w_n\rangle = \langle u, v\rangle$. So we have

$$\liminf I(f_n)(u_n) \geq \langle u, v\rangle - \limsup\ I(f_n^*)(w_n) \geq \langle u, v\rangle - I(f^*)(v),$$

which implies that

$$\liminf I(f_n)(u_n) \geq I(f)(u),$$

that is (6.8).

2) In order to show that $I(f) = (\mathcal{T}_s, L^1(X))\text{-}\lim I(f_n)$, we want to establish that for each $u \in L^1(X)$ with $u \in \mathit{dom}\ I(f)$, $I(f)(u) \geq s\text{-}ls_e I(f_n)(u)$ and that for every $v \in L^\infty(X^*)$, with $v \in \mathit{dom}\ I(f^*)$, $I(f^*)(v) \geq s\text{-}ls_e I(f_n^*)(v)$. The first inequality holds thanks to Theorem 4.1. For the second one, let us consider the additional hypothesis; that is, for almost every $t \in T$ and each $x \in X$, $f_n(t,x) \geq f(t,x)$. Then, for each $x^* \in X^*$, $f_n^*(t,x^*) \leq f^*(t,x^*)$ and for every function $v \in L^\infty(X^*)$,

$$f_n^*(t, v(t)) \leq f^*(t, v(t)) \leq f^{*+}(t, v(t)) \text{ a.s.} \tag{6.22}$$

As $I(f^*)(v) < +\infty$, $f^{*+}(.\,, v(.)) \in L^1(\mathbf{R})$. And from (6.22)

$$\limsup f_n^*(t, v(t)) \leq f^*(t, v(t)). \tag{6.23}$$

Invoking (6.22), (6.23), and then Fatou's lemma, one has $\limsup \int f_n^*(t, v(t))\, d\mu \leq \int \limsup f_n^*(t, v(t))\, d\mu \leq \int f^*(t, v(t))\, d\mu$, and hence

$$\limsup I(f_n^*)(v) = \limsup I^*(f_n)(v) \leq I^*(f)(v). \tag{6.24}$$

Let us deal now with the Lipschitz approximation of the functions $I(f_n^*)$ $(n \geq 1)$. For each $v \in L^\infty(X^*)$ and every $k > 0$, one can write

$$I(f_n^*)(v) = I^*(f_n)(v) \geq I^{*k,1}(f_n)(v). \tag{6.25}$$

From (6.24) and (6.25), it follows that $I^*(f)(v) \geq \limsup I^{*k,1}(f_n)(v)$ for each $k > 0$, which entails

$$I^*(f)(v) \geq \sup_{k>0} \limsup_{n\geq 1} I^{*k,1}(f_n)(v). \tag{6.26}$$

Hypothesis (b) easily implies the existence of some real k and k_0 ($k_0 > 0$) such that $I^*(f_n) \geq -k\|v\|_{L^\infty} - k_0$. Thus, part (2) of Proposition 3.5 can be used, which tells us that $s\text{-}ls_e I^*(f_n)(v) \leq \sup_{k>0} \limsup_{n\geq 1} I^{*k,1}(f_n)(v)$, and then that

$$I^*(f)(v) \geq s\text{-}ls_e I^*(f_n)(v).$$

As $I(f)(v) \geq s\text{-}ls_e I(f_n)(v)$ and $I^*(f)(v) \geq s\text{-}ls_e I^*(f_n)(v)$, from Proposition 3.4, we finally have $I(f) = (\mathcal{T}_s, L^1(X))\text{-}\lim I(f_n)$ and then $I(f^*) = (\mathcal{T}_s^*, L^\infty(X^*))\text{-}\lim I(f_n^*)$. □

Here now is Lemma 6.1.

Lemma 6.1. *Let $(T, \mathcal{A}, \mu)$ be a finite probability space, and let X be a Banach space with strongly separable dual. Let g, $g_n : T \to X^*$ ($n \geq 1$) be some functions such that first the sequence (g_n) is bounded and second the sequence (g_n) converges in measure to g. Then, the sequence (g_n) converges to g with respect to the topology of the uniform convergence on the uniformly integrable and bounded subsets of $L^1(X)$ (which is denoted by ρ).*

Proof. We can only consider the case where $g = 0$. Let Γ be a uniformly integrable subset of $L^1(X)$. For each function $u \in \Gamma$, every $a \in R^+$, and each $n \geq 1$, we write

$$\langle g_n, u\rangle = \int \langle g_n(t), u(t)\rangle d\mu = I + J,$$

where $I = \int_{\{t/\|u(t)\|\geq a\}} \langle g_n(t),\, u(t)\rangle\, d\mu$ and $J = \int_{\{t/\|u(t)\|<a\}} \langle g_n(t),\, u(t)\rangle\, d\mu$. On one hand, see that as by hypothesis there exists $M > 0$ such that $\sup\{\|g_n\|_{L^\infty}/n \geq 1\} \leq M$, one has

$$|I| \leq \int_{\{t/\|u(t)\|\geq a\}} \|g_n(t)\|_{X^*} \|u(t)\|_X \, d\mu \leq M \int_{\{t/\|u(t)\|\geq a\}} \|u(t)\|_X \, d\mu.$$

And $\lim_{a\to+\infty} \int_{\{t/\|u(t)\|\geq a\}} \|u(t)\|_X \, d\mu = 0$ uniformly in u. Then, fixed $\varepsilon > 0$, a can be chosen (we denote it by $a(\varepsilon)$) such that $|I| < \varepsilon/2$ for every $u \in \Gamma$.

On the other hand:

$$\begin{aligned} |J| &\leq \int_{\{t/\|u(t)\|<a(\varepsilon)\}} \langle g_n(t), u(t)\rangle \, d\mu \\ &\leq a(\varepsilon) \int_{\{t/\|u(t)\|<a(\varepsilon)\}} \|g_n(t)\|_{X^*} \, d\mu \leq a(\varepsilon) \int_T \|g_n(t)\|_{X^*} d\mu. \end{aligned}$$

It is easily seen that the sequence $(\|g_n\|_{X^*})$ converges in measure to 0. As $\sup\{\|g_n\|_{L^\infty}/n \geq 1\} \leq M$, we can find n_0 such that $n \geq n_0$, $|J| \leq \varepsilon/2$. Thus, for $n \geq n_0$ and $u \in \Gamma$, $\langle g_n, u\rangle < \varepsilon$ which entails

$$\lim_n \sup \left\{ \int_T \langle g_n(t), u(t)\rangle \, d\mu/u \in \Gamma \right\} = 0. \qquad \square$$

Proof of Theorem 4.4. The proofs of (1) and (2) being similar, we only give the second one (see also [Cou]). Let us define the integrands $\delta_n : (t, x) \to \delta(x, F_n(t))$

$(n \geq 1)$ and $\delta : (t,x) \to \delta(x, F(t))$. According to Beer [Be3, Theorem 3.1], assumption (a) implies that for almost every $t \in T$

$$\delta(\,.\,, F(t)) = (\mathcal{T}_s, X)\text{-}\lim \delta(\,.\,, F_n(t)). \tag{6.27}$$

Moreover, applying the measurable selection theorem with the random sets $t \to F_n(t) \cap B(0; d(0, F_n(t))$ $(n \geq 1)$, we show the existence of a sequence $u_n \in S^p(F_n, \mathcal{A})$ $(n \geq 1)$ with $\|u_n(t)\| \leq k(t)$ a.s. where $k \in L^p(R)$. Then

$$\delta(u_n(t), F_n(t)) = 0 \text{ a.s.}, \quad n \geq 1. \tag{6.28}$$

Moreover, the support function of the random set F_n $(n \geq 1)$, which is the conjugate of the indicator function δ_n, satisfies

$$s(0, F_n(t)) = 0 \text{ a.s.}, \quad n \geq 1. \tag{6.29}$$

As (6.28) and (6.29) are nothing else than, respectively, assumptions (b) and (c) of Theorem 4.2 with $f_n = \delta_n$ and $f_n^* = s(\,.\,, F_n)$, one has $I(\delta) = (\mathcal{T}_s, L^p(X))\text{-}\lim I(\delta_n)$. But since for each function u in $L^p(X)$ $I(\delta)(u) = 0$ (resp., $I(\delta_n)(u) = 0$, $n \geq 1$) if and only if $u(t) \in F(t)$ a.s. (resp., $u(t) \in F_n(t)$ a.s. $n \geq 1$), we have $\delta(\,.\,, S^p(F, \mathcal{A})) = (\mathcal{T}_s, L^p(X))\text{-}\lim\ \delta(\,.\,, S^p(F_n, \mathcal{A}))$. Thus $S^p(F, \mathcal{A}) = \mathcal{T}_s\text{-}\lim S^p(F_n, \mathcal{A})$. □

REFERENCES

[At] H. Attouch, *Variational Convergence for Functions and Operators*, Applicable Math. Series, Pittman (Advanced Publishing Program), Boston, MA, London, UK, 1984.

[AB] H. Attouch and G. Beer, *On the convergence of subdifferentials of convex functions*, Arch. Math. (Basel), 60 (1993), pp. 389–400.

[B] A. Bourass, *Thesis*, Université de Perpignan, Perpignan, France, 1983.

[BV] A. Bourass and M. Valadier, *Condition de croissance associé à l'inclusion des sections*, Séminaire d'Analyse Variationnelle et Applications en Mécanique, Automatique et Contrôle (AVAMAC), n°3, Université de Perpignan, Perpignan, France, 1984.

[Be1] G. Beer, *On Mosco convergence of convex sets*, Bull. Austral. Math. Soc., 38 (1988), pp. 239–253.

[Be2] G. Beer, *Topologies on closed and closed convex sets and Effrös-measurability of set valued functions*, Séminaire d'Analyse Convexe, Montpellier, n°2, Université de Montpellier 2, Montpellier, France, 1991, pp. 2.1–2.44.

[Be3] G. Beer, *The slice topology, a viable alternative to Mosco convergence in non reflexive spaces*, Séminaire d'Analyse Convexe, Montpellier, n°3, Université de Montpellier 2, Montpellier, France, 1991, pp. 3.1–3.33.

[Be4] G. Beer, *Wijsman convergence of convex sets under renorming*, Nonlinear Anal., 22 (1994), pp. 207–216.

[Be5] G. Beer, *Conjugate convex functions and the epi-distance topology*, Proc. Amer. Math. Soc., 108 (1990), pp. 117–126.

[Be6] G. Beer, *Topologies on Closed and Closed Convex Sets*, Math. Appl. 268, Kluwer Academic Publishers, Dordrecht, 1993.

[Be7] G. Beer, *Efficiency and the uniform linear minorization of convex functions*, Monatsh. Math., 115 (1993), pp. 281–290.

[BB] G. Beer and J. Borwein, *Mosco convergence and reflexivity*, Proc. Amer. Math. Soc., 109 (1990), pp. 427–436.

[Ca] C. Castaing, *Topologie de la convergence uniforme sur les parties uniformément intégrables de $L^1(E)$ et théorèmes de compacité faible dans certains espaces du type Kothe-Orlicz*, Séminaire d'Analyse Convexe, Montpellier, n°5, Université de Montpellier 2, Montpellier, France, 1980.

[CV] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer, Berlin, New York, 1977.

[Co] B. Cornet, *Topologies sur les fermés d'un espace métrique*, Cahier de mathématiques de la décision, Université Paris IX, France, 1973.

[Cou] J. Couvreux, *Etude de problèmes de convergence de fonctionnelles intégrales et d'espérances conditionnelles multivoques*, Ph.D. Thesis, Université Paris IX, France, 1995.

[CH] J. Couvreux and C. Hess, *Mosco approximation of integrands and integral functionals*, J. Optim. Theory Appl., 90 (1996), pp. 335–356.

[DM] C. Dellacherie and P. A. Meyer, *Probabilités et potentiels*, Hermann, Paris, 1966.

[DU] J. Diestel and J. J. Uhl, *Vector Measures*, Math. Surveys 15, Amer. Math. Soc., Providence, RI, 1977.

[F] A. Fougères, Comparaison de fonctionnelles intégrales sur les sélections d'une multiapplication mesurable, Séminaire d'Analyse Convexe, Montpellier, n°9, Université de Montpellier 2, Montpellier, France, 1982.

[FT] A. Fougères and A. Truffert, *Regularisation sci et gamma-convergence: approximation inf-convolutives associées à un référentiel*, Ann. Mat. Pura Appl. (4), 152 (1988), pp. 21–51.

[Gr] A. Grothendieck, *Espaces vectoriels topologiques*, Instituto de Matematica da Universidade de Sao Paulo, Sao Paulo, Brazil, 1954.

[He1] C. Hess, *Conditions d' optimalité pour des fonctionnelles intégrales convexes sur les espaces $L^p(E)$*, Cahier de mathématiques de la décision 8203, Université Paris IX, France, 1981.

[He2] C. Hess, *Thèse d'Etat*, Université des sciences et techniques du Languedoc, Languedoc, France, 1986.

[He3] C. Hess, *Measurability and integrability of the weak upper limit of a sequence of multifunctions*, J. Math. Anal. Appl., 153 (1990), pp. 226–249.

[He4] C. Hess, *Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator*, Cahier de mathématiques de la décision 9121, Université Paris IX, France, 1991, and Ann. Statist., 24 (1996), pp. 1298–1315.

[HU] F. Hiai and H. Umegaki, *Integrals, conditional expectations and martingales of multivalued functions*, J. Multivariate Anal., 7 (1977), pp. 149–182.

[Him] C. J. Himmelberg, *Measurable relations*, Fund. Math., 87 (1975), pp. 52–73.

[JT] J. L. Joly and F. de Thélin, *Convergence of convex integrals in L^p spaces*, J. Math. Anal. Appl., 54 (1976), pp. 230–244.

[KA] L. Kantorovitch and G. Akilov, *Analyse fonctionnelle*, Editions de Moscou, "Mir," Moscow, 1977.

[Ku] K. Kuratowski, *Topology. Vol.* I, Academic Press, New York, London, PWN Polish Scientific, Warsaw, 1966.

[KRN] K. Kuratowski and C. Ryll-Nardzewski, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 397–403.

[Mo1] U. Mosco, *Convergence of convex sets and solutions of variational inequalities*, Adv. Math., 3 (1969), pp. 510–585.

[Mo2] U. Mosco, *On the continuity of the Young–Fenchel transform*, J. Math. Anal. Appl., 35 (1971), pp. 518–535.

[Ne] J. Neveu, *Martingales à temps discret*, Masson et Cie, Paris, 1972.

[Ro1] R. T. Rockafellar, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.

[Ro2] R. T. Rockafellar, *Integrals which are convex functionals* II, Pacific J. Math., 39 (1971), pp. 439–469.

[Ro3] R. T. Rockafellar, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

[Ro4] R. T. Rockafellar, *Convex Integral Functionals and Duality. Contribution to Non Linear Functional Analysis*, Academic Press, New York, 1971, pp. 215–236.

[Ro5] R. T. Rockafellar, *Integral functionals, normal integrands and measurable selectors*, in Lecture Notes in Math. 543, Springer, Berlin, 1976, pp. 157–207.

[SW] G. Salinetti and R. Wets, *On the relation between two types of convergences for convex functions*, J. Math. Anal. Appl., 60 (1977), pp. 211–226.

[Sa] A. Salvadori, *On the M-convergence for integral functionals on $L^p(X)$*, Atti. Sem. Mat. Fis. Univ. Modena, 33 (1984), pp. 137–154.

[SZ] Y. Sonntag and C. Zalinescu, *Set convergences. An attempt of classification*, in Proceedings of the International Conference on Differential Equations and Control Theory, Iasi, Romania, 1990, and Trans. Amer. Math. Soc., 340 (1993), pp. 199–226.

[Wi] R. Wijsman, *Convergences of sequences of convex sets, cones and functions* III, Trans. Amer. Math. Soc., 123 (1966), pp. 32–45.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 196–207

DELAY-INDUCED INSTABILITIES IN GYROSCOPIC SYSTEMS[*]

PEDRO FREITAS[†]

Abstract. It is shown that stable linear gyroscopic systems of the form $Mx'' + Tx' + Kx = 0$ $(M > 0)$ always become unstable when an arbitrarily small delay is introduced in the gyroscopic term. In the case where K is negative definite, then the system will be unstable for all positive delays. On the other hand, examples are given showing that some of these systems may actually become asymptotically stable for larger values of the delay parameter.

Key words. gyroscopic forces, time delays, vibrating systems

AMS subject classifications. 34K, 70J25, 93D15

PII. S0363012999357938

1. Introduction.

It is known that certain infinite-dimensional Hamiltonian systems that have been stabilized via distributed or boundary damping are not robust with respect to small time delays which might be present in the feedback stabilization mechanism (see, for example, [DLP, DY, RT]). In these cases, the effect of the introduction of the small delay is to perturb the high frequencies of the (asymptotically) stable system, giving rise to periodic or even exponentially growing solutions. Clearly this phenomenon is made possible by the fact that the spectrum of the system without delay has an infinite number of eigenvalues on an unbounded strip parallel to the imaginary axis.

Another situation where the presence of small delays might be expected to destabilize an otherwise stable system is when there are eigenvalues lying on the imaginary axis. An example of this type is provided by mechanical systems with gyroscopic forces which are modelled by systems of differential equations of the form

$$Mx''(t) + Tx'(t) + Kx(t) = 0. \tag{1.1}$$

Here M and K are real symmetric $n \times n$ matrices usually referred to as the mass and stiffness terms, corresponding to inertial and potential forces, respectively. The mass matrix will be assumed to be positive definite. The matrix T is a real $n \times n$ skew-symmetric matrix which represents the gyroscopic forces.

In order to study the stability of the trivial (zero) solution of (1.1), we look for solutions of the form $x = e^{\lambda t}u$, leading to the eigenvalue problem

$$\lambda^2 Mu + \lambda Tu + Ku = 0.$$

Throughout the paper we consider the usual definitions of different types of stability that may be found in the literature (for instance, in Chapter 5 of [HL]). The trivial solution of a system of the form (1.1) will then be stable if and only if all eigenvalues lie on the imaginary axis and are semisimple, that is, their algebraic and geometric multiplicities are equal. This is, of course, related to the Hamiltonian symmetry of the corresponding spectrum. Such a system will always be stable independent of T

[*] Received by the editors June 15, 1999; accepted for publication (in revised form) April 11, 2000; published electronically August 17, 2000. This research was partially supported by the Fundação para a Ciência e Tecnologia, Portugal.

http://www.siam.org/journals/sicon/39-1/35793.html

[†] Departamento de Matemática, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal (pfreitas@math.ist.utl.pt).

when K is positive definite, in which case all eigenvalues are semisimple [L]. If K is allowed to have both negative and positive eigenvalues, the system may then, under certain conditions, be stabilized by using a suitable matrix T. This has been known for a long time [TT] and, in particular, it is known that a necessary condition for it to be possible is that the number of negative eigenvalues of K be even. Recently there have been several papers giving conditions on the matrix T so that the overall system is stabilized when K is negative definite (see, for instance, [BLM, HKLP]).

In this note, we show that when a small delay parameter is introduced in (1.1), this will force some of the eigenvalues into the right half of the complex plane independently of the gyroscopic, mass, and stiffness matrices, thus making the trivial solution unstable. Under certain quite general conditions given below, we shall see that the presence of small delays in the gyroscopic term always destabilizes the mode of the original system corresponding to the highest frequency. We conjecture that the lowest frequency mode is actually stabilized, but have only been able to prove this under special circumstances (K either positive or negative definite). As mentioned above, this destabilizing effect is not unexpected as all the eigenvalues of the original system are on the imaginary axis. However, it is not clear a priori whether this will always be the case or not, irrespective of the matrices M, T, and K. Besides, this seems to be the first study of a finite-dimensional system where the associated transfer function is proper and which always becomes unstable when arbitrarily small delays are introduced. For some different approaches and situations, see, for instance, [BCD, GS]. Note that in all of the examples given above the destabilizing effect relies on the fact that in the absence of delay the underlying system is already infinite-dimensional.

Under some additional restrictions ($K < 0$), we are able to show that instability then persists for all values of the delay. More surprisingly, we also show that there are situations where the system will actually become asymptotically stable for larger values of the delay parameter—note that although the original system is stable, it is not asymptotically stable. This stabilizing effect of the delay has been identified in other systems. See, for instance, [CG], where stability switching sequences for some second-order scalar differential equations were studied. When n equals two, we give a full description of these sequences in terms of a set of four positive real numbers. These are the only possible crossing points of eigenvalues on the imaginary axis in this case. We then show that although the system might become stable for some values of τ, it ultimately becomes unstable. In fact, we prove that once there are more than four unstable eigenvalues the system never stabilizes again for large values of τ.

We begin by indicating the overall hypotheses which are to be assumed throughout the paper and by presenting the main results in section 2. We then go on to study some properties of the spectrum associated with this problem, with particular focus on the derivatives with respect to the delay at purely imaginary eigenvalues. This is done in section 3. Section 4 deals with the destabilizing effect of delays, and in section 5 we show that under some circumstances it is possible for the system to become asymptotically stable for larger values of the delay. This is done by means of a full description of the 2×2 case. Finally, in section 6 we briefly discuss the results obtained.

2. Preliminaries and main results.

We shall consider only the case where M is positive definite, which implies that system (1.1) can be brought into the form

$$y''(t) + Gy'(t) + Cy(t) = 0, \tag{2.1}$$

by means of the change of variables $x = M^{-1/2}y$. Here $G = M^{-1/2}TM^{-1/2}$ and $C = M^{-1/2}KM^{-1/2}$ are real $n \times n$ skew-symmetric and symmetric matrices, respectively. We now introduce a delay in the gyroscopic term in (2.1) to obtain

$$y''(t) + Gy'(t-\tau) + Cy(t) = 0. \tag{2.2}$$

Throughout the paper we shall make the following assumptions on (2.2).

H1. In the absence of delay ($\tau = 0$) all eigenvalues are on the imaginary axis and are simple.

H2. C is nonsingular.

H3. G and C do not have any common invariant spaces.

The reason for assuming H1 has to do with the fact that if the original system has eigenvalues with positive real parts, then it will always remain unstable for small delays and there is nothing to prove. As mentioned in the introduction this will impose certain conditions on the matrices G and C. In order to have stability, it is also necessary that the eigenvalues are semisimple. To make matters simpler, we shall assume simplicity of eigenvalues, but similar results can be obtained under the weaker assumption that they are semisimple, except that the expressions for the derivatives at a multiple eigenvalue are more complicated.

Note also that H3 may be assumed without loss of generality, as it just means that system (2.2) cannot be decoupled into two or more independent blocks. If this is not the case, then we can just apply the present results to each block separately.

The main results of the paper are contained in the following two theorems. The first concerns the destabilizing effects of the delay.

THEOREM 2.1. *Assume that system* (2.2) *satisfies hypothesis* H1–H3. *Then there exists a positive number τ_1, depending on G and C, such that the trivial solution is unstable for all τ in $(0, \tau_1)$. If C is negative definite, then $\tau_1 = +\infty$; that is, the system is unstable for all positive values of the delay parameter.*

There are, however, cases where the introduction of a delay will also have a stabilizing effect for larger values of the delay parameter.

THEOREM 2.2. *There exist systems of the form* (2.2) *and positive numbers τ_1 and τ_2 with $\tau_1 < \tau_2$ such that the trivial solution is stable when there is no delay, unstable for τ in $(0, \tau_1)$, and asymptotically stable for τ in (τ_1, τ_2).*

3. Some basic facts about the associated spectrum. In order to study the stability of the trivial solution of system (2.2), we look for solutions of the form $y(t) = e^{\lambda t}u$ and obtain the spectral problem

$$L_\tau(\lambda)u := \lambda^2 u + \lambda e^{-\lambda\tau}Gu + Cu = 0. \tag{3.1}$$

This is equivalent to solving the characteristic equation

$$d(\lambda, \tau) := \det\left(\lambda^2 I + \lambda e^{-\lambda\tau}G + C\right) = 0. \tag{3.2}$$

Note that, for fixed τ, the function d defined above is analytic in λ. When τ is zero, this problem will have $2n$ eigenvalues which, since this system is assumed to be stable, will lie on the imaginary axis and be semisimple. For positive values of τ, the presence of the exponential term has the effect of bringing in an infinite number of eigenvalues from the point at infinity, and also of breaking the Hamiltonian symmetry that is present when there is no delay.

That these eigenvalues do not accumulate near the imaginary axis is fundamental in the study of the stability of such problems. This is a consequence of the fact that on the one hand the eigenvalues are zeros of an analytic function, and, on the other, the set of eigenvalues with real part larger than any given number is bounded. These properties are summarized in the following proposition.

PROPOSITION 3.1. *Let α_0 and τ_0 be given real numbers where τ_0 is assumed to be positive. There exists a constant $\kappa = \kappa(\alpha_0, \tau_0)$ such that if* $\mathrm{Re}(\lambda) > \alpha_0$ *and* $\tau \in (0, \tau_0)$, *then $|\lambda| < \kappa$. As a consequence, for any given nonnegative value of the delay, there is only a finite number of eigenvalues with real part larger than α_0.*

For a proof, see [HL], for instance. Finally, recall that the stability of the trivial solution is determined by the eigenvalues (See, for instance, [HL, section 7.6]).

3.1. Purely imaginary eigenvalues. The discussion above implies that the mechanism underlying the changes of stability of the trivial solution is related to eigenvalues which cross the imaginary axis from the left to the right (increasing the number of unstable eigenvalues) or from the right to the left (decreasing this number). Since the system coefficients are real and it is assumed that C is nonsingular, this can only happen at pairs of purely imaginary eigenvalues. We shall now look for points of this type in the spectrum.

Letting $\lambda = \omega i$ ($\omega > 0$) in (3.1) gives

$$-\omega^2 u + \omega i e^{-\omega\tau i} G u + C u = 0. \tag{3.3}$$

Taking the inner product with u and assuming the eigenvectors to be normalized by $\|u\| = 1$, we obtain

$$\omega^2 + \omega e^{-\omega\tau i} g - c = 0,$$

where $ig = (Gu, u)$ and $c = (Cu, u)$. Here $(\cdot, \cdot)$ denotes the usual (complex) inner product. Separating this equation into real and imaginary parts yields

$$\begin{cases} \omega^2 + \omega g \cos(\omega\tau) - c = 0, \\ \omega g \sin(\omega\tau) = 0. \end{cases} \tag{3.4}$$

Since C is taken to be nonsingular, ω cannot vanish, and thus we have from the second equation that either $g = 0$ or $\omega\tau = k\pi$. Let's consider the latter case first. Substituting in (3.3), we get

$$-\omega^2 u + (-1)^k \omega i G u + C u = 0. \tag{3.5}$$

Because the matrices G and C are real, if ωi is an eigenvalue, then the same is true of $-\omega i$. We thus see that we may omit the factor $(-1)^k$ without loss of generality and look for solutions of

$$\omega^2 u - i\omega G u - C u = 0$$

instead. Now this is the same as the equation when τ vanishes, and thus we see that points of this type where eigenvalues may cross the imaginary axis are just the eigenvalues of the original problem, with the same multiplicities.

At this point we shall not consider eigenvalues where the inner product $ig = (Gu, u)$ vanishes, except to remark that when this happens we have from the first equation in (3.4) that $\omega^2 = c$. In particular, this implies that such points cannot exist when the matrix C is negative definite. We shall see below that they exist under other conditions and that they will play an important role in the stabilization mentioned in the introduction.

This discussion suggests the following distinction between two types of crossing points. If a point on the imaginary axis where eigenvalues may cross from one half-plane to the other is an eigenvalue of the system at τ equal zero, we shall say that it is a *primary crossing point*. Otherwise, it will be called a *secondary crossing point*.

Assume that when τ is zero the spectrum of system (2.2) consists of the points $\pm\omega_1 i, \ldots, \pm\omega_n i$, with $(0 <)\omega_1 < \cdots < \omega_n$. Then we have that there are eigenvalues on the imaginary axis at the points $\pm\omega_j$ for $\tau_k^j = k\pi/\omega_j$, $k = 0, 1, \ldots,$ $j = 1, \ldots, n$. Note that for even k each eigenvalue has the same eigenvector as it did at $\tau = 0$, while when k is odd the eigenvector is now the complex conjugate of that same eigenvector.

3.2. Derivatives of eigenvalues at primary crossing points. In order to determine the stability of the trivial solution for τ small, we shall consider the sign of the derivative of eigenvalues on the imaginary axis when τ is zero. First note that the derivative of an eigenvalue λ with respect to τ exists provided that λ is a (finite) semisimple eigenvalue. Thus, if we know the signs of these derivatives when there is no delay, this together with the properties mentioned at the beginning of this section will enable us to study the stability for small positive values of the delay parameter.

In order to proceed, we need to consider the adjoint eigenvalue problem $L_\tau^*(\lambda)v = 0$, where the adjoint operator $L_\tau^*(\lambda)$ is defined by the equality $(L_\tau(\lambda)u, v) = (u, L_\tau^*(\lambda)v)$ for all $u, v \in \mathbb{R}^n$. This gives

$$L_\tau^*(\lambda)v = \overline{\lambda}^2 v - \overline{\lambda} e^{-\overline{\lambda}\tau} Gv + Cv.$$

We now consider the eigenvalue problems for $L_{\tau+\delta}(\lambda)$ and $L_\tau^*(\lambda)$, that is, $L_{\tau+\delta}(\lambda)u = 0$ and $L_\tau^*(\lambda)v = 0$. Taking inner products at $\tau + \delta$ and τ, we get $(L_{\tau+\delta}(\lambda)u, v) = (L_\tau^*(\lambda)v, u) = 0$. Clearly u also depends on τ and δ, but we shall omit this dependence to keep notation simple. If we now take the complex conjugate in the second equality and then subtract, we are led to

$$\left[\lambda^2(\tau+\delta) - \lambda^2(\tau)\right](u, v) + [\lambda(\tau+\delta)e^{-\lambda(\tau+\delta)(\tau+\delta)} - \lambda(\tau)e^{-\lambda(\tau)\tau}](Gu, v) = 0.$$

Dividing by δ and letting it go to zero, we finally obtain

$$\lambda'(\tau) = \frac{\lambda^2(\tau)e^{-\lambda\tau}(Gu, v)}{2\lambda(\tau)(u, v) + [1 - \tau\lambda(\tau)]\, e^{-\lambda\tau}(Gu, v)}. \tag{3.6}$$

Now note that at a point where $\lambda = \omega i$ and $\omega\tau = k\pi$, we have

$$L_\tau(\omega i)u = -\omega^2 u + \omega i e^{-k\pi i} Gu + Cu = 0$$

and

$$L_\tau^*(\omega i)v = -\omega^2 v + \omega i e^{k\pi i} Gv + Cv = 0,$$

so that u and v can actually be taken to be the same. Thus, if we normalize eigenvectors such that $(u, u) = 1$, we obtain that the derivative at a primary crossing point ω for the delay $\tau = k\pi/\omega$ is given by

$$\lambda'(\tau) = \frac{(-1)^{k+1}\omega^2 g}{2\omega + (1 - k\pi i)(-1)^k g},$$

where $ig = (Gu, u)$. Since we have

$$(-1)^k g = \frac{c}{\omega} - \omega,$$

where $c = (Cu, u)$, we finally obtain that

$$\lambda'(\tau) = \frac{\omega^2(\omega^2 - c)}{(\omega^2 + c)^2 + k^2\pi^2(c - \omega^2)^2}\left[(\omega^2 + c) + ik\pi(c - \omega^2)^2\right]. \tag{3.7}$$

From this, we see that the sign of the real part of the derivative of an eigenvalue at a primary point ωi is the same as that of the product $(\omega^2 - c)(\omega^2 + c)$. Furthermore, this sign depends only on the value of the crossing point and is actually independent of the value of τ. These results are summarized in the following proposition.

Proposition 3.2. *The sign of the real part of the derivative of an eigenvalue at a primary crossing point ωi is given for any $\tau = k\pi/\omega$ $(k \in \mathbb{Z})$ by the sign of the product $(\omega^2 - c)(\omega^2 + c)$, where $c = (Cu, u)$, with u a normalized unit eigenvector associated to the corresponding eigenvalue at τ equal to* 0.

4. The destabilizing effect of delays. We shall now show that for the largest (in absolute value) of primary crossing points, the real part of the derivative with respect to τ is positive. Since, as we have seen, the sign of this derivative does not depend on the actual value of τ for which the crossing occurs, we may consider τ to be zero. We shall thus consider the operator $L_0(\lambda)$ for $\lambda = \omega i$. This gives

$$L_0(\omega i) = -\omega^2 I + \omega H + C,$$

where $H = iG$. As G is skew-symmetric, H will be Hermitian and thus $L_0(\omega i)$ is a Hermitian quadratic pencil for real ω. For each real value of ω, there exist n real eigenvalues (counting multiplicities) of $L_0(\omega i)$. The functions describing the behavior of the eigenvalues with respect to the parameter ω are called the eigencurves of the quadratic pencil. Note that eigenvalues of the original problem correspond to zero eigenvalues of $L_0(\omega i)$ and thus to intersections of the eigencurves with the horizontal axis.

Proposition 4.1. *At the largest primary crossing point we have*

$$\lambda'(0) > 0.$$

Proof. Begin by noting that τ equal to zero corresponds to k equal to zero also, and thus from the expression (3.7) for the derivative we see that $\lambda'(0)$ is actually real.

We shall now consider the auxiliary Hermitian spectral problem

$$L_0(\omega i)u = \sigma u,$$

where ω will be taken to be a real parameter. Denote by $c_1 \leq \cdots \leq c_n$ the eigenvalues of C and assume that a change of basis has been carried out such that $C = \mathrm{diag}\{c_1, \ldots, c_n\}$. We have that if $|\omega| \geq \omega_n$, then the matrix $L_0(\omega i)$ is negative semidefinite. In particular, we get

$$L_0(\omega_n i) = -\omega_n^2 I + \omega_n H + C \leq 0,$$

which implies that the diagonal entries of $L_0(\omega_n i)$ are less than or equal to zero. It thus follows that $-w_n^2 + c_n \leq 0$. Since $c_n \geq c = (Cu, u)$ for all u with unit norm, we have that $\omega_n^2 \geq c$. Now equality can hold if and only if $c = c_n$, that is, if u is the eigenvector corresponding to c_n. But as C is in diagonal form, this would mean that $u = (0, \ldots, 0, 1)$, from which we obtain that the entries g_{jn} of G will have to vanish also for all $j = 1, \ldots, n$. This would make G singular with u as an eigenvector associated to the zero eigenvalue, which is not possible by H3. Hence $\omega_n^2 - c > 0$.

We now observe that from $-\omega^2 + \omega(Hu, u) + (Cu, u) = 0$ and the fact that $\omega_n^2 - c > 0$, we obtain that $h := (Hu, u) = \omega_n - c/\omega_n$ is positive.

For the auxiliary spectral problem considered above we have that σ is a differentiable function of the (real) variable ω, and this derivative can easily be seen to be given by

$$\sigma'(\omega) = -2\omega + h.$$

Since $\omega_n i$ is the largest (in absolute value) eigenvalue of the original problem, it follows that $\sigma'(\omega_n)$ is less than or equal to zero. (It is actually negative, since we are assuming that eigenvalues are simple.) This yields that $2\omega_n \geq h(> 0)$ and since

$$(0 <)2\omega_n - h = \pm\sqrt{h^2 + 4c},$$

we have that for the largest eigenvalue the plus sign should be taken. Hence, $4\omega_n^2 + 4c \geq h^2 + 4c > 0$, and finally $\omega_n^2 + c > 0$ as desired. □

When C is negative definite, it is possible to improve the result above and to show that there are other eigenvalues besides that with largest absolute value for which this derivative is also positive. Keep in mind that if C is negative definite, then it is a necessary condition that n be even in order for the system without delay to be stable.

COROLLARY 4.2. *Assume that C is negative definite. Then there exist at least $n/2$ primary crossing points in the upper half-plane for which the real part of the derivative with respect to τ of an eigenvalue at these points is positive.*

Proof. If C is negative definite, we automatically have that $\omega^2 - c$ is positive. On the other hand at $\omega = 0$ all the eigencurves of L_0, are below the horizontal axis. This means that, since eigenvalues are simple (it is actually enough for them to be semisimple), there must exist n intersections of the eigencurves with the horizontal axis, of which $n/2$ have σ' positive, and the remaining $n/2$ have σ' negative. Again, the derivative cannot be zero since this implies double (nonsimple) eigenvalues in the original problem. Proceeding as in the proof of the previous result, it is possible to show that those crossing points which correspond to negative values of σ' will have $\omega_j^2 + c > 0$, and so the derivative with respect to τ at these points is positive. □

Proof of Theorem 2.1. The first part of the result follows immediately from the fact that for the largest eigenvalue $\lambda'(0)$ is positive, and from the continuity of eigenvalues with respect to τ.

To prove that the system will always be unstable when C is negative definite, we begin by noting that in this case there are no secondary crossing points. This follows from the second equation in (3.4) since if g is not zero we obtain a primary crossing point, while if it vanishes we get that $\omega^2 = c$ which is impossible when C is negative definite. Denote the number of eigenvalues with positive real parts by $\nu(\tau)$. We have that $\nu(\tau)$ is completely determined once we know the primary crossing points and their derivatives. This can be done by writing down the values τ_j^k for each crossing point $\omega_j i$, and then ordering them in increasing order and taking into account the sign of the respective derivative. We can thus build a table of the form

$$\begin{array}{ccccc}
\tau_n^1 & < & \tau_n^2 & < & \cdots \\
\tau_{n-1}^1 & < & \tau_{n-1}^2 & < & \cdots \\
 & & \vdots & & \\
\tau_1^1 & < & \tau_1^2 & < & \cdots,
\end{array}$$

where we also have that $\tau_j^k < \tau_{j-1}^k$. From Corollary 4.2 we have that at least $n/2$ of the derivatives are positive. In particular, this is the case for $\omega_n i$. We can see from the table above that the smallest positive value of τ for which there will be a crossing is given by $\tau_n^1 = \pi/\omega_n$. For τ between zero and τ_n^1 we know that there exist at least n eigenvalues with positive real parts and so $\nu(\tau) \geq n$ in this interval. After τ_n^1 and until we reach the next crossing point we shall have $\nu(\tau) \geq n+2$. We now claim that the number of primary points which are smaller than a positive number τ and which have a negative derivative will never exceed by more than one the number of primary points smaller than τ and for which the derivative is positive.

To prove the claim, notice that it holds on each column of the table above since when we consider the eigencurves of the auxiliary spectral problem we see that, travelling from the right to the left along the horizontal axis (travelling down a column), the number of intersections of eigencurves with the horizontal axis which have a negative derivative cannot be exceeded by those with a positive derivative. This means that as we travel down a column the difference can never be greater than one. Since the first point on each column always corresponds to the largest crossing point, the claim follows.

We end this section with a result on the sign of the derivative of an eigenvalue at the smallest primary crossing point.

Proposition 4.3. *If C is either negative or positive definite, then the derivative at the primary crossing points with the smallest absolute value has a negative real part.*

Proof. Consider first the case where C is positive definite. Then we have that $L_0(\omega i)$ is semipositive definite at $\omega = \omega_1$. In a similar fashion to the proof of Proposition 4.1, we obtain that $\omega_1^2 \leq c_1$. Since now $c_1 \leq c = (Cu, u)$ for all u with unit norm, we get that $w_1^2 - c \leq 0$, and, as before, we get that equality cannot hold and thus $w_1^2 - c < 0$. Since C is positive definite, $w_1^2 + c > 0$ and the result follows.

When C is negative definite, $w_1^2 - c$ is always positive. The derivative of the corresponding eigencurve at this point must now be positive and so $\sigma'(\omega_1) = -\omega_1 + h > 0$. Thus $h > \omega_1 (> 0)$. This implies that $2\omega_1 - h = -\sqrt{h^2 + 4c}$, from which we obtain $2\omega_1 + \sqrt{h^2+4c} = h$. Taking squares on both sides gives that $w_1^2 + c = -\omega_1\sqrt{h^2+4c} < 0$, which in turn implies that the real part of the derivative is negative. □

5. Secondary crossing points and the stabilizing effect of delays.

The analysis of secondary crossing points is much more difficult in general than that of primary points. There are several reasons why this is so. On the one hand, now we no longer have that the eigenvectors u and v in the expression (3.6) for the derivative can be taken to be the same. On the other hand, the equation that gives these points is also not easy to obtain explicitly in general. However, and as we shall see below, these points can play a decisive role in the stability of this type of system, and it would thus be interesting to study them in more detail.

5.1. The case of $n = 2$.

We shall now consider the special case where n is two, since this is the only case where a complete study of secondary crossing points may be carried out explicitly. In this case we may assume that

$$G = \begin{bmatrix} 0 & g \\ -g & 0 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}$$

(with $c_1 \leq c_2$), and so the characteristic function is now given by

$$d(\omega i, \tau) = \omega^4 - (c_1 + c_2)\omega^2 + c_1 c_2 - \omega^2 g^2 e^{-2\omega\tau i}.$$

As before, we are interested in the case of real ω. Separating this equation into real and imaginary parts, we have

$$\left\{\begin{array}{l}\omega^4-(c_1+c_2)\omega^2+c_1c_2-\omega^2g^2\cos(2\omega\tau)=0,\\ \omega^2g^2\sin(2\omega\tau)=0.\end{array}\right.$$

If we assume C to be nonsingular, then ω cannot be zero. On the other hand, g must also be different from zero, for otherwise there is no gyroscopic term. It then follows from the second equation that $2\omega\tau=m\pi$. For even m, substituting this in the first equation gives the primary crossing points. We are thus interested in the case of odd m, which leads to the following equation for the secondary points:

$$\omega^4-(c_1+c_2-g^2)\omega^2+c_1c_2=0,$$

from which we obtain that

$$\omega^2=\frac{c_1+c_2-g^2\pm\sqrt{(c_1+c_2-g^2)^2-4c_1c_2}}{2}.$$

When c_1 and c_2 are both negative, we already know that there are no secondary points. This can also be easily seen from the expression above. If $c_1c_2<0$, then the number of negative eigenvalues is odd and thus the system without delay can never be stable, and so we are left with the case where C is positive definite.

Note first that under these conditions the expression for ω^2 given above can only be positive provided that $c_1+c_2-g^2$ is positive. On the other hand, in order to get a positive term under the square root, we must have

$$\left(c_1+c_2-g^2-2\sqrt{c_1c_2}\right)\left(c_1+c_2-g^2+2\sqrt{c_1c_2}\right)>0,$$

which, because of the condition above, reduces to

$$c_1+c_2-g^2-2\sqrt{c_1c_2}>0.$$

We thus have that a necessary and sufficient condition for the existence of four secondary points when C is positive definite is that

$$\sqrt{c_2}-\sqrt{c_1}>g.$$

Denote these points by $\pm\tilde{\omega}_1$ and $\pm\tilde{\omega}_2$, with $\tilde{\omega}_1<\tilde{\omega}_2$. Clearly, then, $\pm\tilde{\omega}_1$ and $\pm\tilde{\omega}_2$ correspond to the minus and plus signs in the expression for ω^2.

The equation giving these roots in λ is

$$\lambda^4+(c_1+c_2)\lambda^2+c_1c_2+\lambda^2g^2e^{-2\lambda\tau}=0,$$

from which the following expression for the derivative of λ with respect to τ can be obtained:

$$\lambda'(\tau)=\frac{\lambda^2g^2e^{-2\lambda\tau}}{2\lambda^2+c_1+c_2+g^2e^{-2\lambda\tau}-\lambda\tau g^2e^{-2\lambda\tau}}.$$

At secondary crossing points $\lambda=\omega i$ and $\omega\tau=m\pi/2$ with m odd. Hence

$$\begin{aligned}\lambda'(\tau) &= \frac{\omega^2g^2}{-2\omega^2+c_1+c_2-g^2+m\pi g^2i/2}\\ &= \frac{\omega^2g^2\left[(-2\omega^2+c_1+c_2-g^2)-m\pi g^2i/2\right]}{(2\omega^2-c_1-c_2+g^2)^2+m^2\pi^2g^4/4},\end{aligned}$$

and the sign of the real part of the derivative is the same as that of $-2\omega^2+c_1+c_2-g^2$. On the other hand, we have

$$2\omega^2 - c_1 - c_2 + g^2 = \pm\sqrt{(c_1+c_2-g^2)^2-4c_1c_2},$$

and so the real parts of the derivatives at $\pm\tilde{\omega}_1$ and $\pm\tilde{\omega}_2$ are positive and negative, respectively. The fact that now the largest of the secondary crossing points has a negative derivative enables the system to be stabilized for positive values of the delays. In particular, we have the following proposition.

PROPOSITION 5.1. *In the case where* n *is equal to two and* C *is positive definite, there will exist numbers* $\tau_1 < \tau_2$ *such that the trivial solution of* (2.2) *is asymptotically stable on the interval* (τ_1, τ_2) *if and only if* $2\tilde{\omega}_2 > \omega_2$.

Proof. Denote the values of the delays for which eigenvalues cross the imaginary axis by τ_j^k and $\tilde{\tau}_j^k$, corresponding, respectively, to ω_j and $\tilde{\omega}_j$, $j = 1,2$. From the previous discussion, we know that $\tau_j^k = k\pi/\omega_j$, while $\tilde{\tau}_j^k = (2k-1)\pi/(2\tilde{\omega}_j)$, $k = 1,\ldots$. The first crossing for positive τ will then occur either at $\tau_2^1 = \pi/\omega_2$ or at $\tilde{\tau}_2^1 = \pi/(2\tilde{\omega}_2)$. Since, at τ equal to zero, one pair of eigenvalues moved to the right, while the other went to the left, we have that $\nu(\tau)$ is two between zero and the first crossing. Clearly, if this happens at $\tilde{\tau}_2^1$, the system becomes stable.

Assume now that the first crossing is at τ_2^1, that is, that $2\tilde{\omega}_2 < \omega_2$. To show that there will always exist at least one pair of eigenvalues with positive real parts, we shall use an argument similar to that used in the proof of Theorem 2.1. To this end, we begin by noting that $\tilde{\omega}_1 > \omega_1$:

$$2(\tilde{\omega}_1^2 - \omega_1^2) = -2g^2 - \sqrt{(c_1+c_2-g^2)^2-4c_1c_2} + \sqrt{(c_1+c_2+g^2)^2-4c_1c_2}.$$

This will be positive if and only if

$$c_1 + c_2 - g^2 > \sqrt{(c_1+c_2+g^2)^2-4c_1c_2},$$

which is clearly satisfied for all positive c_1 and c_2. We shall now build a table similar to that used in the proof of Theorem 2.1, except that because we need to include transitions both at primary and secondary points, it is not possible to ensure that the order of the crossing points in the first column will be the same as in the other columns. However, since we are assuming that $\omega_2 > 2\tilde{\omega}_2$, we get that $\tau_2^k = k\pi/\omega_2 > (2k-1)\pi/(2\tilde{\omega}_2) = \tilde{\tau}_2^k$ for all k. This ensures that the first transition in a column is always from the left to the right. On the other hand, because $\tilde{\omega}_1 > \omega_1$, we have that $\tilde{\tau}_1^k = (2k-1)\pi/(2\tilde{\omega}_1) < k\pi/\omega_1 = \tau_1^k$. This implies that in each column the last element is always one corresponding to a transition from the right to the left. The result now follows in a similar way to the proof of Theorem 2.1.

Finally, note that if $2\tilde{\omega}_2 = \omega_2$, then this cancels out two of the transitions in a column, reducing it to just two terms. As we have seen that the last term in column k is always τ_1^k, we again have instability. □

As a corollary to this result, we obtain that if too many eigenvalues are on the right-hand side for a given value of τ, then the trivial solution cannot become stable again for larger values of τ.

COROLLARY 5.2. *If* $\nu(\tau_*) = \nu_*$ *for some positive* τ_*, *then* $\nu(\tau) \geq \nu_* - 4$ *for all* $\tau > \tau_*$. *In particular, if* $\nu_* = 6$, *the trivial solution is unstable for all* $\tau > \tau_*$. *Furthermore, it will always become unstable for large enough values of the delay parameter.*

Proof. From the proof of the previous result we know that the last element in column k is always τ_1^k and so in order to reach it we must go past $\tilde{\tau}_1^k$. On the other hand,

$$2(\omega_2 - \tilde{\omega}_2) = 2g^2 + \sqrt{(c_1+c_2+g^2)^2 - 4c_1c_2} - \sqrt{(c_1+c_2-g^2)^2 - 4c_1c_2} > 0,$$

which gives that

$$\omega_2 > \tilde{\omega}_2 > \frac{2k}{2k+1}\tilde{\omega}_2,$$

and so $\tilde{\tau}_2^{k+1} > \tau_2^k$. This implies that before we reach $\tau = \tilde{\tau}_2^{k+1}$ we must first pass through τ_2^k. In this way, we see that ν can never decrease by more than four.

To prove the second part of the result, just notice that for large enough values of k we have that the top entry in each column is τ_2^k, as $\omega_2 > \tilde{\omega}_2$. This gives that there exists τ_* sufficiently large for which $\nu(\tau_*)$ is 6 and the result follows. □

With these results it is possible to determine all the stability intervals for any given 2×2 system (with simple eigenvalues), together with the associated stability sequence (the sequence of integers giving the number of eigenvalues with positive real part as the delay parameter is increased).

Example 5.1. Take $c_1 = 4$, $c_2 = 16$, and $g = 1$. We then get

$$\omega_1 \approx 1.9233, \quad \omega_2 \approx 4.1594, \quad \tilde{\omega}_1 \approx 2.0920, \quad \text{and} \quad \tilde{\omega}_2 \approx 3.8241.$$

Clearly, $2\tilde{\omega}_2 > \omega_2$, and in fact there will actually exist three stability intervals:

$$\begin{aligned} (\tilde{\tau}_2^1, \tilde{\tau}_1^1) &\approx (.4108, .7509), \\ (\tilde{\tau}_2^3, \tilde{\tau}_1^2) &\approx (2.0538, 2.2526), \\ \text{and } (\tilde{\tau}_2^5, \tilde{\tau}_1^3) &\approx (3.6969, 3.7543). \end{aligned}$$

The stability sequence in this example is

$$2,0,2,4,2,4,2,0,2,4,2,4,2,0,2,4,2,4,2,4,6,4,6,4,2,4,6,4,6,4,2\ldots,$$

and from Corollary 5.2 we get that these are the only stability intervals.

By changing the coefficients, it is possible to obtain examples with more stable intervals, and also examples for which, although there exist secondary crossing points, the system is always unstable.

Example 5.2. Take $c_1 = 1$, $c_2 = 16$, and $g = 1/2$. Then the stability sequence is given by

$$\begin{gathered} 2,0,2,0,2,4,2,4,2,4,2,0,2,0,2,4,2,4,2,4,2,0,2,0,2,4,2,4,2,4,2,0, \\ 2,0,2,4,2,4,2,4,2,0,2,0,2,4,2,4,2,4,2,0,2,0,2,4,2,4,2,4,2,0,2,0, \\ 2,4,2,4,2,4,2,0,2,4,2,4,6,4,6,4,2,4,2,4,6,4,6,4,6,4,2,4,2,4,\ldots. \end{gathered}$$

Example 5.3. Take $c_1 = 1$, $c_2 = 16$, and $g = 2.9$. Then $\sqrt{c_2} - \sqrt{c_1} - g = .1$, so that there exist secondary crossing points, but now the stability sequence is

$$2,4,2,4,6,8,6,8,10,12,10,12,10,12,10,12,14,16,14,16,18,20,18,\ldots,$$

and so the system is never stable.

6. Conclusions. We have shown that the simple model for gyroscopic systems given by (1.1) is not robust with respect to small delays in the sense that the introduction of an arbitrarily small delay in the gyroscopic term will make the trivial solution unstable under a quite general set of hypotheses. This is the case, in particular, when the gyroscopic term is used to stabilize a system where the matrix K is negative definite and thus the number of eigenvalues with positive real parts for the system without feedback is the maximum possible. For this situation we have seen that the system remains unstable for all values of the delay parameter. It is thus of interest to know under what conditions these are realistic modelling assumptions and whether or not more sophisticated models should be taken into account. These models might include having different delay parameters in the different variables, introducing delays in the stiffness term, or having distributed delays.

Acknowledgment. This work was carried out while I was visiting the Department of Mathematics of the Royal Institute of Technology in Stockholm, Sweden. I would like to thank the people there and, in particular, Ari Laptev, for their hospitality.

REFERENCES

[BLM] L. Barkwell, P. Lancaster, and A. S. Markus, *Gyroscopically stabilized systems: A class of quadratic eigenvalue problems with real spectrum*, Canad. J. Math., 44 (1992), pp. 42–53.

[BCD] J. F. Barman, F. M. Callier, and C. A. Desoer, *L^2-stability and L^2-instability of linear time-invariant distributed feedback systems perturbed by a small delay in the loop*, IEEE Trans. Automat. Control, AC–18 (1973), pp. 479–484.

[CG] K. L. Cooke and Z. Grossman, *Discrete delay, distributed delay and stability switches*, J. Math. Anal. Appl., 86 (1982), pp. 592–627.

[DLP] R. Datko, J. Lagnese, and M. P. Polis, *An example on the effect of time delays in boundary feedback stabilization of wave equations*, SIAM J. Control Optim., 24 (1986), pp. 152–156.

[DY] R. Datko and Y. C. You, *Some second-order vibrating systems cannot tolerate small time delays in their damping*, J. Optim. Theory Appl., 70 (1991), pp. 521–537.

[GS] T. T. Georgiou and M. C. Smith, *w-stability of feedback systems*, Systems Control Lett., 13 (1989), pp. 271–277.

[HL] J. K. Hale and S. V. Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.

[HKLP] R. Hryniv, W. Kliem, P. Lancaster, and C. Pommer, *A precise bound for gyroscopic stabilization*, ZAMM Z. Angew. Math. Mech., to appear.

[L] P. Lancaster, *Lambda-matrices and Vibrating Systems*, Pergamon Press, Oxford, UK, 1966.

[RT] R. Rebarber and S. Townley, *Robustness with respect to delays for exponential stability of distributed parameter systems*, SIAM J. Control Optim., 37 (1999), pp. 230–244.

[TT] W. Thomson and P. G. Tait, *Treatise on Natural Philosophy, Part* 1, Cambridge University Press, Cambridge, UK, 1921.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 208–225

ZERO SUM ABSORBING GAMES WITH INCOMPLETE INFORMATION ON ONE SIDE: ASYMPTOTIC ANALYSIS[*]

DINAH ROSENBERG[†]

Abstract. We prove the existence of the limit of the values of finitely repeated (resp., discounted) absorbing games with incomplete information on one side, as the number of repetitions goes to infinity (resp., the discount factor goes to zero). The main tool is the study of the Shapley operator, for which the value of the λ-discounted game is a fixed point, and of its derivative with respect to λ.

Key words. zero sum games, stochastic games, incomplete information, operator approach, recursive formula

AMS subject classification. 91A15

PII. S0363012999354430

Introduction. We analyze the asymptotic behavior of a class of discounted and finitely repeated two player zero sum games, namely, absorbing games with incomplete information on one side. Indeed, for a given discount factor λ or for a given length n of the game, the existence of the value of the discounted game and of the n-stage game is known. The question is to provide an asymptotic analysis of the values as the players become infinitely patient, or as the length of the game increases.

The existence of the limit of the values of the n-stage games v_n or of the λ-discounted games v_λ has been proved in two main classes of games. In the case of absorbing games with complete information (see Kohlberg [5]) or, more generally, in the case of stochastic games with complete information (see Bewley and Kohlberg [2] or Mertens and Neyman [8]), these are repeated games in which at each stage the actions of the players determine not only their payoffs but also the transitions of a given Markov process; stage payoffs are a function of the pair of actions and of the state of the Markov process at that stage. And in the case of games with no absorbing states but with lack of information on one side (see Aumann and Maschler [1]) or on both sides (Mertens and Zamir [9]), these are repeated games in which the players do not perfectly know the matrix of the game they are playing; this matrix is drawn with a given probability among a given set of matrices, and the players only get partial information on the exact matrix that has been drawn.

The purpose of this paper is to generalize this result to our framework. This question has been solved by Sorin in [14, 15] for the special case of "big match" with incomplete information on one side. These are 2×2 games in which one of the players controls the transition to an absorbing state (see Blackwell and Ferguson [3]).

We analyze the asymptotic behavior of discounted and finitely repeated two player zero sum absorbing games with incomplete information on one side. The game is defined by a family of payoff matrices indexed by two parameters: the first one is private information of player 1 and is kept fixed for the whole interaction; the second one, named the state, is a Markov process in which the transitions depend on the actions of the players. This defines a stochastic game with incomplete information

[*] Received by the editors March 24, 1999; accepted for publication (in revised form) November 1, 1999; published electronically August 17, 2000.
http://www.siam.org/journals/sicon/39-1/35443.html

[†] LAGA, Institut Galilée, Université Paris Nord, Avenue Jean Baptiste Clément, 93430 Villetaneuse, France (dinah@math.univ-paris13.fr). This paper was written during a stay at the Department of Statistics of Tel Aviv University, Israël.

on one side. In addition, the game is absorbing, meaning that all states but one are absorbing: a state is said to be absorbing if, independently of the actions of the players, the probability to leave it once it has been reached is 0. In that framework we prove the existence of the limit of the values of the λ-discounted games and of the n-stage games as λ goes to 0 or as n goes to infinity.

The argument used in this paper is an extension of the approach provided in [5]. It consists of studying a mapping T involved in the recursive formula defining the value of the repeated game. It expresses the total payoff of the repeated game as a weighted average of the first stage payoff and of a continuation payoff. The value v_λ of the λ-discounted game is the unique fixed point of $T(\lambda, .)$ (see Shapley [13]). As the length of the game increases, the weight λ of the first stage payoff tends to zero. The proof studies an expansion of this map with respect to λ. The heuristic idea behind this approach is that the most important thing for a player in a long game is to take care of his future payoffs to keep them above (resp., below) a given level; such a concern is captured in the control of the main part of the expansion. Second, given that he guarantees a good future payoff, a player should maintain the current nonabsorbing payoff above (below) this level; the second term of the expansion expresses this matter.

1. The model and the result.

1.1. The game. We consider an absorbing game with incomplete information on one side. It is defined by two finite sets of actions, I (the set of actions of player 1) and J (the set of actions of player 2), a finite set of parameters K (that represents the private information of player 1), a probability p on K, a finite set of stochastic states $\Omega = \{\omega_0\} \bigcup \Omega^*$ (Ω^* is the set of absorbing states and ω_0 is the unique nonabsorbing state), a transition $q : \Omega \times I \times J \to \Delta(\Omega)$, and an initial state $\omega_1 \in \Omega$. The transition satisfies that for any absorbing state $\omega^* \in \Omega^*$, and for any $(i, j) \in I \times J$, $q(\omega^*|\omega^*, i, j) = 1$. Note that we defined the transition to be independent of the information k. To each pair of actions (i, j) and each state $\omega \in \Omega$ is associated a vector payoff $a^\omega_{ij} = (a^{\omega,k}_{ij})_{k\in K}$. We assume without loss of generality that for all $k \in K$, $\omega \in \Omega$, $i \in I$, and $j \in J$, $a^{\omega,k}_{ij} \in [0, 1]$. The game is played as follows.

At stage 0, a parameter $k \in K$ is drawn according to probability p, and player 1 is informed of the result while player 2 is not. Both players know the probability p and the initial state ω_1.

At stage $m \geq 1$, player 1 chooses $i_m \in I$ and player 2 chooses $j_m \in J$; given the current state ω_m, a new state ω_{m+1} is drawn according to the probability $q(.|\omega_m, i_m, j_m)$, and the triple (i_m, j_m, ω_{m+1}) is announced to both players.

We assume perfect recall, i.e., both players remember what they have done and what they have known in the past. Therefore, by Kuhn's theorem, without loss of generality, one can reduce the strategy sets from mixed strategies to behavior strategies. A behavior strategy of player 1 (resp., of player 2) is a sequence $(\sigma_1, \ldots, \sigma_n, \ldots)$ (resp., $(\tau_1, \ldots, \tau_n, \ldots)$) such that σ_m is a function from $K \times \Omega \times (I \times J \times \Omega)^{m-1}$ to $\Delta(I)$ (resp., τ_m is a function from $\Omega \times (I \times J \times \Omega)^{m-1}$ to $\Delta(J)$). We set $H_n = \Omega \times (I \times J \times \Omega)^n$ and $H_\infty = (I \times J \times \Omega)^{\mathbb{N}}$. Thus $K \times H_\infty$ is the set of plays. Each $h_n \in H_n$ can be identified with a cylinder set of $K \times H_\infty$. Therefore, H_n induces a σ-algebra over $K \times H_\infty$, which we denote by $\mathcal{H}^2_n$ (the information of player 2 at stage n). Similarly, we define $\mathcal{H}^1_n$ (the information available to player 1 at stage n) as the σ-algebra induced by $K \times H_n$ over $K \times H_\infty$. The probability p, the transition q, and a pair of strategies of the players (σ, τ) induce a probability $P_{p,\sigma,\tau}$ over the set of plays.

Any play determines a stream of stage payoffs $(a^{\omega_m,k}_{i_m j_m})_{m\in\mathbb{N}}$ which can be evaluated in different ways.

In the finitely repeated game $G_n(p,\omega_1)$, the payoff is given by the expectation (according to the probability $P_{p,\sigma,\tau}$) of the average payoff: $\frac{1}{n}\big[\sum_{m=1}^{n} a^{\omega_m,k}_{i_m j_m}\big]$.

In the discounted game $G_\lambda(p,\omega_1)$, the payoff is given by the expectation of the discounted sum of payoffs (according to the probability $P_{p,\sigma,\tau}$): $\big[\sum_{m=1}^{\infty}\lambda(1-\lambda)^{m-1}a^{\omega_m,k}_{i_m j_m}\big]$.

By the minmax theorem (see [10, part A]), both $G_n(p,\omega_1)$ and $G_\lambda(p,\omega_1)$ have a value denoted, resp., by $v_n(p,\omega_1)$ and $v_\lambda(p,\omega_1)$. Our goal is to study the asymptotic behavior of these quantities as the game becomes very long, namely, as n goes to infinity or as λ goes to 0.

We denote by NR the subset of $\Delta(I)^K$ of non revealing strategies, namely the subset of $x\in\Delta(I)^K$ such that for all $(k,k')\in K^2$ and all $i\in I$, $x_i^k=x_i^{k'}$. NR can be identified with $\Delta(I)$. We aim at proving the following theorems.

THEOREM 1.1. *The family v_λ converges uniformly to a function v as λ goes to* 0.

THEOREM 1.2. *The sequence v_n converges uniformly to a function v' as n goes to infinity.*

THEOREM 1.3. $v=v'$.

One could also define the *uniform value*, the *maxmin*, and the *minmax* of the infinitely repeated game. The maxmin would be the largest quantity that player 1 can guarantee in all n-stage games for n large enough, with a strategy that is independent of n (see [10] for an exact definition); similarly, one defines the minmax; the uniform value exists if the maxmin equals the minmax. In particular, if such a uniform value exists, one can prove that it is equal both to v and to v'. The existence of such a value implies Theorems 1.1, 1.2, and 1.3.

In the case of perfect information absorbing games, Kohlberg [5] proves that indeed such a value exists. This result has been generalized for finite perfect information stochastic games by Mertens and Neyman [8]. In the case of games with incomplete information on one side, the existence of the uniform value has been proved by Aumann and Maschler [1]. Nevertheless, in the case of absorbing games with incomplete information on one side, Sorin [14] proved that the uniform value needs not exist (the maxmin and the minmax may differ). Here, we prove the weaker result stated in Theorems 1.1, 1.2, and 1.3. There is a conjecture by Sorin [15] and Mertens [7] stating that in stochastic games with incomplete information on one side the maxmin exists and is equal to v and v'. This paper gives no indication about that conjecture in our framework.

1.2. Basic properties of v_λ and v_n. A function from $\Delta(K)\times\Omega$ to $\mathbb{R}$ is said to be concave, continuous, Lipschitz, etc. if it is respectively concave, continuous, and Lipschitz in the first variable for any given $\omega\in\Omega$.

Let us recall the following properties of the functions v_n and v_λ.

LEMMA 1.4.

1. *The functions v_n and v_λ take their values in* $[0,1]$.
2. *The functions v_n and v_λ are Lipschitz with constant* 1 *in p.*
3. *The functions v_n and v_λ are concave in p.*

Proof. As all the payoffs are in $[0,1]$, so is the value that is computed as a convex combination of the payoffs in the matrices. The Lipschitz property follows also, and the concavity expresses the positive value of information in zero sum games. For detailed proofs, see the splitting lemma ([1, Lemma 5.3, p. 25] and [10, Chapter 5, p. 218]). □

We use the following uniform norm on the set of functions from $\Delta(K) \times \Omega$ to $[0, 1]$:

$$\|f\| = \sup_{p \in \Delta(K), \omega \in \Omega} |f(p, \omega)|.$$

Let $\mathcal{B}$ be the set of functions from $\Delta(K) \times \Omega$ to $[0, 1]$ that are Lipschitz with constant 1. $\mathcal{B}$ is a set of uniformly equicontinuous functions, and by Ascoli's theorem it is compact. Therefore, both the family v_λ and the sequence v_n have converging subsequences, and they converge iff all converging subsequences have the same limit.

Notation. Let v be the limit of a converging subsequence of the family $\{v_\lambda\}$. We are going to prove that such a v is unique.

PROPOSITION 1.5. *For any $\omega^* \in \Omega^*$, the functions $v_n(., \omega^*)$ and $v_\lambda(., \omega^*)$ converge uniformly to the same limit $v(., \omega^*)$.*

Proof. The game with initial absorbing state $\omega^* \in \Omega^*$ is reduced to a nonstochastic incomplete information game. In [1], Aumann and Maschler prove the convergence for such games. □

The purpose of this paper is to prove that this result generalizes to the nonabsorbing state ω_0.

2. Basic tools.

2.1. The mapping T. $\mathcal{F}$ is the set of continuous concave functions f from $\Delta(K) \times \Omega$ to $[0, 1]$. Note that the functions in $\mathcal{F}$ do not have to be Lipschitz. Actually all the functions we will use (v_λ, v_n) are not only continuous but Lipschitz with constant 1. Nevertheless, we are going to define on $\mathcal{F}$, for each λ and each pair of strategies $x \in \Delta(I)^K$ and $y \in \Delta(J)$, a map $T_{xy}(\lambda, .)$. This operator maps continuous functions to continuous functions, but it is not clear whether the image of a Lipschitz function with constant 1 is a Lipschitz function with constant 1.

The proof relies on the following mappings T and ϕ. We consider, for $f \in \mathcal{F}$, $\lambda \in [0, 1]$, $\omega_1 \in \Omega$, and $p \in \Delta(K)$ the normal form game $\Gamma(\lambda, f)(p, \omega_1)$ with strategy sets $\Delta(I)^K$, $\Delta(J)$ and where the payoff associated to the pair of strategies $(x, y) \in \Delta(I)^K \times \Delta(J)$ is

$$\begin{aligned} T_{xy}(\lambda, f)(p, \omega_1) = \Bigg[& \lambda \sum_{k \in K} \sum_{\substack{i \in I \\ j \in J}} p^k x_i^k y_j a_{ij}^{\omega_1, k} \\ & + (1 - \lambda) \sum_{k \in K} \sum_{\substack{i \in I \\ j \in J}} \sum_{\omega \in \Omega} p^k x_i^k y_j q(\omega | \omega_1, i, j) f(p_i, \omega) \Bigg], \end{aligned}$$

with $p_i \in \Delta(K)$ and $p_i^k = \frac{p^k x_i^k}{\sum_{l \in K} p^l x_i^l}$ if $\sum_{l \in K} p^l x_i^l \neq 0$ and $p_i = p^*$ if $\sum_{l \in K} p^l x_i^l = 0$ (for a given p^* in $\Delta(K)$).

The above definition of p_i as the conditional probability on K given player 1 played action i will be used throughout the paper. Notice that if x is in NR (meaning that for all $k \in K$, $k' \in K$, $i \in I$, $x_i^k = x_i^{k'}$), then for all $i \in I$, $p_i = p$.

$\Gamma(\lambda, f)(p, \omega_1)$ is a one shot representation of the repeated game in which the stream of payoffs from stage 2 on is evaluated through f. The total payoff of the repeated game is then the weighted average of the first stage payoff (with weight λ) and the continuation payoff (with weight $1-\lambda$).

Lemma 2.1. *If f is in $\mathcal{F}$, then the functions $x \mapsto T_{x,y}(\lambda, f)(p,\omega)$ and $y \mapsto T_{x,y}(\lambda, f)(p,\omega)$ defined, resp., on $\Delta(I)^K$ and $\Delta(J)$ are continuous and, resp., concave and convex.*

Proof. $y \mapsto T_{x,y}(\lambda, f)(p,\omega)$ is linear and continuous. For $x \mapsto T_{xy}(\lambda, f)(p,\omega)$, it is enough to prove that the function

$$\varphi : x \mapsto \sum_{k\in K} \sum_{\substack{i\in I\\ j\in J}} \sum_{\omega'\in\Omega} p^k x_i^k y_j q(\omega'|\omega,i,j) f(p_i,\omega')$$

is concave and continuous. The proof is analog to the one in the nonstochastic case [4].

Assume $x = \alpha x_1 + (1-\alpha)x_2$ with $\alpha \in [0,1]$; denote by $\bar{x}_i$ the quantity $\sum_{k\in K} p^k x_i^k$, and define similarly $\bar{x}_{1i}$ and $\bar{x}_{2i}$; then $p_i = \mu p_{1i} + (1-\mu)p_{2i}$ with $\mu = \alpha\bar{x}_{1i}/\bar{x}_i$ ($\mu \in [0,1]$) and $p_{1i}^k = p^k x_{1i}^k/\bar{x}_{1i}$ ($p_{1i} \in \Delta(K)$; p_{2i} is defined in an analogous way). Concavity of f then implies

$$\begin{aligned}\varphi(x) \quad \le \quad & \sum_{\substack{i\in I\\ j\in J}} \sum_{\omega'\in\Omega} \mu\bar{x}_i y_j q(\omega'|\omega,i,j) f(p_{1i},\omega') \\ & + \sum_{\substack{i\in I\\ j\in J}} \sum_{\omega'\in\Omega} (1-\mu)\bar{x}_i y_j q(\omega'|\omega,i,j) f(p_{2i},\omega').\end{aligned}$$

Using $\mu = \alpha\bar{x}_{1i}/\bar{x}_i$ and $1-\mu = (1-\alpha)\bar{x}_{2i}/\bar{x}_i$, this proves concavity of φ.

For continuity of φ at point x, fix an $\varepsilon > 0$. For all $\eta > 0$, there is an $\alpha > 0$ such that $\|x - x_1\| \le \alpha$ implies for all i satisfying $\bar{x}_i \ge \varepsilon$, $\|p_i - p_{1i}\| \le \eta$. Continuity of f therefore implies that if $\|x - x_1\| \le \alpha$, $\bar{x}_i \ge \varepsilon$, then $\max_{\omega\in\Omega} |f(p_i,\omega) - f(p_{1i},\omega)| \le \varepsilon$. The result follows. □

The sets of strategies $\Delta(I)^K$, $\Delta(J)$ are compact convex. Therefore, the previous lemma and the minmax theorem imply the existence of a value $T(\lambda, f)(p,\omega_1)$ for the game $\Gamma(\lambda, f)(p,\omega_1)$ and of optimal strategies for the players. The sets of optimal strategies of player 1 and of player 2 are denoted, resp., by $X_\lambda[f](p,\omega_1)$ and $Y_\lambda[f](p,\omega_1)$. The following lemma is a restatement of the recursive formula and the fixed point formula characterizing v_λ and v_n, in terms of the mapping T. These formulas justify the introduction of the mappings T.

Lemma 2.2.

(a) *For all $\lambda \in [0,1]$, $T(\lambda, v_\lambda) = v_\lambda$.*

(b) *For all $n \in \mathbb{N}$, $T(\frac{1}{n+1}, v_n) = v_{n+1}$.*

Proof. These formulas are proved in a more general setup in [10, p. 187]. □

Remark. *In the case of finite perfect information stochastic games, Bewley and Kohlberg* [2] *note that these formulas can be restated as:*

$$\begin{cases} \forall\omega\in\Omega,\ \forall j\in J,\ T_{x_\lambda j}(\lambda, v_\lambda)(\omega) \ge v_\lambda(\omega), \\ \forall\omega\in\Omega,\ \forall i\in I,\ T_{iy_\lambda}(\lambda, v_\lambda)(\omega) \le v_\lambda(\omega), \end{cases}$$

where $x_\lambda \in X_\lambda[v_\lambda](p,\omega)$ and $y_\lambda \in Y_\lambda[v_\lambda](p,\omega)$. This system is a finite number of polynomial inequalities and hence proves that v_λ, x_λ, and y_λ are semialgebraic in λ. Therefore they have Puiseux expansions in λ. Together with the fact that these functions are uniformly bounded, this proves Theorem 1.1. *In our framework, under x_λ, the state space can be reduced to a countable subset of $\Delta(K)\times\Omega$. The algebraic* *approach does not extend because the fixed point formula of Lemma* 2.2(a) *does not reduce to a finite number of polynomial inequalities but rather to*

$$\begin{cases} \forall \omega, p,\ \forall j \in J,\ T_{x_\lambda j}(\lambda, v_\lambda)(p,\omega) \geq v_\lambda(p,\omega), \\ \forall \omega, p,\ \forall x \in \Delta(I)^K,\ T_{x y_\lambda}(\lambda, v_\lambda)(p,\omega) \leq v_\lambda(p,\omega). \end{cases}$$

Indeed, $T_{xy_\lambda}(\lambda, v_\lambda)(p,\omega)$ is no more linear in x.

Let us now define the quantity:

$$\phi_{xy}(f)(p,\omega_1) = \frac{T_{xy}(\lambda, f)(p,\omega_1) - T_{xy}(0,f)(p,\omega_1)}{\lambda}.$$

Note that $\phi_{xy}(f)(p,\omega_1)$ is independent of λ. Indeed, for any λ,

$$\begin{aligned} \phi_{xy}(f)(p,\omega_1) &= \sum_{k\in K} \sum_{\substack{i\in I \\ j\in J}} \sum_{\omega\in\Omega} p^k x_i^k y_j q(\omega|\omega_1, i, j)(a_{ij}^{\omega_1,k} - f(p_i,\omega)) \\ &= \sum_{k\in K} \sum_{\substack{i\in I \\ j\in J}} p^k x_i^k y_j a_{ij}^{\omega_1,k} - T_{xy}(0,f)(p,\omega_1). \end{aligned} \tag{2.1}$$

Let now $\Gamma'(f)(p,\omega_1)$ be the game in which the strategy spaces are $X_0[f](p,\omega_1)$ and $Y_0[f](p,\omega_1)$, and the payoff associated to the pair of strategies (x,y) is $\phi_{xy}(f)(p,\omega_1)$.

LEMMA 2.3. *The game* $\Gamma'(f)(p,\omega_1)$ *has a value* $\phi(f)(p,\omega_1)$ *and optimal strategies for both players. Moreover,*

$$\phi(f)(p,\omega_1) = \lim_{\lambda\to 0} \frac{T(\lambda, f)(p,\omega_1) - T(0,f)(p,\omega_1)}{\lambda}.$$

Proof. The existence of $\lim_{\lambda\to 0} \frac{T(\lambda,f)(p,\omega_1)-T(0,f)(p,\omega_1)}{\lambda}$ and of the value of $\Gamma'(f)(p,\omega_1)$ as well as their equality follow from an extension of Mills' result [11] proved in [10, part A, exercise 6, p. 12]. □

This lemma implies that there are two possible interpretations of the introduction of ϕ: the first one is that since $T(\lambda, v_\lambda) = v_\lambda$, a local analysis of the mapping $T(\alpha, .)$ at the point $\alpha = 0$ should characterize the limit of (v_λ). Moreover, this limit is not completely characterized by the condition $T(0,v) = v$. That explains the introduction of ϕ as the derivative of the map T. Mills' result implies, moreover, that this derivative can be viewed as the value of an auxiliary game in which the players take care of their current payoffs under the constraint that they guarantee the value $T(0,v)$.

2.2. Idea of the proof. The functions v_λ are characterized by the fact they are fixed points of the contracting mappings $T(\lambda, f)$. Therefore, the properties of v_λ, such as convergence should follow from a study of the mappings themselves. The idea here is to study an asymptotic expansion of $T(\lambda, f)$ as λ goes to 0. Actually we will need a first-order expansion using $T(0,f)$ and the derivative $\phi(f)$ with respect to λ. Indeed, a necessary condition for v to be the limit of v_λ is that $T(0,v) = v$. But this condition is not sufficient, and we need to go one step further in the expansion. When the values are real numbers (i.e., when K is a singleton), this idea goes back to Kohlberg [5].

In the case of *complete information absorbing games*, any limit v of a converging subsequence of v_λ is a fixed point of $T(0,.)$. But Kohlberg points out that being a fixed point of $T(0,.)$ does not characterize the limit of the family $\{v_\lambda\}$. Indeed, $T(0,v)$ does not depend on nonabsorbing payoffs, and $T(0,v) = v$ implies only that player 1 can guarantee an absorbing payoff of at least v. The control of the nonabsorbing payoff is related to the sign of $\phi(v)$: the limit v is characterized, in this framework, by the following system, where w is any function such that $w(.,\omega^*)$ is equal to the limit $v(.,\omega^*)$ of $v_\lambda(.,\omega^*)$ for all absorbing states $\omega^* \in \Omega^*$:

$$\begin{cases} w > v \ \Rightarrow \ T(0,w) < w \text{ or } \{T(0,w) = w \text{ and } \phi(w) < 0\} & \mathcal{S}_1, \\ w < v \ \Rightarrow \ T(0,w) > w \text{ or } \{T(0,w) = w \text{ and } \phi(w) > 0\} & \mathcal{S}_2. \end{cases}$$

The intuition behind such a system is that if $w > v$, either $T(0,w) < w$, and therefore player 1 cannot maintain the level w, i.e., cannot prevent absorbing with a payoff smaller than w, or $T(0,w) = w$, and $\phi(w) < 0$. Due to the definition of the strategy sets in game $\Gamma'(w)(p,\omega_1)$, this means that player 1 cannot *simultaneously* maintain the level and guarantee a large enough current payoff (compared to w). Hence player 1 cannot guarantee w in the λ-discounted game for λ close to 0.

In the case of complete information absorbing games [5, 12] and in the case of incomplete information games with no absorbing states [1, 12], one can prove that any limit v of a converging subsequence of v_λ satisfies the above system $(\mathcal{S}_1, \mathcal{S}_2)$ and that this system has a unique solution, which proves the result.

The problem in the case of incomplete information games is that v is a function of p, and the study of the equation $T(0,v) = v$ does not reduce to a control of the absorbing payoffs. On the other hand, unlike in nonstochastic games with incomplete information, the equation $T(0,v) = v$ cannot be reduced to a concavity property of v. We are going to prove only that the limit v of any converging subsequence of v_λ satisfies $\mathcal{S}_1$ (see section 2.3). The idea is then to prove that this equation indeed implies that if $w > v$, player 1 cannot guarantee w in the λ-discounted games for λ close enough to 0: then this equation implies that $v = \limsup v_\lambda$, and therefore v is unique (see section 3). More precisely, in section 3 we prove the following (where ε is a positive constant and δ a positive strictly concave function of p, $\bar{\varepsilon}$ denotes the real function on Ω defined by $\bar{\varepsilon}(\omega_0) = \varepsilon$, and for all $\omega^* \in \Omega^*$, $\bar{\varepsilon}(\omega^*) = 0$).

PROPOSITION 2.4. *Let $f \in \mathcal{F}$ satisfy:*

$$T(0, f + \delta + \bar{\varepsilon}) \leq f + \delta + \bar{\varepsilon},$$
$$T(0, f + \delta + \bar{\varepsilon})(p,\omega_0) = f(p,\omega_0) + \delta(p) + \varepsilon \ \Rightarrow \ \phi(f + \delta + \bar{\varepsilon})(p,\omega_0) < 0.$$

Then $f \geq v$.

2.3. Properties. Recall that v is the limit of a converging subsequence of $\{v_\lambda\}$. In this section we prove some general properties of $T(\lambda, f)$. This section aims at proving $\mathcal{S}_1$, namely, that for functions w chosen in some class of functions that always lie above v, $T(0,w) \leq w$ and $T(0,w)(p,\omega) = w(p,\omega)$ implies that $\phi(w)(p,\omega) < 0$. This is the content of Lemmas 2.7 and 2.10. The first lemma is a contracting property of $T(\lambda,.)$ and a regularity property of the correspondences of optimal strategies $X_\lambda[f](p,\omega)$ and $Y_\lambda[f](p,\omega)$.

LEMMA 2.5.

1. *For any $f \in \mathcal{F}$ and any $g \in \mathcal{F}$, for all $p \in \Delta(K)$, $\omega \in \Omega$, and $\lambda \in [0,1]$, $\lambda' \in [0,1]$,*

$$T(\lambda, f)(p,\omega) - T(\lambda', g)(p,\omega) \leq \ (1-\lambda)\max\left[\max_{\substack{q\in\Delta(K)\\ \omega'\in\Omega}} \{f(q,\omega') - g(q,\omega')\}, 0\right] + 2|\lambda - \lambda'|.$$

2. *For any $f \in \mathcal{F}$ and any $\omega \in \Omega$, the correspondences $(\lambda, p) \to X_\lambda[f](p,\omega)$ and $(\lambda, p) \to Y_\lambda[f](p,\omega)$ are upper semicontinuous.*

Remark. *Property* 1 *implies, in particular, the contracting property of* $T(\lambda,.)$ *and the nonexpansiveness of* $T(0,.)$.

$$\|T(\lambda,f)-T(\lambda,g)\| \leq (1-\lambda)\|f-g\|.$$

Proof. Let $(f,g)\in\mathcal{F}^2$ and $(\lambda,\lambda')\in[0,1]^2$. Take $x\in X_\lambda[f](p,\omega)$ and $y\in Y_{\lambda'}[g](p,\omega)$; then

$$T(\lambda,f)(p,\omega)-T(\lambda',g)(p,\omega)\leq T_{xy}(\lambda,f)(p,\omega)-T_{xy}(\lambda',g)(p,\omega).$$

Using the fact that the payoffs and f take their values in [0,1], one gets

$$\begin{aligned}T_{xy}(\lambda,f)(p,\omega)-T_{xy}(\lambda',g)(p,\omega) \quad &\leq \quad 2|\lambda-\lambda'|\\ &+(1-\lambda)\sum_{k\in K}\sum_{\substack{i\in I\\ j\in J}}p^k x_i^k y_j q(\omega'|\omega,i,j)\left(f(p_i,\omega')-g(p_i,\omega')\right),\end{aligned}$$

which implies result 1.

Assume λ_n converges to λ and p_n converges to p. Let $x_n\in X_{\lambda_n}[f](p_n,\omega)$. It is straightforward to check that if x is the limit of a converging subsequence of x_n, then $x\in X_0[f](p,\omega)$. The argument is similar for Y. □

Lemma 2.6. *For any accumulation point* v *of* v_λ, $T(0,v)=v$.

Proof. The result is obtained by letting λ go to 0 in the formula of Lemma 2.2 and by using the continuity of T proved in Lemma 2.5 and the uniform convergence of a subsequence of v_λ to v. □

From now on, we fix $\varepsilon>0$, and we denote by $\bar\varepsilon$ the function from Ω to $[0,\varepsilon]$ satisfying $\bar\varepsilon(\omega_0)=\varepsilon$ and for $\omega^*\in\Omega^*$, $\bar\varepsilon(\omega^*)=0$. Let δ be a fixed 1-Lipschitz strictly concave function from $\Delta(K)$ to $[0,\varepsilon]$. For $f\in\mathcal{F}$, the function $f+\delta+\bar\varepsilon$ therefore satisfies $(f+\delta+\bar\varepsilon)(p,\omega_0)=f(p,\omega_0)+\delta(p)+\varepsilon$, and for $\omega^*\in\Omega^*$, $(f+\delta+\bar\varepsilon)(p,\omega^*)=f(p,\omega^*)+\delta(p)$. Lemmas 2.7 and 2.8 concern properties of the mappings T and the optimal strategies correspondences X and Y that are needed to prove the desired property stated in Lemma 2.10. The idea is that for some functions $w>v$, $T(0,w)(p,\omega_0)=w(p,\omega_0)$ implies that under optimal strategies $X_0[w](p,\omega_0)$ and $Y_0[w](p,\omega_0)$ there is no revelation of information and absorption occurs with probability 0. In addition, in Lemma 2.8 some regularity properties of the correspondences of optimal strategies are stated.

Lemma 2.7.

$$\begin{aligned}&\forall(x,y)\in\Delta(I)^K\times\Delta(J),\ \forall p,\ \forall f\in\mathcal{F},\\ &T_{xy}(0,f+\delta+\bar\varepsilon)(p,\omega_0)\leq T_{xy}(0,f)(p,\omega_0)+\delta(p)+\varepsilon.\end{aligned}$$

Assume $f\in\mathcal{F}$ *and* $p\in\Delta(K)$ *satisfy* $T(0,f)(p,\omega_0)\leq f(p,\omega_0)$; *then*

$$T(0,f+\delta+\bar\varepsilon)(p,\omega_0)\leq f(p,\omega_0)+\delta(p)+\varepsilon.$$

In particular,

$$T(0,v+\delta+\bar\varepsilon)(p,\omega_0)\leq v(p,\omega_0)+\delta(p)+\varepsilon.$$

Proof. Let $x\in\Delta(I)^K$ and $y\in\Delta(J)$. Then

$$\begin{aligned}T_{xy}(0,f+\delta+\bar\varepsilon)(p,\omega_0) \quad &\leq \quad \sum_{k\in K}\sum_{\substack{i\in I\\ j\in J}}\sum_{\omega'\in\Omega}p^k x_i^k y_j q(\omega'|\omega_0,i,j)\left(f(p_i,\omega')+\delta(p_i)+\varepsilon\right)\\ &\leq \quad T_{xy}(0,f)(p,\omega)+\sum_{k\in K}\sum_{\substack{i\in I\\ j\in J}}p^k x_i^k y_j\left(\delta(p_i)+\varepsilon\right).\end{aligned}$$

Therefore, using the concavity of δ, one gets the first assertion. By choosing $x \in X_0[f+\delta+\bar{\varepsilon}](p,\omega_0)$ and $y \in Y_0[f](p,\omega_0)$, one gets the second assertion. Lemma 2.6 implies the third assertion. $\square$

This lemma would be a direct consequence of the contracting property of $T(\lambda, v)$ if δ were a constant. As δ is a function of p, one has to take care of the splitting that occurs in $\Gamma(\lambda, v+\delta+\bar{\varepsilon})(p,\omega)$. The concavity of δ then implies that such a splitting does not increase the payoffs. We denote by $\mathcal{C}_\varepsilon(p)$ the following condition.

Condition $\mathcal{C}_\varepsilon(p)$. p is in the interior of $\Delta(K)$ and $T(0, v+\delta+\bar{\varepsilon})(p,\omega_0) = v(p,\omega_0) + \delta(p) + \varepsilon$.

The following lemma states a tightness property. The basic tightness property would be the following: if h is a positive constant and $T(0,v)(p,\omega_0) = v(p,\omega_0)$ and $T(0,v+h)(p,\omega_0) = v(p,\omega_0)+h$, then for any $h' \leq h$, $T(0,v+h')(p,\omega_0) = v(p,\omega_0)+h'$. In Lemma 2.8 we get a more precise result that allows us, in particular, to let p vary.

LEMMA 2.8. *Under condition* $\mathcal{C}_\varepsilon(p)$, *we have the following.*

(a) $X_0[v+\delta+\bar{\varepsilon}](p,\omega_0) \subset NR$.

(b) $X_0[v+\delta+\bar{\varepsilon}](p,\omega_0) \subset X_0[v+\delta+\bar{\varepsilon}'](p',\omega_0)$, *where* $\varepsilon' < \varepsilon$ *and* $\bar{\varepsilon}' : \Omega \to [0,\varepsilon]$ *is defined by* $\bar{\varepsilon}'(\omega^*) = 0$ *for* $\omega^* \in \Omega^*$ *and* $\bar{\varepsilon}'(\omega_0) = \varepsilon'$; *and*

$$|p - p'| = \sum_{k\in K} |p^k - p'^k| \leq \frac{\varepsilon - \varepsilon'}{4}.$$

Moreover,

$$T(0, v+\delta+\bar{\varepsilon}')(p',\omega_0) = v(p',\omega_0) + \delta(p') + \varepsilon'.$$

(c) $Y_0[v+\delta+\bar{\varepsilon}'](p,\omega_0) \subset Y_0[v+\delta+\bar{\varepsilon}](p,\omega_0)$ *for* $\bar{\varepsilon}'$ *defined as above.*

Notation. For a pair $(x,y) \in \Delta(I)^K \times \Delta(J)$ of mixed moves of the players, let μ^*_{xy} denote the measure on Ω^* defined by

$$\mu^*_{xy}(\omega^*) = \sum_{k\in K} \sum_{\substack{i\in I\\ j\in J}} p^k x^k_i y_j q(\omega^*|\omega_0, i, j).$$

$\mu^*_{xy}(\omega^*)$ is the probability to reach state ω^* if we are in state ω_0, and the players play (x,y). Note that μ^* is a measure but not a probability distribution on Ω^*.

Proof.

(a) Let $x \in X_0[v+\delta+\bar{\varepsilon}](p,\omega_0)$ and $y \in Y_0[v](p,\omega_0)$. Condition $\mathcal{C}_\varepsilon(p)$ and Lemma 2.6 imply

$$v(p,\omega_0) + \delta(p) + \varepsilon \quad \leq \quad T_{xy}(0, v+\delta+\bar{\varepsilon})(p,\omega_0).$$

The proof of Lemma 2.7 and the definition of y imply

$$T_{xy}(0, v+\delta+\bar{\varepsilon})(p,\omega_0) \leq v(p,\omega_0) + \sum_{\substack{k\in K\\ i\in I}} p^k x^k_i \delta(p_i) + \varepsilon.$$

These two inequalities lead to

$$\sum_{\substack{k\in K\\ i\in I}} p^k x^k_i \delta(p_i) \geq \delta(p).$$

Therefore, by strict concavity of δ, and since p is in the interior of $\Delta(K)$, the result follows.

(b) Let $x \in X_0[v+\delta+\bar{\varepsilon}](p,\omega_0)$ and $y \in \Delta(J)$. Condition $\mathcal{C}_\varepsilon(p)$ leads to:

$$T_{xy}(0,v+\delta+\bar{\varepsilon})(p,\omega_0) \geq v(p,\omega_0)+\delta(p)+\varepsilon.$$

By (a) and $x \in NR$, this inequality can be rephrased:

$$\sum_{\omega^*\in\Omega^*} \mu^*_{xy}(\omega^*)\left(v(p,\omega^*)+\delta(p)\right) \geq \sum_{\omega^*\in\Omega^*} \mu^*_{xy}(\omega^*)\left(v(p,\omega_0)+\delta(p)+\varepsilon\right).$$

By definition of p' and using the Lipschitz property of δ and v,

$$\begin{aligned}\sum_{\omega^*\in\Omega^*} \mu^*_{xy}(\omega^*)\left(v(p',\omega^*)+\delta(p')\right) &\geq \sum_{\omega^*\in\Omega^*} \mu^*_{xy}(\omega^*)\left(v(p',\omega_0)+\delta(p')+\varepsilon-4|p-p'|\right)\\ &\geq \sum_{\omega^*\in\Omega^*} \mu^*_{xy}(\omega^*)\left(v(p',\omega_0)+\delta(p')+\varepsilon'\right).\end{aligned}$$

This implies $T_{xy}(0,v+\delta+\bar{\varepsilon}')(p',\omega_0) \geq v(p',\omega_0)+\delta(p')+\varepsilon'$. By Lemma 2.7, $v(p',\omega_0)+\delta(p')+\varepsilon' \geq T(0,v+\delta+\bar{\varepsilon}')(p',\omega_0)$. Hence

$$\begin{aligned}T(0,v+\delta+\bar{\varepsilon}')(p',\omega_0) &\geq \inf_{y\in\Delta(J)} T_{xy}(0,v+\delta+\bar{\varepsilon}')(p',\omega_0)\\ &\geq v(p',\omega_0)+\delta(p')+\varepsilon'\\ &\geq T(0,v+\delta+\bar{\varepsilon}')(p',\omega_0).\end{aligned}$$

This inequality drives the desired result.

(c) Let $x \in \Delta(I)^K$ and $y \in Y_0[v+\delta+\bar{\varepsilon}'](p,\omega_0)$.

$$\begin{aligned}T_{xy}(0,v+\delta+\bar{\varepsilon})(p,\omega_0) &= T_{xy}(0,v+\delta+\bar{\varepsilon}')(p,\omega_0)\\ &+ \sum_{k\in K}\sum_{\substack{i\in I\\ j\in J}} p^k x_i^k y_j q(\omega_0|\omega_0,i,j)(\varepsilon-\varepsilon')\\ &\leq v(p,\omega_0)+\delta(p)+\varepsilon'+(\varepsilon-\varepsilon')\end{aligned}$$

by Lemma 2.7. By condition $\mathcal{C}_\varepsilon(p)$, this last inequality is the result. □

Let $d(y,y') = \max_{j\in J}|y_j - y'_j|$ for $y \in \Delta(J)$, $y' \in \Delta(J)$. For $Y \subset \Delta(J)$, $d(y,Y) = \inf_{y'\in Y} d(y,y')$ is the distance between y and the set Y.

LEMMA 2.9. *Under condition $\mathcal{C}_\varepsilon(p)$, there is an $\alpha > 0$ such that for all $y \in \Delta(J)$,*

$$\begin{aligned}&d\left(y, Y_0[v+\delta+\bar{\varepsilon}](p,\omega_0)\right) \geq \varepsilon/4\\ \Rightarrow\ &\\ &\exists x \in \Delta(I)^K,\ T_{xy}(0,v+\delta+\bar{\varepsilon})(p,\omega_0) \geq v(p,\omega_0)+\delta(p)+\varepsilon+\alpha.\end{aligned}$$

Proof. This lemma follows from the compactness of $\Delta(J)$ and the continuity of $T_{xy}(0,v+\delta+\bar{\varepsilon})(p,\omega_0)$ in y. □

The following is the key lemma.

LEMMA 2.10. *Under condition $\mathcal{C}_\varepsilon(p)$,*

$$\phi(v+\delta+\bar{\varepsilon})(p,\omega_0) < 0.$$

The idea of the proof is the following. Assume that for some p, the result is false; then there is a small enough λ such that given any strategy τ of player 2, player 1 has a reply that ensures him a payoff strictly greater than $v_\lambda(p,\omega_0)$. Indeed, if the mixed move y of player 2 is far from $Y_0[v+\delta+\bar{\varepsilon}](p,\omega_0)$, then by Lemma 2.9, player 1 has a reply that strictly increases the level of future payoffs: absorption occurs with a positive probability and a high payoff. If, on the other hand, y is close to $Y_0[v+\delta+\bar{\varepsilon}](p,\omega_0)$, then the hypothesis that $\phi(v+\delta+\bar{\varepsilon})(p,\omega_0)\geq 0$ implies that player 1 has a reply that gives him at the same time a good absorbing and a good nonabsorbing payoff.

Notation. For a function $f\in\mathcal{F}$, let us denote by $\bar{X}_0[f](p,\omega_0)$ the set of optimal strategies in $\Gamma'[f](p,\omega_0)$. This is a subset of $X_0[f](p,\omega_0)$.

Proof. Let us prove the result by contradiction, assuming that for some p, $T(0,v+\delta+\bar{\varepsilon})(p,\omega_0)=v(p,\omega_0)+\delta(p)+\varepsilon$ and $\phi(v+\delta+\bar{\varepsilon})(p,\omega_0)\geq 0$.

For ε, δ, and p, fix α as in Lemma 2.9. Then choose a λ such that

$$\begin{cases} \|v_\lambda - v\| \leq \beta = \min(\varepsilon/4,\alpha/4), \\ \lambda\leq \alpha/8. \end{cases}$$

For any $y\in\Delta(J)$, we will exhibit an $x\in\Delta(I)^K$ with $T_{xy}(\lambda,v_\lambda)(p,\omega_0)>v_\lambda(p,\omega_0)$. Hence $T(\lambda,v_\lambda)(p,\omega_0)>v_\lambda(p,\omega_0)$, which contradicts Lemma 2.2.

- If $d\left(y,Y_0[v+\delta+\bar{\varepsilon}](p,\omega_0)\right)\geq\varepsilon/4$, then by Lemma 2.9, there is an $x\in\Delta(I)^K$ such that

  $$T_{xy}(0,v+\delta+\bar{\varepsilon})(p,\omega_0)\geq v(p,\omega_0)+\delta(p)+\varepsilon+\alpha.$$

  Hence, by Lemma 2.7,

  $$T_{xy}(0,v)(p,\omega_0)\geq v(p,\omega_0)+\alpha.$$

  Let us now compute $T_{xy}(\lambda,v_\lambda)$. Lemma 2.5 implies

  $$\begin{aligned} T_{xy}(\lambda,v_\lambda)(p,\omega_0) &\geq T_{xy}(0,v)(p,\omega_0)-\alpha/4-\beta \\ &\geq v(p,\omega_0)+\alpha-\alpha/2 \\ &\geq v_\lambda(p,\omega_0)+\alpha/4. \end{aligned}$$

- If $d(y,Y_0[v+\delta+\bar{\varepsilon}](p,\omega_0)<\varepsilon/4$, let $x\in\bar{X}_0[v+\delta+\bar{\varepsilon}](p,\omega_0)$. By Lemma 2.8, x is nonrevealing; thus $T_{xy}(0,v+\delta+\bar{\varepsilon})(p,\omega_0)\geq v(p,\omega_0)+\delta(p)+\varepsilon$ and the concavity of δ imply

  $$\sum_{\omega^*\in\Omega^*}\mu^*_{xy}(\omega^*)v(p,\omega^*)\geq\sum_{\omega^*\in\Omega^*}\mu^*_{xy}(\omega^*)(v(p,\omega_0)+\varepsilon).$$

  By definition of λ,

  $$\sum_{\omega^*\in\Omega^*}\mu^*_{xy}(\omega^*)v_\lambda(p,\omega^*)\geq\sum_{\omega^*\in\Omega^*}\mu^*_{xy}(\omega^*)(v_\lambda(p,\omega_0)+\varepsilon-2\beta),$$

  and therefore

  $$T_{xy}(0,v_\lambda)(p,\omega_0)\geq v_\lambda(p,\omega_0). \tag{2.2}$$

  Define now $y_0\in Y_0[v+\delta+\bar{\varepsilon}](p,\omega_0)$ such that $\|y-y_0\|\leq\varepsilon/4$. Then

  $$\sum_{k\in K}\sum_{\substack{i\in I\\ j\in J}}p^k x_i y_{0j}(a^{\omega_0 k}_{ij}-v(p,\omega_0)-\delta(p)-\varepsilon)=\phi_{xy_0}(v+\delta+\bar{\varepsilon})(p,\omega_0)\geq 0.$$

  Thus, by continuity,

  $$\sum_{k\in K}\sum_{\substack{i\in I\\ j\in J}}p^k x_i y_j a^{\omega_0 k}_{ij} \quad\geq\quad v_\lambda(p,\omega_0)+\varepsilon-\beta-\varepsilon/4.$$

This last inequality and (2.2) imply

$$T_{xy}(\lambda, v_\lambda)(p, \omega_0) \geq v_\lambda(p, \omega_0) + \lambda\varepsilon/2.$$

- Hence for any y, we have an x such that $T_{xy}(\lambda, v_\lambda)(p, \omega_0) > v_\lambda(p, \omega_0)$, and this contradicts the recursive formula and proves the result. □

3. The proof of Theorem 1.1. Our goal is to prove that v is the unique limit of a converging subsequence of $\{v_\lambda\}$.

If in addition to Lemma 2.10, one had a symmetric result, namely that $T(0, v - \bar{\varepsilon})(p, \omega_0) \geq v(p, \omega_0) - \varepsilon$, and $T(0, v - \bar{\varepsilon})(p, \omega_0) = v(p, \omega_0) - \varepsilon \ \Rightarrow\ \phi(v - \bar{\varepsilon})(p, \omega_0) > 0$, then uniqueness of such a v would be a consequence of the definition of ϕ as the derivative of T, and of the contracting property of $T(\lambda, .)$ stated in Lemma 2.5. This is the case for incomplete information games with no absorbing states, and for absorbing games with complete information (see [5], [12]). In the case of absorbing games with incomplete information, in order to prove this property by contradiction, one should assume that for some p, $T(0, v - \bar{\varepsilon})(p, \omega_0) = v(p, \omega_0) - \varepsilon$ and $\phi(v - \bar{\varepsilon})(p, \omega_0) < 0$. But the difficulty in concluding that player 2 has a best response to any strategy of player 1 that leads to a payoff strictly inferior to v_λ in a λ-discounted game arises from the fact that since the strategies of player 1 may be revealing, the equation $T(0, v - \bar{\varepsilon})(p, \omega_0) = v(p, \omega_0) - \varepsilon$ does not lead to a bound on the absorbing payoffs.

But, without such a property Lemma 2.10 implies that player 1 cannot guarantee $v + \delta + \bar{\varepsilon}$ in a λ-discounted game with λ small, because he cannot simultaneously push his absorbing and nonabsorbing payoffs above $v + \delta + \varepsilon$. More precisely, the intuition for the end of the proof is the following: the function $v + \delta + \bar{\varepsilon}$ is such that for every p, either $T(0, v + \delta + \bar{\varepsilon})(p, \omega_0) < v(p, \omega_0) + \delta(p) + \varepsilon$ or $T(0, v + \delta + \varepsilon)(p, \omega) = v(p, \omega_0) + \delta(p) + \varepsilon$, and $\phi(v + \delta + \bar{\varepsilon})(p, \omega_0) < 0$. By Lemma 2.3, this implies that for λ small enough (depending on p)

$$T(\lambda, v + \delta + \bar{\varepsilon})(p, \omega_0) \leq v(p, \omega_0) + \delta(p) + \varepsilon.$$

Would such a λ be uniform in p, then such an inequality would prove $v_\lambda(p, \omega_0) \leq v(p, \omega_0) + \delta(p) + \varepsilon$. But λ needs not be uniform in p. Nevertheless, the remainder of the proof aims at establishing that for λ small enough $v_\lambda(p, \omega_0) \leq v(p, \omega_0) + \delta(p) + \varepsilon$. Uniformity of λ is replaced by the following proposition.

PROPOSITION 3.1. *For any sequences $\lambda_n \in [0, 1]$ and $p_{\lambda_n} \in \Delta(K)$ such that λ_n converges to 0 as n goes to infinity and p_{λ_n} converges to $\bar{p}$ in the interior of $\Delta(K)$,*

$$\exists N,\ \forall n \geq N,\ T(\lambda_n, v + \delta + \bar{\varepsilon})(p_{\lambda_n}, \omega_0) < v(p_{\lambda_n}, \omega_0) + \delta(p_{\lambda_n}) + \varepsilon.$$

Proof. Assume by contradiction that there is a sequence λ_n converging to 0 and a sequence p_{λ_n} converging to $\bar{p}$ in the interior of $\Delta(K)$ such that

$$\forall N,\ \exists n \geq N,\ T(\lambda_n, v + \delta + \bar{\varepsilon})(p_{\lambda_n}, \omega_0) \geq v(p_{\lambda_n}, \omega_0) + \delta(p_{\lambda_n}) + \varepsilon.$$

Note that, by concavity, one always has $v(p_{\lambda_n}, \omega_0) + \delta(p_{\lambda_n}) + \varepsilon \geq T(0, v + \delta + \bar{\varepsilon})(p_{\lambda_n}, \omega_0)$. In what follows, let us denote by Λ the (infinite) subset of $(\lambda_n)_{n \in \mathbb{N}}$ satisfying

$$(3.1)\quad T(\lambda, v + \delta + \bar{\varepsilon})(p_\lambda, \omega_0) \geq v(p_\lambda, \omega_0) + \delta(p_\lambda) + \varepsilon \geq T(0, v + \delta + \bar{\varepsilon})(p_\lambda, \omega_0),$$

and by $\bar{p}$ the limit of a converging subsequence of $(p_\lambda)_{\lambda \in \Lambda}$. The proposition will be a consequence of Lemmas 3.2 and 3.3. Indeed, they imply that for some λ, $T(0, v + \delta + \bar{\varepsilon}/2)(p_\lambda, \omega_0) = v(p_\lambda, \omega_0) + \delta(p_\lambda) + \varepsilon/2$ and $\phi(v + \delta + \bar{\varepsilon}/2)(p_\lambda, \omega_0) \geq 0$. This contradicts Lemma 2.10. □

Lemma 3.2.

(a) *For* $\lambda \in \Lambda$, *for all* $x_\lambda \in X_\lambda[v+\delta+\bar{\varepsilon}](p_\lambda,\omega_0)$, $y_\lambda \in Y_0[v+\delta+\bar{\varepsilon}](p_\lambda,\omega_0)$,

$$\phi_{x_\lambda y_\lambda}(v+\delta+\bar{\varepsilon})(p_\lambda,\omega_0) \geq 0.$$

(b) *For* $\varepsilon' \leq \varepsilon$,

$$T(0,v+\delta+\bar{\varepsilon}')(\bar{p},\omega_0) = v(\bar{p},\omega_0)+\delta(\bar{p})+\varepsilon'.$$

(c) *For all* ε' *satisfying* $\varepsilon/8 \leq \varepsilon' < \varepsilon$, *there is a* λ^* *such that for* $\lambda \in \Lambda$, $\lambda \leq \lambda^*$,

$$T(0,v+\delta+\bar{\varepsilon}')(p_\lambda,\omega_0) = v(p_\lambda,\omega_0)+\delta(p_\lambda)+\varepsilon',$$

$$X_0[v+\delta+\bar{\varepsilon}](\bar{p},\omega_0) \subset X_0[v+\delta+\bar{\varepsilon}'](p_\lambda,\omega_0).$$

Proof.

(a)

$$\phi_{x_\lambda y_\lambda}(v+\delta+\bar{\varepsilon})(p_\lambda,\omega_0) \quad = \frac{T_{x_\lambda y_\lambda}(\lambda,v+\delta+\bar{\varepsilon})(p_\lambda,\omega_0) - T_{x_\lambda y_\lambda}(0,v+\delta+\bar{\varepsilon})(p_\lambda,\omega_0)}{\lambda}.$$

Therefore inequality (3.1) implies the result.

(b) Inequality (3.1) and the continuity of T (Lemma 2.5) imply by going to the limit:

$$T(0,v+\delta+\bar{\varepsilon})(\bar{p},\omega_0) = v(\bar{p},\omega_0)+\delta(\bar{p})+\varepsilon.$$

This is condition $\mathcal{C}_\varepsilon(\bar{p})$. Lemma 2.8(b) leads to the result.

(c) Lemma 3.2(b) is condition $\mathcal{C}_{\varepsilon'}(\bar{p})$; Lemma 2.8(b) then implies the result. □

Lemma 3.3. *Fix* $\varepsilon' = \varepsilon/2$, *and define* λ^* *as in Lemma* 3.2(c). *There is a* $\lambda \leq \lambda^*$, $\lambda \in \Lambda$ *such that*

$$\phi(v+\delta+\bar{\varepsilon}/2)(p_\lambda,\omega_0) \geq 0.$$

Proof. Define $y_\lambda \in Y_0[v+\delta+\bar{\varepsilon}/2](p_\lambda,\omega_0)$ to be optimal in $\Gamma'[v+\delta+\bar{\varepsilon}/2](p_\lambda,\omega_0)$. For $\lambda \in \Lambda$, $\lambda \leq \lambda^*$, let $x_\lambda \in X_\lambda[v+\delta+\bar{\varepsilon}](p_\lambda,\omega_0)$, and define x to be the limit of a converging subsequence of x_λ. Choose λ such that $\|x-x_\lambda\| \leq \varepsilon/2$. By Lemma 2.5, $x \in X_0[v+\delta+\bar{\varepsilon}](\bar{p},\omega_0)$. By Lemma 3.2c, $x \in X_0[v+\delta+\bar{\varepsilon}/2](p_\lambda,\omega_0)$.

Therefore,

$$\phi_{xy_\lambda}(v+\delta+\bar{\varepsilon}/2)(p_\lambda,\omega_0) \leq \phi(v+\delta+\bar{\varepsilon}/2)(p_\lambda,\omega_0). \tag{3.2}$$

Recall that by Lemma 2.8, $x \in NR$ and $y_\lambda \in Y_0[v+\delta+\bar{\varepsilon}](p_\lambda,\omega_0)$. By (2.1) and Lemma 3.2(c),

$$\begin{aligned}\phi_{xy_\lambda}(v+\delta+\bar{\varepsilon}/2)(p_\lambda,\omega_0) \quad &= \quad \sum_{\substack{i\in I\\ j\in J}}\sum_{k\in K} p_\lambda^k x_i y_{\lambda j} a_{ij}^{k\omega_0} - (v(p_\lambda,\omega_0)+\delta(p_\lambda)+\varepsilon/2)\\ &\geq \quad \phi_{x_\lambda y_\lambda}(v+\delta+\bar{\varepsilon})(p_\lambda,\omega_0)+\varepsilon/2-\|x-x_\lambda\|.\end{aligned}$$

Lemma 3.2(a) implies $\phi_{x_\lambda y_\lambda}(v+\delta+\bar{\varepsilon})(p_\lambda,\omega_0) \geq 0$, and the result follows. □

We now establish Theorem 1.1. The proof proceeds by induction on the cardinality of K, *card* (K).

If *card* $(K) = 1$, the game is a complete information game and by [5] v_λ converges.

Assume the result is true for *card* $(K) \leq \mathcal{K}-1$, and let *card* $(K) = \mathcal{K}$. The induction hypothesis implies that on the boundary of $\Delta(K)$, v_λ converges.

Let us prove the convergence of v_λ by contradiction. Assume that there is a sequence λ_n such that λ_n goes to 0 and such that v_{λ_n} converges uniformly to some function v' different from v. By Proposition 1.5, $v(p,\omega)$ and $v'(p,\omega)$ are equal as soon as $\omega \in \Omega^*$. So, we assume the following.

Hypothesis **H**. There is a $p_0 \in \Delta(K)$ such that $v'(p_0,\omega_0) > v(p_0,\omega_0)$.

By induction hypothesis, p_0 is in the interior of $\Delta(K)$. Consider a sequence λ_n such that $\lim_{n\to\infty} \lambda_n = 0$ and v_{λ_n} converges uniformly to v' as n goes to infinity.

Let $\varepsilon > 0$ satisfy $v'(p_0,\omega_0) > v(p_0,\omega_0) + 4\varepsilon$; let δ be a strictly concave continuous function from $\Delta(K)$ to $[0,\varepsilon]$.

Now, for all λ, let $\hat{p}_\lambda$ belong to $\arg\max_{p\in\Delta(K)} [v_\lambda(p,\omega_0) - v(p,\omega_0) - \delta(p)]$. Assume $\|v_\lambda - v'\| \leq \varepsilon/16$. Then **H** and the definition of ε imply $v_\lambda(\hat{p}_\lambda,\omega_0) - v(\hat{p}_\lambda,\omega_0) - \delta(\hat{p}_\lambda) - \varepsilon > 0$.

Define $\hat{p}$ as the limit of a converging subsequence $\hat{p}_{\lambda_{s(n)}}$ of $\hat{p}_{\lambda_n}$. Then, $v'(\hat{p},\omega_0) - v(\hat{p},\omega_0) - \delta(\hat{p}) - \varepsilon \geq 0$. By induction hypothesis, $\hat{p}$ is in the interior of $\Delta(K)$.

Set Λ_0 the subset of $(\lambda_{s(n)})_{n\in\mathbb{N}}$ such that for all $\lambda \in \Lambda_0$,

$$\begin{cases} v_\lambda(\hat{p}_\lambda,\omega_0) > v(\hat{p}_\lambda,\omega_0) + \delta(\hat{p}_\lambda) + \varepsilon, \\ \hat{p}_\lambda \text{ is in the interior of } \Delta(K), \\ \|v_\lambda - v'\| \leq \varepsilon/16. \end{cases}$$

Lemma 3.4. *For all* $\lambda \in \Lambda_0$ *and all* $\varepsilon/8 \leq \varepsilon' \leq \varepsilon$,

$$T(\lambda, v+\delta+\bar{\varepsilon}')(\hat{p}_\lambda,\omega_0) \geq v(\hat{p}_\lambda,\omega_0) + \delta(\hat{p}_\lambda) + \varepsilon'.$$

Proof. For $p \in \Delta(K)$ and $\omega^* \in \Omega^*$, $v_\lambda(p,\omega^*) \leq v'(p,\omega^*) + \varepsilon/16 = v(p,\omega^*) + \varepsilon/16$. Therefore, by definition of $\hat{p}_\lambda$,

$$\begin{aligned} &T(\lambda, v_\lambda)(p,\omega_0) - T(\lambda, v+\delta+\bar{\varepsilon}')(p,\omega_0) \\ &\quad \leq \max\left\{0, \max_{\substack{r\in\Delta(K)\\ \omega'\in\Omega}} [v_\lambda(r,\omega') - v(r,\omega') - \delta(r) - \bar{\varepsilon}'(\omega')]\right\} \\ &\quad \leq v_\lambda(\hat{p}_\lambda,\omega_0) - v(\hat{p}_\lambda,\omega_0) - \delta(\hat{p}_\lambda) - \varepsilon'. \end{aligned}$$

Since $T(\lambda, v_\lambda) = v_\lambda$, the last inequality implies, for $p = \hat{p}_\lambda$,

$$T(\lambda, v+\delta+\bar{\varepsilon}')(\hat{p}_\lambda,\omega_0) \geq v(\hat{p}_\lambda,\omega_0) + \delta(\hat{p}_\lambda) + \varepsilon',$$

which is the desired result. □

Proof of Theorem 1.1. We have proved that $\hat{p}$ is in the interior of $\Delta(K)$. Therefore $(\lambda_{s(n)})$ satisfies the hypothesis of Proposition 3.1. But Lemma 3.4 and Proposition 3.1 are in contradiction. Hence the assumption **H** is impossible, and for any v' such that there is a subsequence of v_λ that converges uniformly to v',

$$\forall p,\ v'(p,\omega_0) \leq v(p,\omega_0).$$

But since v is any limit of a uniformly converging subsequence of v_λ, this implies that $v' = v$. The uniqueness of such a limit gives the result, i.e., implies the convergence of v_λ as λ goes to 0 when *card* $(K) = \mathcal{K}$. The theorem is then proved by induction. □

The previous results do not depend on the exact definition of v. The only important property of v is that $T(0, v+\delta+\bar{\varepsilon}) \leq v+\delta+\bar{\varepsilon}$ and $T(0, v+\delta+\bar{\varepsilon})(p,\omega_0) =$

$v(p,\omega_0)+\delta(p)+\varepsilon \ \Rightarrow\ \phi(v+\delta+\bar{\varepsilon})(p,\omega_0) < 0$. This property implies that there cannot be a subsequence of v_λ and p_λ such that $v(p_\lambda) < v_\lambda(p_\lambda)$. Therefore, we have proved the following proposition.

Proposition 3.5. *Let $f \in \mathcal{F}$ satisfy*

$$\begin{aligned} &T(0, f+\delta+\bar{\varepsilon}) \le f+\delta+\bar{\varepsilon},\\ &T(0, f+\delta+\bar{\varepsilon})(p,\omega_0) = f(p,\omega_0)+\delta(p)+\varepsilon \ \Rightarrow\ \phi(f+\delta+\bar{\varepsilon})(p,\omega_0) < 0; \end{aligned}$$

then $f \ge v$.

This proposition will be used in the following section.

4. The case of finitely repeated games. The same argument will apply to prove the convergence of v_n to v. Although, since v_n is no longer a fixed point of T, the proofs might be longer. We only sketch them and mention the complete argument when it is different from above.

Denote by w the limit (in the sense of the uniform norm) of a fixed converging subsequence of v_n and by v the limit of v_λ. We prove that $v = w$ and therefore that w is unique. Let $\underline{w}$ be the lim inf of (v_n). $\underline{w}$ is 1-Lipschitz and concave.

Let ε, $\bar{\varepsilon}$, and δ be defined as previously.

Lemma 4.1. *For all $p \in \Delta(K)$, we have the following.*

(a) $T(0,w)(p,\omega_0) = w(p,\omega_0)$.

(b) $T(0,\underline{w}+\delta+\bar{\varepsilon})(p,\omega_0) \le \underline{w}(p,\omega_0)+\delta(p)+\varepsilon$.

(c) $T(0,\underline{w}+\delta+\bar{\varepsilon})(p,\omega_0) = \underline{w}(p,\omega_0)+\delta(p)+\varepsilon \ \Rightarrow\ \phi(\underline{w}+\delta+\bar{\varepsilon})(p,\omega_0) < 0$.

Proof.

(a) If $(v_{s(n)})$ is a subsequence of (v_n) that converges uniformly to w, then $(v_{s(n)+1})$ also converges uniformly to w, and therefore for any p, $T(0,w)(p,\omega_0) = w(p,\omega_0)$.

(b) Lemma 4.1(a) implies that for any w and any p,

$$T(0,\underline{w})(p,\omega_0) \le T(0,w)(p,\omega_0) = w(p,\omega_0).$$

But $\underline{w}(p,\omega_0)$ is the infimum of all possible $w(p,\omega_0)$. Hence $T(0,\underline{w})(p,\omega_0) \le \underline{w}(p,\omega_0)$. The result follows from Lemma 2.7.

(c) This step is analog to Lemma 2.10. One proceeds by contradiction and assumes

$$T(0,\underline{w}+\delta+\bar{\varepsilon})(p,\omega_0) = \underline{w}(p,\omega_0)+\delta(p)+\varepsilon \text{ and } \phi(\underline{w}+\delta+\bar{\varepsilon})(p,\omega_0) \ge 0.$$

Define α as in Lemma 2.9 (applied for the function $\underline{w}$), and let $\eta \le \min(\alpha/8, \varepsilon/4)$. By definition of $\underline{w}$, there is an N such that for all $n \ge N$, $v_n \ge \underline{w} - \eta$ and $\frac{1}{n} \le \eta$.

Fix $y \in \Delta(J)$. There are two cases (see Lemma 2.9): either there is an $x \in \Delta(I)^K$ such that $T(0,\underline{w}+\delta+\bar{\varepsilon})(p,\omega_0) \ge \underline{w}(p,\omega_0)+\delta(p)+\varepsilon+\alpha$, or there is an $x \in X_0[\underline{w}+\delta+\bar{\varepsilon}](p,\omega_0)$ such that $\phi_{xy}(\underline{w}+\delta+\bar{\varepsilon})(p,\omega_0) \ge -\varepsilon/2$.

In the first case, there is an $x \in \Delta(I)^K$ such that $T(0,\underline{w}+\delta+\bar{\varepsilon})(p,\omega_0) \ge \underline{w}(p,\omega_0)+\delta(p)+\varepsilon+\alpha$. Therefore, for $n \ge N$,

$$T_{xy}\left(\frac{1}{n+1}, v_n\right)(p,\omega_0) \ge -\frac{2}{n+1} + T_{xy}(0,\underline{w}-\eta)(p,\omega_0).$$

Hence

$$T_{xy}\left(\frac{1}{n+1}, v_n\right)(p,\omega_0) \ge -\frac{2}{n+1} + w(p,\omega_0)+\alpha-\eta. \tag{4.1}$$

In the second case, there is an $x \in X_0[\underline{w} + \delta + \bar{\varepsilon}](p, \omega_0)$ such that $\phi_{xy}(\underline{w} + \delta + \bar{\varepsilon})(p, \omega_0) \geq -\varepsilon/2$. By Lemma 2.8a, $x \in NR$. Hence

$$\sum_{\substack{i\in I\\ j\in J}} \sum_{k\in K} p^k x_i y_j a_{ij}^{k\omega_0} \geq \underline{w}(p, \omega_0) + \delta(p) + \varepsilon/2.$$

Moreover, $T_{xy}(0, \underline{w} + \delta + \bar{\varepsilon})(p, \omega_0) \geq \underline{w}(p, \omega_0) + \delta(p) + \varepsilon$; therefore

$$\sum_{\omega^*\in\Omega^*} \mu_{xy}^*(\omega^*)\underline{w}(p, \omega^*) \geq \sum_{\omega^*\in\Omega^*} \mu_{xy}^*(\omega^*)(\underline{w}(p, \omega_0) + \varepsilon).$$

This leads to

$$\begin{array}{rcl} T_{xy}\left(\dfrac{1}{n+1}, v_n\right)(p, \omega_0) & \geq & \dfrac{1}{n+1}(\underline{w}(p, \omega_0) + \delta(p) + \varepsilon/2) \\ & + & \dfrac{n}{n+1} \displaystyle\sum_{\omega^*\in\Omega^*} \mu_{xy}^*(\omega^*)(\underline{w}(p, \omega_0) + \varepsilon - \eta) \\ & + & \dfrac{n}{n+1} \displaystyle\sum_{\substack{i\in I\\ j\in J}} x_i y_j q(\omega_0|\omega_0, i, j) v_n(p, \omega_0). \end{array} \tag{4.2}$$

There is an $n_0 \geq N$ such that $|v_{n_0}(p, \omega_0) - \underline{w}(p, \omega_0)| \leq \eta$. Equations (4.1) and (4.2) imply that for any n such that $|v_n(p, \omega_0) - \underline{w}(p, \omega_0)| \leq \eta$, and for any y, there is an x satisfying $T_{xy}(\frac{1}{n+1}, v_n)(p, \omega_0) \geq v_n(p, \omega_0) + \min(\eta, \frac{\varepsilon}{2n})$. Hence $v_{n+1}(p, \omega_0) \geq v_n(p, \omega_0) + \min(\eta, \frac{\varepsilon}{2n})$.

Therefore there is an $n_1 \geq N$ such that $v_{n_1}(p, \omega_0) \geq \underline{w}(p, \omega_0) + \eta$. Equations (4.1) and (4.2) imply that for any n such that $v_n(p, \omega_0) \geq \underline{w}(p, \omega_0) + \eta$, and for any y, there is an x satisfying, $T_{xy}(\frac{1}{n+1}, v_n)(p, \omega_0) \geq \underline{w}(p, \omega_0) + \eta$. Hence $v_{n+1}(p, \omega_0) \geq \underline{w}(p, \omega_0) + \eta$. Thus for any $n \geq n_1$, $v_n(p, \omega_0) \geq \underline{w}(p, \omega_0) + \eta$, which contradicts the definition of $\underline{w}$. □

LEMMA 4.2. *$\underline{w} \geq v$.*

Proof. This is a corollary of the previous lemma and Proposition 2.4 applied for the function $\underline{w}$. □

LEMMA 4.3. *$w \leq v$.*

Proof. Let us denote by p_n an element of $\arg\max\{v_n(p, \omega_0) - v(p, \omega_0) - \delta(p)\}$. Denote by d_n the quantity $v_n(p_n, \omega_0) - v(p_n, \omega_0) - \delta(p_n) - \bar{\varepsilon}(\omega_n)$. Notice that Proposition 1.5 implies that for all p, $\omega^* \in \Omega^*$, $v(p, \omega^*) = w(p, \omega^*)$. We prove the claim by induction on the cardinality of K. If *card* $(K) = 1$, it is a consequence of [5]. Assume it is true for *card* $(K) = \mathcal{K} - 1$, and let *card* $(K) = \mathcal{K}$. The induction hypothesis implies that if $p_{s(n)}$ is any converging subsequence of p_n, then its limit is in the interior of $\Delta(K)$. Thus Proposition 3.1 (with $\lambda_n = 1/n$) implies that there is an integer N such that for all $n \geq N$,

$$T\left(\frac{1}{n}, v + \delta + \bar{\varepsilon}\right)(p_n, \omega_0) < v(p_n, \omega_0) + \delta(p_n) + \varepsilon. \tag{4.3}$$

In the case of finitely repeated games, we have to consider this case that was ruled out for discounted games by Lemma 3.4. Inequality 4.3 implies

$$\begin{aligned} T\left(\frac{1}{n+1}, v + \delta + \bar{\varepsilon}\right)(p_{n+1}, \omega_0) &\leq v(p_{n+1}, \omega_0) + \delta(p_{n+1}) + \varepsilon \\ &\leq v_{n+1}(p_{n+1}, \omega_0) - d_{n+1} \\ &\leq T\left(\frac{1}{n+1}, v_n\right)(p_{n+1}, \omega_0) - d_{n+1}. \end{aligned}$$

Hence, in any case, $(n+1)\max(d_{n+1},0) \leq n\max(d_n,0)$. Thus there is a constant $C>0$ such that for all $n \geq N$, $d_n \leq \frac{C}{n}$. The result follows. □

These results imply that for any uniform limit w of a subsequence of (v_n), we have $w=v$; the result follows.

5. Concluding remarks.

1) The proof provided above gives no indication about the speed of convergence of v_n and v_λ to their limit.
2) There is no immediate generalization of the proof to the case where the transitions q depend on the parameter k. Indeed, in this case the class of nonrevealing strategies may change, and all arguments may become much more involved.
3) The limit of v_n and the limit of v_λ are equal in the situation under concern in this paper. This property is conjectured to generalize to all stochastic games with incomplete information and finite state and action sets: in [6], Lehrer and Sorin proved that for one player games the uniform convergence of v_λ is equivalent to that of v_n.
4) It is conjectured by Sorin [15] and Mertens [7] that the maxmin of infinitely repeated games with lack of information on one side exists and is also equal to v. In the case of big match games [14, 15], Sorin proved that this conjecture is satisfied. The present proof gives no insight about this conjecture, though the mappings T and ϕ, by separating the study of the absorbing and nonabsorbing payoffs, seem to be an appropriate tool to study the maxmin of the infinitely repeated game.

Acknowledgments. This work is part of my Ph.D. and has benefited from the constant help and support of my advisor Sylvain Sorin. I also owe a lot of thanks to Nicolas Vieille for the very fruitful discussions we had on the subject, the proof, and its presentation.

REFERENCES

[1] R. J. Aumann and M. Maschler, with the collaboration of R.E. Stearns, *Repeated Games with Incomplete Information*, MIT Press, Cambridge, MA, 1995.

[2] T. Bewley and E. Kohlberg, *The asymptotic theory of stochostic games*, Math. Oper. Res., 1 (1976), pp. 197–208.

[3] D. Blackwell and T. Ferguson, *The big match,* Ann. Math. Statist., 33 (1968), pp. 882–886.

[4] B. De Meyer, *Repeated games and partial differential equations*, Math. Oper. Res., 21 (1996), pp. 209–236.

[5] E. Kohlberg, *Repeated games with absorbing states*, Ann. Statist., 2 (1974), pp. 724–738.

[6] E. Lehrer and S. Sorin, *A uniform tauberian theorem in dynamic programming*, Math. Oper. Res., 17 (1992), pp. 303–307.

[7] J.-F. Mertens, *Repeated games*, in Proceedings of the International Congress of Mathematicians, Berkeley, CA, 1986, AMS, Providence, RI, 1987, pp. 1528–1577.

[8] J.-F. Mertens and A. Neyman, *Stochastic games*, Internat. J. Game Theory, 10 (1981), pp. 53–56.

[9] J.-F. Mertens and S. Zamir, *The value of two person zero sum repeated games with lack of information on both sides*, Internat. J. Game Theory, 1 (1971–72), pp. 39–64.

[10] J.-F. Mertens, S. Sorin, and S. Zamir, *Repeated Games, Parts A, B, and C*, CORE D.P. 9420–9422, 1994.

[11] H.D. Mills, *Marginal values of matrix games and linear programs*, in Linear Inequalities and Related Systems, H.W. Kuhn and A.W. Tucker, eds., Ann. Math. Study 38, Princeton University Press, Princeton NJ, 1956, pp. 183–194.

[12] D. Rosenberg and S. Sorin, *An operator approach to zero sum repeated games*, Cahiers du Laboratoire d'Econométrie de l'Ecole Polytechnique 494, 1999, Israel J. Math., to appear.

[13] L.S. Shapley, *Stochastic games*, Proc. Natl. Acad. Sci. USA., 39 (1953), pp. 1095–1100.

[14] S. Sorin, *Big match with lack of information on one side* (*part*1), Internat. J. Game Theory, 13 (1984), pp. 201–255.

[15] S. Sorin, *Big match with lack of information on one side* (*part*2), Internat. J. Game Theory, 14 (1985), pp. 173–204.

[16] S. Zamir, *Repeated games of incomplete information: Zero-sum*, in Handbook of Game Theory with Economic Applications, R.J. Aumann and S. Hart, eds., North-Holland, Amsterdam, 1992, pp. 109–154.

SIAM J. Control Optim.
Vol. 39, No. 1, pp. 226–249

LYAPUNOV CHARACTERIZATIONS OF INPUT TO OUTPUT STABILITY[*]

EDUARDO SONTAG[†] AND YUAN WANG[‡]

Abstract. This paper presents necessary and sufficient characterizations of several notions of input to output stability. Similar Lyapunov characterizations have been found to play a key role in the analysis of the input to state stability property, and the results given here extend their validity to the case when the output, but not necessarily the entire internal state, is being regulated.

Key words. Lyapunov functions, output stability, ISS, robust control

AMS subject classifications. 93D05, 93D20, 93D09, 34D20

PII. S0363012999350213

1. Introduction.

This paper concerns itself with systems with outputs of the general form

$$\dot{x}(t) = f(x(t), u(t)), \quad y(t) = h(x(t)), \tag{1.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ are both locally Lipschitz continuous, $f(0,0) = 0$, and $h(0) = 0$. In [19] (see also [17]), the authors introduced several notions of output stability for such systems. All these notions serve to formalize the idea of a "stable" dependence of outputs y upon inputs (which may be thought of as disturbances, actuator or measurement errors, or regulation signals). They differ in the precise formulation of the decay estimates and the overshoot, or transient behavior, characteristics of the output. Among all of them, the one of most interest is probably the one singled out for the name *input to output stability*, or IOS, for short.

Our main theorem in this paper provides a necessary and sufficient characterization of the IOS property in terms of Lyapunov functions. In the process of obtaining this characterization, we derive as well corresponding results for the variants of IOS discussed in [19]. (The relationships between those variants, shown in [19], play a role in our proofs, but otherwise the two papers are independent of each other.)

In the very special case when $y = x$, our concepts all reduce to the input to state stability (ISS) property. Much of ISS control design (cf. [2, 3, 4, 5, 6, 7, 9, 10, 13, 14, 15, 20]) relies upon the Lyapunov characterizations first obtained in [12, 16]. Thus, it is reasonable to expect a similar impact from the results given here for the more general case.

In order to review the different i/o stability concepts, let us make the following notational conventions. Euclidean norms will be denoted as $|x|$, and $\|u\|$ denotes the L^m_∞-norm (possibly infinite) of an input u (i.e., a measurable and locally essentially bounded function $u : \mathcal{I} \to \mathbb{R}^m$, where $\mathcal{I}$ is a subinterval of $\mathbb{R}$ which contains the origin;

[*] Received by the editors January 15, 1999; accepted for publication (in revised form) April 16, 2000; published electronically August 24, 2000.
http://www.siam.org/journals/sicon/39-1/35021.html

[†] Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (sontag@control.rutgers.edu). The research of this author was supported in part by U.S. Air Force grant F49620-98-1-0242.

[‡] Department of Mathematics, Florida Atlantic University, Boca Raton, FL 33431 (ywang@math.fau.edu). The research of this author was supported by NSF grants DMS-9457826 and DMS-9403924.

if we do not specify the domain $\mathcal{I}$ of an input u, we mean implicitly that $\mathcal{I} = \mathbb{R}_{\geq 0}$). For each initial state $\xi \in \mathbb{R}^n$ and input u, we let $x(\cdot, \xi, u)$ be the unique maximal solution of the initial value problem $\dot{x} = f(x, u)$, $x(0) = \xi$, and write the corresponding output function $h(x(t, \xi, u))$ simply as $y(\cdot, \xi, u)$. Given a system with control-value set $\mathbb{R}^m$, we often consider the same system but with controls restricted to take values in some subset $\Omega \subseteq \mathbb{R}^m$; we use $\mathcal{M}_\Omega$ for the set of all such controls. As usual, by a $\mathcal{K}$ function we mean a function $\gamma : [0, \infty) \to [0, \infty)$ that is strictly increasing and continuous and satisfies $\gamma(0) = 0$, by a $\mathcal{K}_\infty$ function one that is in addition unbounded, and we let $\mathcal{KL}$ be the class of functions $[0, \infty)^2 \to [0, \infty)$ which are of class $\mathcal{K}$ on the first argument and decrease to zero on the second argument. When we state the various properties below, we always interpret the respective estimates as holding for all inputs u and for all initial states $\xi \in \mathbb{R}^n$.

Recall that a system is said to be *forward complete* if for every initial state ξ and input u, the solution $x(t, \xi, u)$ is defined for all $t \geq 0$.

The following four output stability properties were discussed in [19]. A forward complete system is:

- IOS, or *input to output stable*, if there exist a $\mathcal{KL}$-function β and a $\mathcal{K}$-function γ such that

  $$|y(t, \xi, u)| \;\leq\; \beta(|\xi|, t) + \gamma(\|u\|) \qquad \forall t \geq 0 \tag{1.2}$$

  (the term $\gamma(\|u\|)$ can be replaced by the norm of the restriction to past inputs $\gamma(\|u\|_{[0,t]})$, and the sum could be replaced by a “max” or two analogous terms);
- OLIOS, or *output-Lagrange input to output stable*, if it is IOS and, in addition, there exist some $\mathcal{K}$-functions σ_1, σ_2 such that

  $$|y(t, \xi, u)| \;\leq\; \max\{\sigma_1(|h(\xi)|), \sigma_2(\|u\|)\} \qquad \forall\, t \geq 0; \tag{1.3}$$
- SIIOS, or *state-independent input to output stable*, if there exist some $\beta \in \mathcal{KL}$ and some $\gamma \in \mathcal{K}$ such that

  $$|y(t, \xi, u)| \;\leq\; \beta(|h(\xi)|, t) + \gamma(\|u\|) \qquad \forall\, t \geq 0; \tag{1.4}$$
- ROS, or *robustly output stable*, if there are a smooth $\mathcal{K}_\infty$-function λ and a $\beta \in \mathcal{KL}$ such that the system

  $$\dot{x} = g(x, d) := f(x, d\lambda(|y|)), \quad y = h(x), \tag{1.5}$$

  is forward complete, and the estimate

  $$|y_\lambda(t, \xi, d)| \;\leq\; \beta(|\xi|, t) \qquad \forall\, t \geq 0 \tag{1.6}$$

  holds for all $d \in \mathcal{M}_\mathcal{B}$, where $\mathcal{B} = \{|\mu| \leq 1\} \subset \mathbb{R}^m$, and where $y_\lambda(\cdot, \xi, d)$ denote the output function of system (1.5).

The last concept corresponds to the preservation of output stability under output feedback with “robustness margin” λ. It was shown in [19] that SIIOS $\Rightarrow$ OLIOS $\Rightarrow$ IOS $\Rightarrow$ ROS, and no converses hold. It was also remarked in section 2.2 of [19] that the OLIOS property is equivalent to the existence of a $\mathcal{KL}$-function β and a $\mathcal{K}$-function γ such that the estimate

$$|y(t, \xi, u)| \leq \beta\left(|h(\xi)|, \, \frac{t}{1 + \rho(|\xi|)}\right) + \gamma(\|u\|) \qquad \forall\, t \geq 0$$

holds for all trajectories of the system. We now introduce the associated Lyapunov concepts.

Definition 1.1. *With respect to the system* (1.1), *a smooth function* $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ *is:*

- *an* IOS*-Lyapunov function if there exist* $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ *such that*

$$\alpha_1(|h(\xi)|) \leq V(\xi) \leq \alpha_2(|\xi|) \qquad \forall\, \xi \in \mathbb{R}^n \tag{1.7}$$

and there exist $\chi \in \mathcal{K}$ *and* $\alpha_3 \in \mathcal{KL}$ *such that*

$$V(\xi) \geq \chi(|\mu|) \Rightarrow DV(\xi)f(\xi,\mu) \leq -\alpha_3(V(\xi), |\xi|) \quad \forall\, \xi,\ \forall\, \mu, \tag{1.8}$$

- *an* OLIOS*-Lyapunov function if there exist* $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ *such that*

$$\alpha_1(|h(\xi)|) \leq V(\xi) \leq \alpha_2(|h(\xi)|) \qquad \forall\, \xi \in \mathbb{R}^n \tag{1.9}$$

and there exist $\chi \in \mathcal{K}$ *and* $\alpha_3 \in \mathcal{KL}$ *such that* (1.8) *holds,*

- *an* SIIOS*-Lyapunov function if there exist* $\chi \in \mathcal{K}$ *and* $\alpha_3 \in \mathcal{K}$ *such that*

$$V(\xi) \geq \chi(|\mu|) \Rightarrow DV(\xi)f(\xi,\mu) \leq -\alpha_3(V(\xi)) \quad \forall\, \xi,\ \forall \mu \tag{1.10}$$

and there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ *such that* (1.9) *holds,*

- *an* ROS*-Lyapunov function if there exist* $\chi \in \mathcal{K}$ *and* $\alpha_3 \in \mathcal{KL}$ *such that*

$$|h(\xi)| \geq \chi(|\mu|) \ \Rightarrow DV(\xi)f(\xi,\mu) \leq -\alpha_3(V(\xi), |\xi|) \quad \forall\, \xi,\ \forall\, \mu \tag{1.11}$$

and there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ *such that* (1.7) *holds.*

Observe that, if an estimate (1.7) holds, then (1.11) is implied by (1.8) in the sense that if χ and α_1 are as in the former, then $\widetilde{\chi} := \alpha_1^{-1} \circ \chi$ can be used as "χ" for the latter. Note also that, provided that (1.9) holds, condition (1.8) is equivalent to the existence of $\chi \in \mathcal{K}$ and $\alpha_3 \in \mathcal{KL}$ so that

$$|h(\xi)| \geq \chi(|\mu|) \ \Rightarrow \ DV(\xi)f(\xi,\mu) \leq -\alpha_3(V(\xi), |\xi|).$$

Our main results can be summarized as follows. We say that system (1.1) is *uniformly bounded input bounded state* stable, and write UBIBS for short, if it is forward complete and, for some function σ of class $\mathcal{K}$, the following estimate holds for all solutions:

$$|x(t,\xi,u)| \ \leq \ \max\{\sigma(|\xi|),\ \sigma(\|u\|)\} \qquad \forall t \geq 0. \tag{1.12}$$

Theorem 1.2. *A* UBIBS *system is:*

1. IOS *if and only if it admits an* IOS*-Lyapunov function;*
2. OLIOS *if and only if it admits an* OLIOS*-Lyapunov function;*
3. ROS *if and only if it admits an* ROS*-Lyapunov function; and*
4. SIIOS *if and only if it admits an* SIIOS*-Lyapunov function.*

The proofs are provided in section 4.

2. Remarks on rates of decrease. In properties (1.8) and (1.11), the decay rate of $V(x(t))$ depends on the state and on the value of $V(x(t))$. The main role of α_3 is to allow for slower convergence if $V(x(t))$ is very small or if $x(t)$ is very large. We first note two simplifications.

REMARK 2.1. Inequality (1.8) holds for some $\alpha_3 \in \mathcal{KL}$ if and only if there exist $\mathcal{K}$-functions κ_1, κ_2 such that

$$V(\xi) \geq \chi(|\mu|) \;\Rightarrow\; DV(\xi)f(\xi,\mu) \leq -\frac{\kappa_1(V(\xi))}{1+\kappa_2(|\xi|)} \tag{2.1}$$

for all $\xi \in \mathbb{R}^n$ and all $\mu \in \mathbb{R}^m$. This follows from Lemma A.2, proved in the appendix. A similar remark applies to (1.11).

REMARK 2.2. Suppose V is an IOS-Lyapunov function for the system satisfying (1.7) with some $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ and satisfying (2.1) with some $\chi, \kappa_1, \kappa_2 \in \mathcal{K}$. By the proof of Lemma 11 together with Lemma 12 in [14], one sees that there exists a C^1 $\mathcal{K}_\infty$-function ρ such that $\rho'(s)\kappa_1(s) \geq \rho(s)$ for all $s \geq 0$. Let $W = \rho \circ V$. Then W is a C^1 function satisfying the following:

$$\rho(\alpha_1(|h(\xi)|)) \leq W(\xi) \leq \rho(\alpha_2(|\xi|)) \qquad \forall\, \xi \in \mathbb{R}^n,$$

and

$$W(\xi) \geq \chi_1(|\mu|) \;\Rightarrow\; DW(\xi)f(\xi,\mu) \leq -\frac{W(\xi)}{1+\kappa_2(|\xi|)} \tag{2.2}$$

for all $\xi \in \mathbb{R}^n$ and all $\mu \in \mathbb{R}^m$, where $\chi_1 = \rho \circ \chi \in \mathcal{K}$. This shows that if a system admits an IOS-Lyapunov function, then it admits one satisfying inequality (2.2). A similar remark applies to (1.11).

Obviously, a function which satisfies a decay estimate of the stronger form

$$V(\xi) \geq \chi(|\mu|) \;\Rightarrow\; DV(\xi)f(\xi,\mu) \leq -\alpha(V(\xi)) \tag{2.3}$$

for some $\chi, \alpha \in \mathcal{K}$ is in particular an IOS Lyapunov function. It is thus natural to ask if there always exists, for an IOS system, a function with this stronger property. We now show, by means of an example, that such functions do not in general exist. Consider for that purpose the following two-dimensional single-input system:

$$\dot{x}_1 = 0, \quad \dot{x}_2 = -\frac{2x_2+u}{1+x_1^2}, \quad y = x_2. \tag{2.4}$$

This system is IOS, because with $V(x) := x_2^2$, it holds that

$$V(\xi) \geq \mu^2 \;\Rightarrow\; DV(\xi)f(\xi,\mu) = -2x_2\,\frac{2x_2+u}{1+x_1^2} \leq -\frac{2V(\xi)}{1+x_1^2}.$$

Namely, V is an IOS-Lyapunov function for the system.

Suppose that system (2.4) would admit an IOS-Lyapunov function W with a decay estimate as in (2.3), i.e., there exist some $\chi, \alpha \in \mathcal{K}$ such that

$$W(\xi) \geq \chi(|\mu|) \;\Rightarrow\; DW(\xi)f(\xi,\mu) \leq -\alpha(W(\xi)). \tag{2.5}$$

Without loss of generality, we may assume that $\chi \in \mathcal{K}_\infty$. In particular, we have that

$$DW(\xi)f(\xi, -\chi^{-1}(W(\xi))) \leq -\alpha(W(\xi)) \tag{2.6}$$

for all $\xi \in \mathbb{R}^2$. Fix any $\xi_1 \in \mathbb{R}$, and consider the one-dimensional differential equation

$$\dot{x}_2 = -\frac{2x_2 - \chi^{-1}(W(\xi_1, x_2))}{1+\xi_1^2}. \tag{2.7}$$

Since $W(\xi_1, x_2(t)) \to 0$ (because of (2.6)) and as $\alpha_1(|\xi_2|) \leq W(\xi_1, \xi_2)$ for all ξ (for some $\alpha_1 \in \mathcal{K}$), it follows that $x_2(t) \to 0$ as $t \to \infty$. This implies that $W(\xi_1, \xi_2) < \chi(2\xi_2)$ for all $\xi_1 \in \mathbb{R}$ and $\xi_2 > 0$. Together with (2.5), this implies that there exists some $\beta \in \mathcal{KL}$ such that, for every trajectory of (2.4) with $u(t) \equiv 0$, it holds that

$$|x_2(t)| \leq \beta(|x_2(0)|, t)$$

for all $\xi = (x_1(0), x_2(0))$ such that $x_2(0) > 0$. This is impossible, as it can be seen that, when $u(t) \equiv 0$, $x_2(t) = x_2(0)e^{-2t/(1+(x_1(0))^2)}$, whose decay rate depends on both $x_2(0)$ and $x_1(0)$.

Observe that, if we let $U(\xi_1, \xi_2) := [(1+\xi_1^2)\,|\xi_2|]^{(1+\xi_1^2)}$, then one obtains the following estimate:

$$|\xi_2| \geq |\mu| \;\Rightarrow\; DU(\xi)f(\xi,\mu) \leq -U(\xi) \tag{2.8}$$

for all $\xi_1 \in \mathbb{R}$, $\xi_2 \neq 0$, and all $\mu \in \mathbb{R}$. (The function U is not smooth on the set where $U(\xi) = 0$, but, using a routine smoothing argument, one may easily modify U to get a smooth Lyapunov function.) This U is not an example of a W as here (which, in any case, we know cannot exist), because (2.8) only means that U is an ROS-Lyapunov function, not necessarily an IOS-Lyapunov function (since the comparison is between $|\xi_2|$ and $|\mu|$ rather than between a function of U and $|\mu|$).

Finally, we observe that property (1.8) in the IOS-Lyapunov definition may be rephrased as follows:

$$V(\xi) > \widetilde{\chi}(|\mu|) \;\Rightarrow\; DV(\xi)f(\xi,\mu) < 0 \qquad \forall \xi \in \mathbb{R}^n,\ \forall \mu \in \mathbb{R}^m, \tag{2.9}$$

where $\widetilde{\chi}(s) := \rho\,\chi(s)$ (for any arbitrary chosen $\rho \in (0,1)$). This statement is obviously implied by (1.8). Conversely, if V satisfies this property, then there is an $\alpha \in \mathcal{KL}$ so that (1.8) holds; this follows from Lemma A.5 given in the appendix.

3. Uniform stability notions.

There is a key technical result which underlies the proofs of all our converse Lyapunov theorems. It requires yet another set of definitions, which correspond to stability uniformly on all "disturbance" inputs.

Definition 3.1. *A system* (1.1) *is* uniformly output stable *with respect to inputs in $\mathcal{M}_\Omega$, where Ω is a compact subset of $\mathbb{R}^m$, if*

- *it is forward complete, and*
- *there exists a $\mathcal{KL}$-function β such that*

$$|y(t,\xi,u)| \leq \beta(|\xi|, t) \qquad \forall t \geq 0 \tag{3.1}$$

holds for all u and all $\xi \in \mathbb{R}^n$.

If, in addition, there exists $\sigma \in \mathcal{K}$ such that

$$|y(t,\xi,u)| \leq \sigma(|h(\xi)|) \qquad \forall t \geq 0 \tag{3.2}$$

holds for all trajectories of the system with $u \in \mathcal{M}_\Omega$, then the system is output-Lagrange uniformly output stable with respect to inputs in $\mathcal{M}_\Omega$. Finally, if one strengthens (3.1) *to*

$$|y(t,\xi,u)| \leq \beta(|h(\xi)|, t) \qquad \forall t \geq 0 \tag{3.3}$$

holding for all trajectories of the system with $u \in \mathcal{M}_\Omega$, then the system is state-independent uniformly output stable with respect to inputs in $\mathcal{M}_\Omega$.

Theorem 3.2. *Let Ω be a compact subset of $\mathbb{R}^m$, and suppose that a system* (1.1) *is uniformly output stable with respect to inputs in $\mathcal{M}_\Omega$. Then the system admits a smooth Lyapunov function V satisfying the following properties:*

- *there exist* $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ *such that*

$$\alpha_1(|h(\xi)|) \le V(\xi) \le \alpha_2(|\xi|) \qquad \forall\, \xi \in \mathbb{R}^n; \tag{3.4}$$

- *there exists* $\alpha_3 \in \mathcal{KL}$ *such that*

$$DV(\xi)f(\xi,\mu) \le -\alpha_3(V(\xi),|\xi|) \qquad \forall\, \xi \in \mathbb{R}^n,\ \forall\, \mu \in \Omega. \tag{3.5}$$

Moreover, if the system is output-Lagrange uniformly output stable with respect to inputs in $\mathcal{M}_\Omega$*, then* (3.4) *can be strengthened to*

$$\alpha_1(|h(\xi)|) \le V(\xi) \le \alpha_2(|h(\xi)|) \qquad \forall\, \xi \tag{3.6}$$

for some $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$*. Finally, if the system is state-independent uniformly output stable with respect to inputs in* $\mathcal{M}_\Omega$*, then* (3.4) *can be strengthened to* (3.6) *and also* (3.5) *can be strengthened to*

$$DV(\xi)f(\xi,\mu) \le -\alpha_4(V(\xi)) \qquad \forall\, \xi \in \mathbb{R}^n,\ \forall\, \mu \in \Omega \tag{3.7}$$

for some $\alpha_4 \in \mathcal{K}$*.*

The proof of this theorem will be postponed until section 4.5.

4. Proof of Theorem 1.2.

In the proofs of the various parts of the theorem, we need the following small gain lemma for output-Lagrange stability (see Lemma 8 of [19]).

LEMMA 4.1. *For every system which satisfies* (1.3)*, there exist a* $\mathcal{K}$*-function* σ *and a* $\mathcal{K}_\infty$*-function* λ *such that the system*

$$\dot{x} = f(x,\, d\lambda(|y|)), \quad y = h(x), \tag{4.1}$$

where $d \in \mathcal{M}_\mathcal{B}$*, is forward complete, and*

$$|y_\lambda(t,\xi,d)| \le \sigma(|h(\xi)|) \tag{4.2}$$

for all $\xi \in \mathbb{R}^n$*, all* $t \ge 0$*, and all* $d \in \mathcal{M}_\mathcal{B}$*.*

4.1. Proof of Theorem 1.2, part 1.

Necessity. Consider an OLIOS system (1.1). By Lemma 4.1, there exist a smooth $\mathcal{K}_\infty$-function λ_1 and a $\mathcal{K}$-function σ such that the system

$$\dot{x} = f(x, d\lambda_1(|y|)), \qquad y = h(x),$$

where $d \in \mathcal{M}_\mathcal{B}$, is forward complete, and (4.2) holds.

Since the system is OLIOS, and, in particular, IOS, and since, as shown in [19], any IOS system is necessarily also ROS, there exists some smooth $\mathcal{K}_\infty$-function λ_2 such that the system

$$\dot{x} = f(x, d\lambda_2(|y|)),\ y = h(x), \tag{4.3}$$

where $d \in \mathcal{M}_\mathcal{B}$, is forward complete, and there exists some $\beta \in \mathcal{KL}$ such that, for all trajectories $x_{\lambda_2}(t,\xi,u)$ with the output functions $y_{\lambda_2}(t,\xi,u)$, it holds that

$$\left|y_{\lambda_2}(t,\xi,d)\right| \le \beta(|\xi|\,,t) \qquad \forall\, t \ge 0,\ \forall\, \xi \in \mathbb{R}^n,\ \forall\, d \in \mathcal{M}_\mathcal{B}.$$

Let $\lambda_3(s) = \min\{\lambda_1(s), \lambda_2(s)\}$, and let $\lambda(\cdot)$ be any smooth $\mathcal{K}_\infty$-function so that $\lambda(s) \leq \lambda_3(s)$ for all s. Then, the system

$$\dot{x} = f(x, d\lambda(|y|)), \quad y = h(x), \tag{4.4}$$

where $d \in \mathcal{M}_{\mathcal{B}}$, is forward complete, and it holds that

$$\left|y_\lambda(t,\xi,d)\right| \leq \beta(|\xi|, t) \quad \text{and} \quad \left|y_\lambda(t,\xi,d)\right| \leq \sigma(|h(\xi)|) \qquad \forall\, t \geq 0.$$

Applying Theorem 3.2, one sees that there exists some smooth function V such that:

- there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that

$$\alpha_1(|h(\xi)|) \leq V(\xi) \leq \alpha_2(|h(\xi)|) \qquad \forall\, \xi; \tag{4.5}$$

- there exist some $\alpha_3 \in \mathcal{KL}$ such that

$$DV(\xi)f(\xi, \nu\lambda(|h(\xi)|)) \leq -\alpha_3(V(\xi), |\xi|) \tag{4.6}$$

for all $\xi \in \mathbb{R}^n$ and all $|\nu| \leq 1$.

It then follows that

$$DV(\xi)f(\xi,\mu) \leq -\alpha_3(V(\xi), |\xi|)$$

whenever $|\mu| \leq \lambda(|h(\xi)|)$, or, equivalently, whenever $|h(\xi)| \geq \lambda^{-1}(|\mu|)$. Let $\chi = \alpha_2^{-1} \circ \lambda^{-1}$. Then one has

$$V(\xi) \geq \chi(|\mu|) \Rightarrow DV(\xi)f(\xi,\mu) \leq -\alpha_3(V(\xi), |\xi|)$$

for all ξ and all μ. Hence, V is an OLIOS-Lyapunov function for the system. □

Sufficiency. Let V be an OLIOS-Lyapunov function for system (1.1). Let $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that (1.9) holds. By (1.8), and arguing as in Remark 2.1, one also knows that there exist some κ_1 and $\kappa_2 \in \mathcal{K}_\infty$ such that

$$V(\xi) \geq \chi(|\mu|) \Rightarrow DV(\xi)f(\xi,\mu) \leq -\frac{\kappa_1(V(\xi))}{1+\kappa_2(|\xi|)} \tag{4.7}$$

for all ξ and μ.

Let $\beta \in \mathcal{KL}$ be as in Lemma A.4 for the function κ_1. Pick any initial state ξ and any u. Let $x(t)$ and $y(t)$ denote the ensuing trajectory and output function, respectively. If for some $t_1 \geq 0$, $V(x(t_1)) \leq \chi(\|u\|)$, then $V(x(t)) \leq \chi(\|u\|)$ for all $t \geq t_1$. (Proof: pick any $\varepsilon > 0$. If $t_2 := \inf\{t > t_1 \,|\, V(x(t)) > \chi(\|u\|) + \varepsilon\}$ is finite, then $V(x(t)) > \chi(\|u\|)$ for all t in some left neighborhood of t_2, so $DV(x(t))/dt < 0$ and $V(x(t)) > V(x(t_2))$ for such t, contradicting its minimality. As ε was arbitrary, the claim follows.) Now let

$$\widetilde{t} = \inf\{t \geq 0 : \ V(x(t)) \leq \chi(\|u\|)\}$$

with the understanding that $\widetilde{t} = \infty$ if $V(x(t)) > \chi(\|u\|)$ for all $t \geq 0$. Then

$$V(x(t)) \leq \chi(\|u\|) \qquad \forall\, t \geq \widetilde{t}, \tag{4.8}$$

and on $[0, \widetilde{t})$, it holds that

$$\frac{d}{dt}V(x(t)) \leq -\frac{\kappa_1(V(x(t)))}{1+\kappa_2(|x(t)|)}.$$

Since the system is UBIBS, there exists some σ such that (1.12) holds. Hence,

$$\frac{d}{dt}V(x(t)) \leq -\frac{\kappa_1(V(x(t)))}{1+\max\{\widetilde{\kappa}_2(|\xi|),\ \widetilde{\kappa}_2(\|u\|)\}}$$

for all $t \in [0, \widetilde{t})$, where $\widetilde{\kappa}_2 = \kappa_2 \circ \sigma$. It then follows Lemma A.4 that

$$V(x(t)) \leq \beta\left(V(\xi),\ \frac{t}{1+\max\{\widetilde{\kappa}_2(|\xi|),\ \widetilde{\kappa}_2(\|u\|)\}}\right)$$

for all $t \in [0, \widetilde{t})$.

Let $v_0(s) = \max_{|\xi|\leq s} V(\xi)$. Then v_0 is nondecreasing, $v_0(0) = 0$, and $V(\xi) \leq v_0(|\xi|)$. Note then that

$$\begin{aligned}
&\beta\left(V(\xi),\ \frac{t}{1+\max\{\widetilde{\kappa}_2(|\xi|),\ \widetilde{\kappa}_2(\|u\|)\}}\right)\\
&\qquad \leq \max\left\{\beta\left(V(\xi),\ \frac{t}{1+\widetilde{\kappa}_2(|\xi|)}\right),\ \beta\left(v_0(\|u\|),\ \frac{t}{1+\widetilde{\kappa}_2(\|u\|)}\right)\right\}\\
&\qquad \leq \max\left\{\beta\left(V(\xi),\ \frac{t}{1+\widetilde{\kappa}_2(|\xi|)}\right),\ \beta\left(v_0(\|u\|),\ 0\right)\right\}
\end{aligned}$$

(consider two cases: $|\xi| \geq \|u\|$ and $|\xi| \leq \|u\|$). This shows that

$$V(x(t)) \leq \max\left\{\beta\left(V(\xi),\ \frac{t}{1+\widetilde{\kappa}_2(|\xi|)}\right),\ \widetilde{\beta}_0(\|u\|)\right\}$$

for all $t \in [0, \widetilde{t})$, where $\widetilde{\beta}_0(s) = \beta(v_0(s), 0)$. Combining this with (4.8), one sees that

$$V(x(t)) \leq \max\left\{\beta\left(V(\xi),\ \frac{t}{1+\widetilde{\kappa}_2(|\xi|)}\right),\ \widetilde{\gamma}(\|u\|)\right\} \tag{4.9}$$

for all $t \geq 0$, where $\widetilde{\gamma}(s) = \widetilde{\beta}(s) + \chi(s)$. Using the fact that $|h(\xi)| \leq \alpha_1^{-1}(V(\xi))$, we conclude that

$$|y(t)| \leq \max\left\{\widetilde{\beta}\left(|h(\xi)|,\frac{t}{1+\widetilde{\kappa}_2(|\xi|)}\right),\ \gamma(\|u\|)\right\} \tag{4.10}$$

for all $t \geq 0$, where $\widetilde{\beta}(s,r) = \alpha_1^{-1}(\beta(\alpha_2(s), r))$, and $\gamma(s) = \alpha_1^{-1}(\widetilde{\gamma}(s))$. □

4.2. Proof of Theorem 1.2, part 2.

Necessity. Consider an IOS system (1.1). By Theorem 1 in [19], there exist some locally Lipschitz map h_0 and $\chi \in \mathcal{K}_\infty$ with the property that $h_0(\xi) \geq \chi(|h(\xi)|)$ such that the system

$$\dot{x} = f(x,u), \quad y = h_0(x) \tag{4.11}$$

is OLIOS. By part 1 of this theorem, system (4.11) admits an OLIOS-Lyapunov function V. This means that there exist $\alpha_1, \alpha_2, \rho \in \mathcal{K}_\infty$, and $\alpha_3 \in \mathcal{KL}$ such that

$$\alpha_1(|h_0(\xi)|) \leq V(\xi) \leq \alpha_2(|h_0(\xi)|) \qquad \forall\, \xi \in \mathbb{R}^n,$$

and

$$V(\xi) \geq \rho(|\mu|) \Longrightarrow DV(\xi)f(\xi,\mu) \leq -\alpha_3(V(\xi), |\xi|).$$

To show that V is an IOS-Lyapunov function, it remains only to show that $V(\xi) \geq \widetilde{\alpha}_1(|h(\xi)|)$ for some $\widetilde{\alpha}_1 \in \mathcal{K}_\infty$. But this follows immediately from the fact that $|h(\xi)| \leq \chi^{-1}(h_0(\xi))$. So one can let $\widetilde{\alpha}_1 := \alpha_1 \circ \chi$. Hence, V is indeed an IOS-Lyapunov function for system (1.1). □

Sufficiency. Let V be an IOS-Lyapunov function for system (1.1). From the proof of part 1 of Theorem 1.2 (sufficiency), one can see that if V satisfies (4.7) for some $\chi, \kappa_1, \kappa_2 \in \mathcal{KL}$, then there exist $\widetilde{\beta} \in \mathcal{KL}, \widetilde{\kappa}_2, \ \widetilde{\gamma} \in \mathcal{K}_\infty$ such that (4.9) holds. This means that the system

$$\dot{x} = f(x, u), \quad y = V(x)$$

is OLIOS. Since $V(x) \geq \alpha_1(|h(\xi)|)$ for some $\alpha_1 \in \mathcal{K}_\infty$, it follows that system (1.1) is IOS. □

4.3. Proof of Theorem 1.2, part 3.

Necessity. Since the system (1.1) is ROS, there is a smooth $\mathcal{K}_\infty$-function λ such that system (1.5) is forward complete, and (1.6) holds for the corresponding system (1.5). That is, system (1.5) is uniformly output stable. By Theorem 3.2, system (1.5) admits a smooth Lyapunov function V satisfying (3.4) and

$$DV(\xi)f(\xi, \mu\lambda(|y|)) \leq -\alpha_3(V(\xi), |\xi|) \qquad \forall\, \xi \in \mathbb{R}^n,\ \forall\, |\mu| \leq 1$$

for some $\alpha_3 \in \mathcal{KL}$. This is equivalent to

$$|y| \geq \lambda^{-1}(|\nu|) \ \Rightarrow\ DV(\xi)f(\xi, \nu) \leq -\alpha_3(V(\xi), |\xi|) \qquad \forall\, \xi \in \mathbb{R}^n,\ \forall\, |\nu| \in \mathbb{R}^m.$$

Hence, one concludes that V is an ROS-Lyapunov function for system (1.1).

Sufficiency. Let V be an ROS-Lyapunov function. As in Remark 2.1, there exist $\chi, \kappa_1, \kappa_2 \in \mathcal{K}_\infty$ such that

$$DV(\xi)f(\xi, \mu) \leq -\frac{\kappa_1(V(\xi))}{1 + \kappa_2(|\xi|)}$$

whenever $|h(\xi)| \geq \chi(|\mu|)$. Let $\lambda = \chi^{-1}$. Without loss of generality, one may assume that λ is smooth. (Otherwise, one can always replace λ by a smooth $\mathcal{K}_\infty$-function that is majorized by λ.) It then follows that

$$DV(\xi)f(\xi, \nu\lambda(|h(\xi)|)) \leq -\frac{\kappa_1(V(\xi))}{1 + \kappa_2(|\xi|)}$$

for all $\xi \in \mathbb{R}^n$ and all $|\nu| \leq 1$. This implies that for any trajectory $x_{\lambda}(t) = x_{\lambda}(t, \xi, d)$ of the system

$$\dot{x} = f(x, d\lambda(|y|)), \quad y = h(x),$$

where $d \in \mathcal{M}_{\mathcal{B}}$, it holds that

$$\frac{d}{dt}V(x_{\lambda}(t)) \leq -\frac{\kappa_1(V(x_{\lambda}(t)))}{1 + \kappa_2(|x_{\lambda}(t)|)} \tag{4.12}$$

for all $t \geq 0$. It follows immediately that $V(x_{\lambda}(t)) \leq V(\xi)$ for all $t \geq 0$. Since $V(\xi) \geq \alpha_1(|h(\xi)|)$ for some $\alpha_1 \in \mathcal{K}_\infty$, it follows that, for some $\sigma \in \mathcal{K}_\infty$,

$$|y_{\lambda}(t)| \leq \sigma(|\xi|) \qquad \forall\, t \geq 0. \tag{4.13}$$

Since the system is UBIBS, there exists some $\sigma_0 \in \mathcal{K}$ such that

$$\left|x_\lambda(t,\xi,d)\right| \le \max\{\sigma_0(|\xi|), \sigma_0(\|u_d\|)\} \qquad \forall\, t \ge 0,$$

where $u_d(t) = d(t)\lambda(|y(t)|)$. Combining this with (4.13), it follows that

$$\left|x_\lambda(t,\xi,d)\right| \le \widetilde{\sigma}(|\xi|) \qquad \forall\, t \ge 0,$$

where $\widetilde{\sigma}(s) = \max\{\sigma_0(s), \sigma_0(\lambda(\sigma(s)))\}$. Substituting this back into (4.12), one has

$$\frac{d}{dt}V(x_\lambda(t)) \le -\frac{\kappa_1(V(x_\lambda(t)))}{1+\kappa_3(|\xi|)} \qquad \forall\, t \ge 0,$$

where $\kappa_3(s) = \kappa_2(\widetilde{\sigma}(s))$. Again, by Lemma A.4, one knows that there exists some $\beta \in \mathcal{KL}$ (which depends only upon κ_1) such that

$$V(x_\lambda(t)) \le \beta\left(V(\xi),\ \frac{t}{1+\kappa_3(|\xi|)}\right) \qquad \forall\, t \ge 0.$$

Together with the fact that $|h(\xi)| \le \alpha_1^{-1}(V(\xi))$, this yields

$$\left|y_\lambda(t,\xi,d)\right| \le \widetilde{\beta}(|\xi|\,,t) \qquad \forall\, t \ge 0,$$

where $\widetilde{\beta}(s,r) = \alpha_1^{-1}[\beta(\alpha_2(s), t/(1+\kappa_3(s)))]$ is in $\mathcal{KL}$, and α_2 is any $\mathcal{K}_\infty$-function such that $V(\xi) \le \alpha_2(|\xi|)$ for all ξ. This shows that the system is ROS. □

4.4. Proof of Theorem 1.2, part 4.

Necessity. Assume that a UBIBS system (1.1) admits an estimate (1.4) for some $\beta \in \mathcal{KL}$ and some $\gamma \in \mathcal{K}$. Without loss of generality, one may assume that

$$|y(t,\xi,u)| \le \max\{\beta(|h(\xi)|\,,t),\ \gamma(\|u\|)\}.$$

Let $\sigma_1(s) = \beta(s,0)$, and let $\sigma_2(s) = \gamma(s)$. Note then that (1.3) holds. By Lemma 8 in [19], there exists some smooth $\mathcal{K}_\infty$-function such that the corresponding system (1.5) is forward complete, and it holds that

$$\sigma_2(|d(t)|\,\lambda(|y_\lambda(t,\xi,d)|)) \le \frac{1}{2}\,|h(\xi)|$$

for all $\xi \in \mathbb{R}^n$, all $t \ge 0$, and all $d \in \mathcal{M}_{\mathcal{B}}$. One then can show that for the system

$$\dot{x}(t) = f(x(t), d(t)\lambda(|y(t)|)), \quad y(t) = h(x(t)),$$

there exists $\widetilde{\beta} \in \mathcal{KL}$ so that, for all trajectories $x_\lambda(t,\xi,d)$, it holds that

$$\left|y_\lambda(t,\xi,d)\right| \le \widetilde{\beta}(|h(\xi)|\,,t)$$

for all $t \ge 0$. Applying the last part of Theorem 3.2, one sees that there exists V satisfying (3.6) for some $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ and

$$DV(\xi)f(\xi, \nu\lambda(|y(\xi)|)) \le -\alpha_3(V(\xi))$$

for all ξ and all $|\nu| \le 1$. This is equivalent to the existence of $\chi \in \mathcal{K}_\infty$ such that

$$V(\xi) \ge \chi(|\mu|) \Rightarrow DV(\xi)f(\xi,u) \le -\alpha_3(V(\xi)). \tag{4.14}$$

Sufficiency. It is routine to show that if there is a smooth function V satisfying (3.6) and (4.14), then the system admits an estimate of type (1.4). □

REMARK 4.1. Note that in all the proofs of the necessity implications of Theorem 1.2, the UBIBS property is not needed. That is, to show the existence of various Lyapunov functions for the corresponding stability properties, one does not need the UBIBS property. However, the UBIBS property is indeed required in the proofs of the sufficiency implications regarding the IOS, OLIOS, and the ROS properties. It is not hard to find examples where a system admits an IOS-, OLIOS-, or ROS-Lyapunov function, without satisfying the UBIBS property, and fails to be IOS, OLIOS, or ROS, respectively.

It should also be noticed that part 4 of Theorem 1.2 also holds for all forward complete systems (not necessarily UBIBS). Without the UBIBS assumption, this result recovers the converse Lyapunov theorem obtained in [12] for systems that are uniformly globally asymptotically stable with respect to closed invariant sets, when applied using as output the distance to a closed invariant set. In fact, part 4 of Theorem 1.2 yields a more general result than the one in [12]. Because of the techniques used in the proofs in [12], the systems were required to be backward complete. Due to part 4 of Theorem 1.2, it can be seen that the backward completeness assumption is redundant.

4.5. Proof of Theorem 3.2. Consider the system

$$\dot{x}(t) = f(x(t), u(t)), \quad y = h(x(t)), \tag{4.15}$$

where the input u takes values in a compact subset Ω of $\mathbb{R}^m$. Assume that the system is UBIBS and there exists some $\beta \in \mathcal{KL}$ such that (3.1) holds for all trajectories of (4.15). Let $\omega : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be defined by

$$\omega(\xi) := \sup \left\{ |y(t, \xi, u)| : \ t \geq 0, u \in \mathcal{M}_\Omega \right\}. \tag{4.16}$$

It then holds that

$$|h(\xi)| \leq \omega(\xi) \leq \beta_0(|\xi|) \qquad \forall \, \xi \in \mathbb{R}^n, \tag{4.17}$$

where $\beta_0(s) = \beta(s, 0)$. Moreover, if there exists some $\sigma \in \mathcal{K}$ such that (3.2) holds for all trajectories, then the above can be strengthened to

$$|h(\xi)| \leq \omega(\xi) \leq \sigma(|h(\xi)|) \qquad \forall \, \xi \in \mathbb{R}^n. \tag{4.18}$$

Observe that, for any $\xi \in \mathbb{R}^n$, $u \in \mathcal{M}_\Omega$, and $t_1 \geq 0$,

$$\omega(x(t_1, \xi, u)) \leq \sup_{t \geq 0, v \in \mathcal{M}_\Omega} |y(t_1 + t, \xi, v)| \leq \beta(|\xi|, t_1). \tag{4.19}$$

Also ω decreases along trajectories, i.e.,

$$\omega(x(t, \xi, u)) \leq \omega(\xi) \qquad \forall t \geq 0, \ \xi \in \mathbb{R}^n, \ u \in \mathcal{M}_\Omega. \tag{4.20}$$

Define

$$\mathcal{D} := \{\xi : \ y(t, \xi, u) = 0 \quad \forall \, t \geq 0, \ \forall \, u \in \mathcal{M}_\Omega\}.$$

Then $\omega(\xi) = 0$ if and only if $\xi \in \mathcal{D}$. For $\xi \notin \mathcal{D}$, it holds that

$$\omega(\xi) = \sup_{0 \leq t \leq t_\xi, \ u \in \mathcal{M}_\Omega} |y(t, \xi, u)|, \tag{4.21}$$

where $t_\xi = T_{|\xi|}(\omega(\xi)/2)$, and $T_r(s)$ is defined as in Lemma A.1 associated with the function β.

LEMMA 4.2. *The function $\omega(\xi)$ is locally Lipschitz on $\mathbb{R}^n \setminus \mathcal{D}$ and continuous everywhere.*

Proof. First notice that

$$\varliminf_{\xi\to\xi_0} \omega(\xi) \geq \omega(\xi_0) \qquad \forall\, \xi_0 \in \mathbb{R}^n; \tag{4.22}$$

that is, $\omega(\xi)$ is lower semicontinuous on $\mathbb{R}^n$. Indeed, pick ξ_0 and let $c := \omega(\xi_0)$. Take any $\varepsilon > 0$. Then there are some u_0 and t_0 so that $|y(t_0, \xi_0, u_0)| \geq c - \varepsilon/2$. By continuity of $y(t_0, \cdot, u_0)$, there is some neighborhood $\widetilde{U}_0$ of ξ_0 so that $|y(t_0, \xi, u_0)| \geq c - \varepsilon$ for all $\xi \in \widetilde{U}_0$. Thus $\omega(\xi) \geq c - \varepsilon$ for all $\xi \in \widetilde{U}_0$, and this establishes (4.22).

Fix any $\xi_0 \in \mathbb{R}^n \setminus \mathcal{D}$, and let $c_0 = \omega(\xi_0)/2$. Then there exists a bounded neighborhood U_0 of ξ_0 such that

$$\omega(\xi) \geq c_0 \qquad \forall\, \xi \in U_0.$$

Let s_0 be such that $|\xi| \leq s_0$ for all $\xi \in U_0$. Then

$$\omega(\xi) = \sup\{|y(t, \xi, u)| : \ t \in [0, t_1], u \in \mathcal{M}_\Omega\} \qquad \forall\, \xi \in U_0,$$

where $t_1 = T_{s_0}(c_0/2)$. By [12, Proposition 5.5], one knows that $x(t, \xi, u)$ is locally Lipschitz in ξ uniformly on $u \in \mathcal{M}_\Omega$ and on $t \in [0, t_1]$, and therefore, so is $y(t, \xi, u)$. Let C be a constant such that

$$|y(t, \xi, u) - y(t, \eta, u)| \leq C\, |\xi - \eta| \qquad \forall\, \xi, \eta \in U_0,\ \forall\, 0 \leq t \leq t_1,\ \forall\, u \in \mathcal{M}_\Omega.$$

For any $\varepsilon > 0$ and any $\xi \in U_0$, there exist some $t_{\xi,\varepsilon} \in [0, t_1]$ and some $u_{\xi,\varepsilon}$ such that

$$\omega(\xi) \leq |y(t_{\xi,\varepsilon},\, \xi,\, u_{\xi,\varepsilon})| + \varepsilon.$$

It then follows that, for any $\xi, \eta \in U_0$, and for any $\varepsilon > 0$,

$$\omega(\xi) - \omega(\eta) \leq |y(t_{\xi,\varepsilon},\, \xi,\, u_{\xi,\varepsilon})| + \varepsilon - |y(t_{\xi,\,\varepsilon}, \eta,\, u_{\xi,\varepsilon})| \leq C\, |\xi - \eta| + \varepsilon.$$

Consequently,

$$\omega(\xi) - \omega(\eta) \leq C\, |\xi - \eta| \qquad \forall\, \xi, \eta \in U_0.$$

By symmetry,

$$\omega(\eta) - \omega(\xi) \leq C\, |\xi - \eta| \qquad \forall\, \xi, \eta \in U_0.$$

This proves that ω is locally Lipschitz on $\mathbb{R}^n \setminus \mathcal{D}$.

We now show that ω is continuous on $\mathcal{D}$. Fix $\xi_0 \in \mathcal{D}$. One would like to show that

$$\lim_{\xi\to\xi_0} \omega(\xi) = 0. \tag{4.23}$$

Assume that this does not hold. Then there exists some $\varepsilon_0 > 0$ and a sequence $\{\xi_k\}$ with $\xi_k \to \xi_0$ such that $\omega(\xi_k) > \varepsilon_0$ for all k. Without loss of generality, one may assume that

$$|\xi_k| \leq s_1 \qquad \forall\, k$$

for some $s_1 \geq 0$. It then follows that

$$\omega(\xi_k) = \sup\left\{|y(t, \xi_k, u)| : \; t \in [0, t_2], u \in \mathcal{M}_\Omega\right\},$$

where $t_2 = T_{s_1}(\varepsilon_0/2)$. Hence, for each k, there exists some $u_k \in \mathcal{M}_\Omega$ and some $\tau_k \in [0, t_2]$ such that

$$|y(\tau_k, \xi_k, u_k)| \geq \omega(\xi_k) - \varepsilon_0/2 \geq \varepsilon_0/2.$$

Again, by the locally Lipschitz continuity of the trajectories, one knows that there is some $C_1 > 0$ such that

$$|y(t, \xi_k, u) - y(t, \xi_0, u)| \leq C_1 \, |\xi_k - \xi_0| \qquad \forall\, k \geq 0, \; \forall\, 0 \leq t \leq t_2, \; \forall\, u \in \mathcal{M}_\Omega.$$

Hence,

$$|y(\tau_k, \xi_0, u_k)| \geq \varepsilon_0/4$$

for k large enough, contradicting the fact that $y(t, \xi_0, u) \equiv 0$ for all $u \in \mathcal{M}_\Omega$. This shows that (4.23) holds if $\xi_0 \in \mathcal{D}$. □

Next, we pick any smooth and bounded function $k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ whose derivative is everywhere positive, and define $W : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ by

$$W(\xi) := \sup\left\{\omega(x(t, \xi, u))k(t) : \; t \geq 0, \; u \in \mathcal{M}_\Omega\right\}. \tag{4.24}$$

Corresponding to k there are two positive real numbers $c_1 < c_2$ such that $k(t) \in [c_1, \, c_2]$ for all $t \geq 0$, and so

$$c_1\omega(\xi) \leq W(\xi) \leq c_2\omega(\xi) \qquad \forall\, \xi \in \mathbb{R}^n,$$

which implies that

$$c_1 \, |h(\xi)| \leq W(\xi) \leq c_2\beta_0(|\xi|) \qquad \forall\, \xi \in \mathbb{R}^n. \tag{4.25}$$

Note, for future reference, that it is always possible to find a bounded, positive, and decreasing continuous function $\tau(\cdot)$ with $\tau(t) \to 0$ as $t \to \infty$ such that

$$k'(t) \geq \tau(t) \quad \forall\, t \geq 0. \tag{4.26}$$

By (4.19), one knows that $\omega(x(t, \xi, u)) \to 0$ as $t \to \infty$. It follows that there is some $\tau_\xi \geq 0$ such that

$$W(\xi) = \sup\left\{\omega(x(t, \xi, u))k(t) : \; u \in \mathcal{M}_\Omega, \; 0 \leq t \leq \tau_\xi\right\}. \tag{4.27}$$

Furthermore, one can get the following estimate, where $\{T_r\}$ is a family of functions associated to β as in Lemma A.4.

Lemma 4.3. *For any $\xi \notin \mathcal{D}$ with $|\xi| \leq r$,*

$$W(\xi) = \sup\left\{\omega(x(t, \xi, u))k(t) : \; u \in \mathcal{M}_\Omega, \; 0 \leq t \leq \tau_\xi\right\},$$

where $\tau_\xi = T_r(\frac{c_1}{2c_2}\omega(\xi))$.

Proof. If the statement is not true, then for any $\varepsilon > 0$, there exists some $t_\varepsilon > T_r(\frac{c_1}{2c_2}\omega(\xi))$ and some $u_\varepsilon \in \mathcal{M}_\Omega$ such that

$$W(\xi) \leq \omega(x(t_\varepsilon, \xi, u_\varepsilon))k(t_\varepsilon) + \varepsilon.$$

This implies the following:

$$\begin{aligned}\omega(\xi) &\leq \frac{1}{c_1}W(\xi) \leq \frac{1}{c_1}\omega(x(t_\varepsilon, \xi, u_\varepsilon))k(t_\varepsilon) + \frac{\varepsilon}{c_1}\\ &\leq \frac{c_2}{c_1}\omega(x(t_\varepsilon, \xi, u_\varepsilon)) + \frac{\varepsilon}{c_1} \leq \frac{c_2}{c_1}\cdot\frac{c_1}{2c_2}\omega(\xi) + \frac{\varepsilon}{c_1}\\ &= \frac{\omega(\xi)}{2} + \frac{\varepsilon}{c_1}.\end{aligned}$$

Taking the limit as $\varepsilon \to 0$ results in a contradiction.

LEMMA 4.4. *The function $W(\cdot)$ is locally Lipschitz on $\mathbb{R}^n \setminus \mathcal{D}$ and continuous everywhere.*

Proof. Fix $\xi_0 \notin \mathcal{D}$. Let K_0 be a compact neighborhood of ξ_0 such that $K_0 \cap \mathcal{D} = \emptyset$. Since ω is continuous, it follows that there is some $r_0 > 0$ such that $\omega(\xi) > r_0$ for all $\xi \in K_0$, and hence, $W(\xi) > r_1 := c_1 r_0$ for all $\xi \in K_0$. Let

$$T_0 = T_{s_0}\left(\frac{r_1}{8c_2}\right),$$

where $s_0 > 0$ is such that $|\xi| \leq s_0$ for all $\xi \in K_0$. Let $C > 0$ be such that

$$|y(t, \xi, u) - y(t, \eta, u)| \leq C\,|\xi - \eta| \quad \forall\, t \in [0, T_0],\ \forall\, \xi, \eta \in K_0,\ \forall\, u \in \mathcal{M}_\Omega.$$

Let

$$K_1 = K_0 \cap \left\{\xi :\ |\xi - \xi_0| \leq \frac{r_1}{16Cc_2}\right\}.$$

Fix any $\varepsilon \in (0, r_1/4)$. Then, for any $\xi \in K_1$, there exist $t_{\xi,\varepsilon} \in [0, T_0]$ and $u_{\xi,\varepsilon} \in \mathcal{M}_\Omega$ such that

$$W(\xi) \leq \omega(x(t_{\xi,\varepsilon},\, \xi,\, u_{\xi,\varepsilon}))k(t_{\xi,\varepsilon}) + \varepsilon.$$

Claim. For any $\xi, \eta \in K_1$, $\omega(x(t_{\xi,\varepsilon}, \eta, u_{\xi,\varepsilon})) \geq \frac{r_1}{8c_2}$.

Proof. First we note that for any $\xi \in K_1 \subset K_0$,

$$\omega(x(t_{\xi,\varepsilon}, \xi, u_{\xi,\varepsilon})) \geq \frac{W(\xi) - \varepsilon}{c_2} \geq \frac{W(\xi)}{2c_2} \geq r_2,$$

where $r_2 := \frac{r_1}{2c_2}$. Thus, for each $\xi \in K_1$, there exists some $v_\xi \in \mathcal{M}_\Omega$ and some $\tau_\xi > 0$ such that

$$|y(\tau_\xi, x(t_{\xi,\varepsilon}, \xi, u_{\xi,\varepsilon}), v_\xi)| \geq \omega(x(t_{\xi,\varepsilon}, \xi, u_{\xi,\varepsilon})) - r_2/2 \geq r_2/2.$$

Observe that

$$y(\tau_\xi, x(t_{\xi,\varepsilon}, \xi, u_{\xi,\varepsilon}), v_\xi) = y(\tau_\xi + t_{\xi,\varepsilon}, \xi, \overline{v}_{\xi,\varepsilon}),$$

where $\overline{v}_{\xi,\varepsilon}$ is the concatenation of $u_{\xi,\varepsilon}$ and v_ξ, i.e.,

$$\overline{v}_{\xi,\varepsilon}(t) = \begin{cases} u_{\xi,\varepsilon}(t), & \text{if } 0 \leq t < t_{\xi,\varepsilon},\\ v_\xi(t - t_{\xi,\varepsilon}), & \text{if } t \geq t_{\xi,\varepsilon}.\end{cases}$$

Noticing that $|y(t,\xi,u)| \le r_2/2$ for all $t \ge T_{s_0}(r_2/4)$, one concludes that $\tau_\xi + t_{\xi,\varepsilon} < T_{s_0}(r_2/4) = T_0$. Note also that for any $\eta \in K_1$,

$$\begin{aligned}
&|y(\tau_\xi, x(t_{\xi,\varepsilon},\eta,u_{\xi,\varepsilon}), v_\xi)| = |y(\tau_\xi + t_{\xi,\varepsilon},\eta,\overline{v}_{\xi,\varepsilon})| \\
&\ge |y(\tau_\xi + t_{\xi,\varepsilon},\xi,\overline{v}_{\xi,\varepsilon})| - |y(\tau_\xi + t_{\xi,\varepsilon},\eta,\overline{v}_{\xi,\varepsilon}) - y(\tau_\xi + t_{\xi,\varepsilon},\xi,\overline{v}_{\xi,\varepsilon})| \\
&\ge \frac{r_2}{2} - C\,|\xi-\eta| \\
&\ge \frac{r_2}{2} - 2C\,\frac{r_1}{16Cc_2} = \frac{r_1}{4c_2} - \frac{r_1}{8c_2} = \frac{r_1}{8c_2}.
\end{aligned}$$

This implies that $\omega(x(t_{\xi,\varepsilon},\eta,u_{\xi,\varepsilon})) \ge \frac{r_1}{8c_2}$ for all $\xi,\eta \in K_1$, as claimed.

According to [12, Proposition 5.1], there is some compact set K_2 such that $x(t,\xi,u) \in K_2$ for all $0 \le t \le T_0$, all $\xi \in K_1$, and all $u \in \mathcal{M}_\Omega$. Let

$$K_3 = K_2 \cap \{\xi :\ \omega(\xi) \ge r_1/8c_2\}.$$

Applying Lemma 4.2, one knows that there is some $C_1 > 0$ such that

$$|\omega(\zeta_1) - \omega(\zeta_2)| \le C_1\,|\zeta_1 - \zeta_2| \qquad \forall\,\zeta_1,\zeta_2 \in K_3.$$

Since for all $\xi,\eta \in K_1$, and all $0 < \varepsilon < r_1/4$, $x(t_{\xi,\varepsilon},\eta,u_{\xi,\varepsilon}) \in K_3$, we have

$$|\omega(x(t_{\xi,\varepsilon},\xi,u_{\xi,\varepsilon})) - \omega(x(t_{\xi,\varepsilon},\eta,u_{\xi,\varepsilon}))| \le C_1\,|x(t_{\xi,\varepsilon},\xi,u_{\xi,\varepsilon}) - x(t_{\xi,\varepsilon},\eta,u_{\xi,\varepsilon})|$$

for all $\xi,\eta \in K_1$, and all $\varepsilon \in (0, r_1/4)$. Hence,

$$\begin{aligned}
W(\xi) - W(\eta) &\le \omega(x(t_{\xi,\varepsilon},\xi,u_{\xi,\varepsilon}))k(t_{\xi,\varepsilon}) - \omega(x(t_{\xi,\varepsilon},\eta,u_{\xi,\varepsilon}))k(t_{\xi,\varepsilon}) + \varepsilon \\
&\le c_2\,|\omega(x(t_{\xi,\varepsilon},\xi,u_{\xi,\varepsilon})) - \omega(x(t_{\xi,\varepsilon},\eta,u_{\xi,\varepsilon}))| + \varepsilon \\
&\le c_2C_1\,|x(t_{\xi,\varepsilon},\xi,u_{\xi,\varepsilon}) - x(t_{\xi,\varepsilon},\eta,u_{\xi,\varepsilon})| + \varepsilon \\
&\le c_2C_1C_2\,|\xi-\eta| + \varepsilon,
\end{aligned}$$

where $C_2 > 0$ is such a constant that $|x(t,\xi,u) - x(t,\eta,u)| \le C_2\,|\xi-\eta|$ for all $\xi,\eta \in K_3$, all $t \in [0,T_0]$, and all $u \in \mathcal{M}_\Omega$. Note that the above holds for any $\varepsilon \in (0, r_1/4)$, and thus,

$$W(\xi) - W(\eta) \le C_3\,|\xi-\eta|$$

for all $\xi,\eta \in K_1$, where $C_3 = c_2C_1C_2$. By symmetry, one proves that

$$W(\eta) - W(\xi) \le C_3\,|\xi-\eta|$$

for all $\xi,\eta \in K_1$.

To prove the continuity of W on $\mathcal{D}$, it is enough to notice that for any $\xi \in \mathcal{D}$, $W(\xi) = 0$ and

$$|W(\xi) - W(\eta)| \le c_2\omega(\eta) \to 0, \qquad \text{as} \quad \eta \to \xi.$$

The proof of Lemma 4.4 is thus concluded. □

Below we show that W is decreasing along trajectories. Pick any $\xi \notin \mathcal{D}$. Let $\theta_0 > 0$ be such that

$$\omega(x(t,\xi,\mathbf{v})) \ge \omega(\xi)/2 \quad \forall\,t \in [0,\theta_0],\ \forall\,v \in \Omega,$$

where $\mathbf{v}$ denotes the constant function $\mathbf{v}(t) \equiv v$. (Observe that such a θ_0 exists because ω is continuous.) Pick any $\theta \in [0, \theta_0]$, and let $\eta_{\mathbf{v}} = x(\theta, \xi, \mathbf{v})$. For any $\varepsilon > 0$, there exists some $t_{\mathbf{v},\varepsilon}$ and $u_{\mathbf{v},\varepsilon} \in \mathcal{M}_\Omega$ such that

$$\begin{aligned} W(\eta_{\mathbf{v}}) &\le \omega(x(t_{\mathbf{v},\varepsilon}, \eta_{\mathbf{v}}, u_{\mathbf{v},\varepsilon}))k(t_{\mathbf{v},\varepsilon}) + \varepsilon \\ &= \omega(x(t_{\mathbf{v},\varepsilon} + \theta, \xi, \overline{u}_{\mathbf{v},\varepsilon}))k(t_{\mathbf{v},\varepsilon} + \theta)\left(1 - \frac{k(t_{\mathbf{v},\varepsilon} + \theta) - k(t_{\mathbf{v},\varepsilon})}{k(t_{\mathbf{v},\varepsilon} + \theta)}\right) + \varepsilon \\ &\le W(\xi)\left(1 - \frac{k(t_{\mathbf{v},\varepsilon} + \theta) - k(t_{\mathbf{v},\varepsilon})}{c_2}\right) + \varepsilon, \end{aligned} \tag{4.28}$$

where $\overline{u}_{\mathbf{v},\varepsilon}$ denotes the concatenation of $\mathbf{v}$ and $u_{\mathbf{v},\varepsilon}$. Still for the fixed ξ and θ, and for any $r > |\xi|$, define

$$T^r_{\xi,\theta} := \max_{\widetilde{v} \in \Omega} T_r\left(\frac{c_1}{2c_2}\omega(x(\theta, \xi, \widetilde{\mathbf{v}}))\right). \tag{4.29}$$

Notice that $x(\theta, \xi, \widetilde{\mathbf{v}})$ is jointly continuous as a function of $(\theta, \xi, \widetilde{v})$. Since ω and T_r are both continuous, this maximum is well defined and, moreover, $T^r_{\xi,\theta}$ is continuous as a function of θ, so, in particular,

$$\lim_{\theta \to 0^+} T^r_{\xi,\theta} \;=\; T_r\left(\frac{c_1}{c_2}\omega(\xi)\right). \tag{4.30}$$

Claim. $t_{\mathbf{v},\varepsilon} + \theta \le T^r_{\xi,\theta}$ for all $v \in \Omega$ and for all $\varepsilon \in (0, \frac{c_1}{4}\omega(\xi))$.

Proof. Assume that this is not true. Then there is some $v \in \Omega$ and some $\varepsilon \in \big(0, \frac{c_1}{4}\omega(\xi)\big)$ such that $t_{\mathbf{v},\varepsilon} + \theta > T^r_{\xi,\theta}$, and, in particular,

$$t_{\mathbf{v},\varepsilon} + \theta \ge T_r\left(\frac{c_1}{2c_2}\omega(x(\theta, \xi, \mathbf{v}))\right),$$

from which it follows that

$$\omega(x(t_{\mathbf{v},\varepsilon}, \eta_{\mathbf{v}}, u_{\mathbf{v},\varepsilon})) = \omega(x(t_{\mathbf{v},\varepsilon} + \theta, \xi, \overline{u}_{\mathbf{v},\varepsilon})) \le \frac{c_1}{2c_2}\omega(x(\theta, \xi, \mathbf{v})) = \frac{c_1}{2c_2}\omega(\eta_{\mathbf{v}})$$

for some input function $\overline{u}_{\mathbf{v},\varepsilon}$ (which we can take to be the concatenation of $\mathbf{v}$ and $u_{\mathbf{v},\varepsilon}$; note that the inequality follows from (4.19) and the definition of the functions T_r).

By the definition of W, one has

$$\begin{aligned} \omega(\eta_{\mathbf{v}}) &\le \frac{1}{c_1}W(\eta_{\mathbf{v}}) \le \frac{1}{c_1}\omega(t_{\mathbf{v},\varepsilon}, \eta_{\mathbf{v}}, u_{\mathbf{v},\varepsilon})k(t_{\mathbf{v},\varepsilon}) + \frac{\varepsilon}{c_1} \\ &\le \frac{c_2}{c_1}\omega(t_{\mathbf{v},\varepsilon} + \theta, \xi, \overline{u}_{\mathbf{v},\varepsilon}) + \frac{\varepsilon}{c_1} \\ &\le \frac{1}{2}\omega(\eta_{\mathbf{v}}) + \frac{\varepsilon}{c_1}, \end{aligned}$$

which is impossible, since $\varepsilon < \frac{c_1}{4}\omega(\xi) \le \frac{c_1}{2}\omega(\eta_{\mathbf{v}})$. This proves the claim.

From (4.28), we have, for any $v \in \mathcal{D}$ and for any ε small enough,

$$W(x(\theta, \xi, \mathbf{v})) - W(\xi) \;\le\; -\frac{W(\xi)}{c_2}\,\tau(t_{\mathbf{v},\varepsilon} + c\theta)\theta + \varepsilon$$

for some $c \in (0,1)$, where we used the mean value theorem in order to estimate the change in k, and where τ is a function as in (4.26). Using the monotonicity of $\tau(\cdot)$ and the above claim, one concludes

$$W(x(\theta,\xi,\mathbf{v})) - W(\xi) \;\leq\; -\frac{W(\xi)}{c_2}\,\tau\left(T^r_{\xi,\theta}\right)\theta + \varepsilon$$

for all ε small enough. Letting $\varepsilon \to 0$, one obtains

$$W(x(\theta,\xi,\mathbf{v})) - W(\xi) \;\leq\; -\frac{W(\xi)}{c_2}\,\tau\left(T^r_{\xi,\theta}\right)\theta \quad \forall\, v \in \Omega.$$

Thus one concludes that for any $v \in \Omega$ and any $\theta > 0$,

$$\frac{W(x(\theta,\xi,\mathbf{v})) - W(\xi)}{\theta} \leq -\frac{W(\xi)}{c_2}\,\tau(T^r_{\xi,\theta}).$$

Since W is locally Lipschitz on $\mathbb{R}^n \backslash \mathcal{D}$, it is differentiable almost everywhere on $\mathbb{R}^n \backslash \mathcal{D}$, and hence, for any $v \in \Omega$, any $r > |\xi|$, and any ξ at which W is differentiable,

$$\begin{aligned} DW(\xi)f(\xi,v) &= \lim_{\theta\to 0^+} \frac{W(x(\theta,\xi,\mathbf{v})) - W(\xi)}{\theta} \leq -\lim_{\theta\to 0^+} \frac{W(\xi)}{c_2}\,\tau(T^r_{\xi,\theta}) \\ &= -\frac{W(\xi)}{c_2}\,\tau\Big(\lim_{\theta\to 0^+} T^r_{\xi,\theta}\Big) = -\frac{W(\xi)}{c_2}\,\tau\left(T_r\left(\frac{c_1}{c_2}\omega(\xi)\right)\right) \\ &\leq -\frac{W(\xi)}{c_2}\,\tau\left(T_r\left(\frac{c_1}{c_2^2}W(\xi)\right)\right) = -\widetilde{\alpha}_3(W(\xi),r), \end{aligned} \tag{4.31}$$

where $\widetilde{\alpha}_3(s,r) = \frac{s}{c_2}\,\tau(T_r(c_3 s))$ with $c_3 = c_1/c_2^2$. Since (4.31) holds for all $r > |\xi|$, it follows that

$$DW(\xi)f(\xi,v) \leq -\widetilde{\alpha}_3(W(\xi), 2\,|\xi|) \tag{4.32}$$

for all $v \in \Omega$ and for almost all $\xi \in \mathbb{R}^n \setminus \mathcal{D}$.

Since $T_r(s)$ is defined for all $r \geq 0$ and $s > 0$, one sees that $\widetilde{\alpha}_3$ is defined on $\mathbb{R}_{>0} \times \mathbb{R}_{\geq 0}$. Extend $\widetilde{\alpha}_3$ to $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ by letting $\widetilde{\alpha}_3(0,r) := 0$ for all $r \geq 0$. By the continuity property of τ and $T_r(\cdot)$, one sees that $\widetilde{\alpha}_3(\cdot,r)$ is continuous for each r. (The continuity at $s = 0$ follows from $\widetilde{\alpha}_3(s,r) = s\tau(T_r(c_3 s))/c_2 \leq s\tau(0)/c_2$ for all $s > 0$.) Furthermore, since $\tau(T_r(c_3 s))$ is nondecreasing in s, it follows that $\widetilde{\alpha}_3(s,r)$ is of class $\mathcal{K}$ in s. Let $\breve{\alpha}_3(s,r) = \widetilde{\alpha}_3(s,2r)/(1+r)$. This function tends to zero as $r \to \infty$, because $\widetilde{\alpha}_3(s,r)$ is nonincreasing in r; thus $\breve{\alpha}_3(s,r)$ is of class $\mathcal{KL}$. Moreover,

$$DW(\xi)f(\xi,v) \leq -\breve{\alpha}_3(W(\xi),|\xi|) \qquad \forall\, \xi \in \mathbb{R}^n \setminus \mathcal{D},\ \forall\, v \in \Omega.$$

By Corollary A.3, there exists a continuous $\mathcal{KL}$-function $\widehat{\alpha}_3$ such that

$$DW(\xi)f(\xi,v) \leq -\widehat{\alpha}_3(W(\xi),|\xi|) \qquad \forall\, \xi \in \mathbb{R}^n \setminus \mathcal{D},\ \forall\, v \in \Omega. \tag{4.33}$$

To complete the proof, we follow the strategy used in [12] to find a smooth approximation of W. First of all, by Theorem B.1 in [12], applied on $\mathbb{R}^n \setminus \mathcal{D}$, there is a continuous function W_1 that is smooth on $\mathbb{R}^n \setminus \mathcal{D}$ such that

$$|W_1(\xi) - W(\xi)| \leq \frac{W(\xi)}{2} \qquad \forall\, \xi \in \mathbb{R}^n \setminus \mathcal{D}, \tag{4.34}$$

and

$$DW_1(\xi)f(\xi,v) \le -\widehat{\alpha}_3(W(\xi),|\xi|)/2 \qquad \forall\, \xi \in \mathbb{R}^n \setminus \mathcal{D},\ \forall\, v \in \Omega. \tag{4.35}$$

We extend W_1 to all of $\mathbb{R}^n$ by letting $W_1 \equiv 0$ on $\mathcal{D}$; thus, the approximation (4.34) holds on all of $\mathbb{R}^n$. (Note that W and $\widehat{\alpha}_3(V(\xi),|\xi|)$ are both continuous, so the result in [12] can indeed be applied.)

Next, we appeal to Lemma 4.3 in [12]. This shows that there exists some $\rho \in \mathcal{K}_\infty$ with $\rho'(s) > 0$ for all $s > 0$ such that $\rho \circ W_1$ is smooth everywhere. Let $V = \rho \circ W_1$. It follows from (4.25) and (4.34) that

$$\alpha_1(|h(\xi)|) \le V(\xi) \le \alpha_2(|\xi|) \qquad \forall\, \xi \in \mathbb{R}^n,$$

where $\alpha_1(s) = \rho(c_1 s/2)$, $\alpha_2(s) = \rho(2c_2\beta_0(s))$, and it follows from (4.34) and (4.35) that

$$DV(\xi)f(\xi,\mu) \le -\rho'(W_1(\xi))\widehat{\alpha}_3(W(\xi),|\xi|)/2 \le -\alpha_3(V(\xi),|\xi|) \tag{4.36}$$

for all $\xi \in \mathbb{R}^n \setminus \mathcal{D}$ and all $\mu \in \Omega$, where

$$\alpha_3(s,r) = \frac{\rho'(\rho^{-1}(s))\widehat{\alpha}_3(\rho^{-1}(V(\xi))/2,r)}{2}.$$

Since V has local (actually, global) minima at all points in $\mathcal{D}$, it follows that $DV(\xi) \equiv 0$ on $\mathcal{D}$, so we know that the estimate (4.36) also holds on all of $\mathbb{R}^n$.

Finally, observe that if there exists $\sigma \in \mathcal{K}$ such that (3.2) holds for all trajectories of the system, then (4.18) holds for all ξ, which, in turn, implies that

$$c_1\,|h(\xi)| \le W(\xi) \le c_2\sigma(|h(\xi)|) \qquad \forall\, \xi \in \mathbb{R}^n. \tag{4.37}$$

This results in the desired inequality

$$\alpha_1(|h(\xi)|) \le V(\xi) \le \sigma_1(|h(\xi)|) \qquad \forall\, \xi \in \mathbb{R}^n, \tag{4.38}$$

where $\sigma_1(s) = \rho(2c_2\sigma(s))$. This shows that if (3.2) holds for some $\sigma \in \mathcal{K}$, then property (3.4) can be strengthened to property (3.6).

Finally, suppose that, in the above proof, one strengthens (3.1) to (3.3). Associated to the function β there are, as before, functions $\{T_r\}$. Since we also have an estimate as in (3.1), there are functions $\{T_r\}$ associated to a β as in (3.1); without loss of generality, we will assume that the same T_r's work for both. Thus, we know that, provided $t \ge T_r(s)$, $|y(t,\xi,u)| \le s$ whenever $|h(\xi)| \le r$ or $|\xi| \le r$. The claim stated after (4.30) holds now for all $r > |h(\xi)|$ (instead of merely if $r > |\xi|$), because (4.19) can be strengthened to

$$\omega(x(t_1,\xi,u)) \;\le\; \beta(|h(\xi)|\,,t_1).$$

We now repeat the above proof to get a function $W(\xi)$ satisfying (4.37), and corresponding to (4.33), one has now also

$$DW(\xi)f(\xi,v) \le -\widehat{\alpha}_3(W(\xi),|h(\xi)|) \le -\widehat{\alpha}_3\left(W(\xi),\frac{W(\xi)}{c_1}\right)$$

for all $\xi \in \mathbb{R}^n \setminus \mathcal{D}$ and all $v \in \Omega$. Therefore, on $\mathbb{R}^n \setminus \mathcal{D}$,

$$DW(\xi)f(\xi,v) \le -\alpha_4(W(\xi)),$$

where $\widetilde{\alpha}_4(s) = \widetilde{\alpha}_3(s, s/c_1)$ is a continuous positive definite function. Using the same smoothing argument as earlier, we can show that there is a smooth function V such that (4.38) holds for some $\sigma_1, \sigma_2 \in \mathcal{K}_\infty$, and (4.36) can be strengthened to

$$DV(\xi)f(\xi, v) \leq -\widehat{\alpha}_4(V(\xi)) \tag{4.39}$$

for all $\xi \in \mathbb{R}^n$ and all $v \in \Omega$, where $\widehat{\alpha}_4(\cdot)$ is some continuous positive definite function.

Now we modify the function V to get V_1 so that V_1 satisfies inequalities of type (4.37) and (4.39) with $\widehat{\alpha}_4$ replaced by a $\mathcal{K}_\infty$ function α_5. For this purpose, let $\rho_0(\cdot)$ be a smooth $\mathcal{K}_\infty$-function such that $\rho_0(s)\widehat{\alpha}_4(s) \geq 1$ for $s \geq 1$, and let

$$\rho_1(s) = e^{\int_0^s \rho_0(s_1)\, ds_1} - 1.$$

Define $V_1(\xi) = \rho_1(V(\xi))$. It holds that

$$\widehat{\alpha}_1(|h(\xi)|) \leq V_1(\xi) \leq \widehat{\alpha}_2(|h(\xi)|) \qquad \forall\, \xi \in \mathbb{R}^n,$$

where $\widehat{\alpha}_1(s) = \rho_1(\alpha_1(s))$, $\widehat{\alpha}_2(s) = \rho_1(\alpha_2(s))$, and

$$DV_1(\xi)f(\xi, v) = -(V_1(\xi) + 1)\rho_0(V(\xi))\widehat{\alpha}_4(V(\xi)) \leq -\alpha_5(V_1(\xi))$$

for all $\xi \in \mathbb{R}^n$ and all $v \in \Omega$, where α_5 is any $\mathcal{K}_\infty$ function with the property that

$$\alpha_5(\rho_1(s)) \ \leq \ (\rho_1(s) + 1)\rho_0(s)\widehat{\alpha}_4(s)$$

for all $s \geq 0$ (such a $\mathcal{K}_\infty$-function exists because $(s + 1)\rho_0(s)\widehat{\alpha}_4(s) \geq s$ for all $s \geq 1$). Using V_1 as a Lyapunov function, this completes the proof. □

5. Remarks. The concept of IOS does not distinguish between "measured outputs," which may be used to provide information about the state of a system, and "target outputs," which are often the object of control, nor does it allow for the consideration of "robustness" to disturbances. A more general concept can be studied as well, as follows. Suppose that, instead of systems as in (1.1), we study more general systems of the following form:

$$\dot{x}(t) = f(x(t), u(t), d(t)), \quad y(t) = h(x(t)), \quad w(t) = k(x(t)), \tag{5.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}^n$, $h : \mathbb{R}^n \to \mathbb{R}^p$, and $k : \mathbb{R}^n \to \mathbb{R}^q$ are all locally Lipschitz continuous (for some nonnegative integers n, m, r, p, q). We think of the functions $d(\cdot)$ and $w(\cdot)$ as disturbances and measured outputs, respectively. Even more generality is gained if one considers, as mentioned in [17], a "measure" for states (in the sense of [11]), which we denote by $|x|_{\mathcal{A}}$ in analogy to the distance to a set $\mathcal{A}$ as in previous extensions of the ISS notion. Then, a natural definition of relative stability is given by the requirement that there should exist a $\mathcal{KL}$-function β and $\mathcal{K}$-functions γ_1 and γ_2 such that, for each initial state ξ and inputs (u, d), and for all t in the domain of definition of the corresponding maximal solution $x(\cdot)$ of (5.1),

$$|y(t)| \ \leq \ \beta(|\xi|_{\mathcal{A}}, t) + \gamma_1(\|u\|) + \gamma_2(\|w\|), \tag{5.2}$$

where y and w are the functions $h(x(\cdot))$ and $k(x(\cdot))$, respectively. Observe that, when d does not appear in the equations and when $k \equiv 0$, we recover (if $|\cdot|_{\mathcal{A}} = |\cdot|$) the IOS definition. When, again, d does not appear in the equations, but now $h(x) = x$, we recover (if $|\cdot|_{\mathcal{A}} = |\cdot|$) the input/output to state stability (IOSS) notion of zero-detectability discussed in [18] and recently completely characterized in [8]. (These notions are related by the fact that a system is ISS if and only if it is both IOSS and IOS, which generalizes the linear systems theory fact that internal stability is equivalent to detectability plus external stability.) A sufficient Lyapunov-theoretic condition for our general notion (which could be called "input/measurement to output stability") is the existence of a smooth $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ such that, for some $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$,

$$\alpha_1(|h(\xi)|) \leq V(\xi) \leq \alpha_2(|\xi|_{\mathcal{A}}) \qquad \forall\, \xi \in \mathbb{R}^n \tag{5.3}$$

and there exist $\chi_1, \chi_2 \in \mathcal{K}$, and $\alpha_3 \in \mathcal{KL}$ such that

$$DV(\xi)f(\xi,\mu,\delta) \leq -\alpha_3(V(\xi),|\xi|) + \chi_1(|\mu|) + \chi_2(|h(\xi)|) \quad \forall\, \xi,\ \forall \mu,\ \forall\, \delta, \tag{5.4}$$

or obvious variations of this inequality. We leave the formulation of converse theorems for future work.

Appendix A. Some facts regarding $\mathcal{KL}$ functions.

The following simple observation is proved in [19] and will be needed here too.

Lemma A.1. *For any $\mathcal{KL}$-function β, there exists a family of mappings $\{T_r\}_{r\geq 0}$ such that*

- *for each fixed $r > 0$, $T_r : \mathbb{R}_{>0} \xrightarrow{\text{onto}} \mathbb{R}_{>0}$ is continuous and strictly decreasing, and $T_0(s) \equiv 0$;*
- *for each fixed $s > 0$, $T_r(s)$ is strictly increasing as r increases and is such that $\beta(r, T_r(s)) < s$, and consequently, $\beta(r,t) < s$ for all $t \geq T_r(s)$.*

Lemma A.2. *For any $\mathcal{KL}$ function β, there exist two $\mathcal{K}$ functions κ_1 and κ_2 so that*

$$\beta(s,t) \;\geq\; \frac{\kappa_1(s)}{1+\kappa_2(t)} \tag{A.1}$$

for all $s \geq 0$ and all $t \geq 0$.

Proof. We assume that $b := \sup_s \beta(s,0) < \infty$ (otherwise, we first find a $\beta_0 \leq \beta$ with that property and prove the result for β_0). We define, for all $s \geq 0$ and $t \geq 0$,

$$\widetilde{\beta}(s,t) := \int_t^{t+1} \beta(s,\tau)\, d\tau.$$

Note that $\widetilde{\beta}$ is again of class $\mathcal{KL}$, and $\widetilde{\beta}(s,t) \leq \beta(s,t)$ for all s,t. Let

$$\widetilde{\alpha}(t) := \sup_{s\geq 0} \widetilde{\beta}(s,t).$$

This is finite everywhere, since it is bounded by b. Moreover, it is a continuous function, because

$$\widetilde{\alpha}(t) := \int_t^{t+1} \alpha(\tau)\, d\tau,$$

where α is the decreasing function (not necessarily strictly) defined by $\alpha(t) := \sup_{s\geq 0} \beta(s,t)$. We will write from now on $\widetilde{\beta}(\infty,t)$ instead of $\widetilde{\alpha}(t)$. Finally, we let

$$\rho(x) := \max\{x, 0\}$$

for all $x \in \mathbb{R}$ and introduce the following function:

$$c :\ \mathbb{R}^2 \to \mathbb{R} :\ (x,y) \mapsto -\ln \widetilde{\beta}\left(\frac{1}{\rho(x)}, \rho(y)\right) - \rho(-x) - \rho(-y),$$

where we understand $\widetilde{\beta}(\frac{1}{0}, t)$ as $\widetilde{\alpha}(t)$. As in [1], we let $\mathcal{N}$ denote the class of all functions $k : \mathbb{R} \to \mathbb{R}$ that are nondecreasing, continuous, and unbounded below. Note that c is of class $\mathcal{N}$ on each variable separately. (Continuity follows from the continuity of each of $\widetilde{\beta}(\infty, \cdot)$, $\widetilde{\beta}(s, \cdot)$ for each $s \geq 0$, and $\widetilde{\beta}(\cdot, t)$ for each $t \geq 0$ as well as continuity of ρ. The nondecreasing property is clear, using that $\widetilde{\beta}(\cdot, t)$ for each $t \geq 0$ and ρ are nondecreasing, and that $\widetilde{\beta}(\infty, \cdot)$ and $\widetilde{\beta}(s, \cdot)$ for each $s \geq 0$ are nonincreasing. Unbounded below follows from the fact that for $x \to -\infty$ we have $c(x, y_0) = a + x$, where $a = \widetilde{\beta}(\infty, \rho(y_0)) - \rho(-y_0)$ and for $y \to -\infty$ we have $c(x_0, y) = a + y$, where $a = -\ln \widetilde{\beta}\left(\frac{1}{\rho(x_0)}, 0\right) - \rho(-x_0)$.

By Proposition 3.4 in [1], there is some $k \in \mathcal{N}$ such that $c(x, y) \leq k(x) + k(y)$ for all x, y. So, we can write, after using that $\beta \geq \widetilde{\beta}$: $\beta(1/x, y) \geq e^{-k(x)} e^{-k(y)}$ for all $x, y > 0$. Equivalently,

$$\beta(s, t) \;\geq\; \frac{\kappa_1(s)}{1 + \kappa_2(t)}$$

for all $s, t > 0$, when we define

$$\kappa_1(s) := \; e^{-k(1/s) - k(0)}$$

for all $s > 0$ and

$$\kappa_2(t) := \; e^{k(t) - k(0)} - 1$$

for all $t \geq 0$. Observe that both of these functions are continuous, nondecreasing, and nonnegative. Moreover, $\kappa_2(0) = 0$, so κ_2 is in $\mathcal{K}$. From the inequality

$$[1 + \kappa_2(0)]\, \beta(s, 0) \;\geq\; \kappa_1(s)$$

for all $s > 0$, and the fact that $\beta(0, 0) = 0$, we conclude that $\lim_{s \to 0^+} \kappa_1(s) = 0$, so we may extend κ_1 by defining $\kappa_1(0) = 0$, and thus κ_1 is in $\mathcal{K}$ as well. □

As κ_1 and κ_2 in Lemma A.2 are continuous, we have, in particular, the following corollary.

COROLLARY A.3. *For any $\mathcal{KL}$-function β, there is a (jointly) continuous $\mathcal{KL}$-function β_1 such that $\beta(s, r) \geq \beta_1(s, r)$ for all $(s, r) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$.*

The following is a generalization of the comparison lemma given in [12]. It plays a role in the proofs of sufficiency, which are the easier parts of the theorems.

LEMMA A.4. *For any $\mathcal{K}$-function κ, there exists a $\mathcal{KL}$ function β such that if $y(\cdot)$ is any locally absolutely continuous function defined on some interval $[0, T]$ with $y(t) \geq 0$, and if $y(\cdot)$ satisfies the differential inequality*

$$\text{(A.2)} \qquad \dot{y}(t) \leq -c\,\kappa(y(t)) \;\; \textit{for almost all} \;\; t \in [0, T]$$

for some $c \geq 0$ with $y(0) = y_0 \geq 0$, then it holds that

$$y(t) \leq \beta(y_0, ct)$$

for all $t \in [0, T]$.

Proof. First, by Lemma 4.4 in [12], for each $\kappa \in \mathcal{K}$, there exists $\beta \in \mathcal{KL}$ such that for any locally absolutely continuous function $z(t) \geq 0$, if it satisfies the inequality

$$\dot{z}(t) \leq -\kappa(z(t))$$

on $[0,T]$, it holds that $z(t) \leq \beta(z(0),t)$ for all t. (The statement in that reference applies to z defined on all of $[0,\infty)$, but exactly the same proof works for a finite interval.)

Let $y(t)$ be a function as in the statement of the lemma for some $c > 0, T > 0$. Let $\widetilde{y}(t) = y(t/c)$. Then $\widetilde{y}$ is again locally absolutely continuous and nonnegative on $[0, cT]$. Moreover, $\widetilde{y}$ satisfies the inequality

$$\frac{d}{dt}\widetilde{y}(t) \leq -\kappa(\widetilde{y}(t)).$$

Hence,

$$\widetilde{y}(t) \leq \beta(\widetilde{y}(0), t)$$

for all $t \in [0, cT]$. This then implies that

$$y(t) \leq \beta(y(0), ct)$$

for all $t \in [0,T]$. □

Finally, we have the following fact, mentioned when discussing decrease conditions.

Lemma A.5. *Let $V : \mathbb{R}^n \to \mathbb{R}$ be a C^1 positive definition function with the following property: for some $\mathcal{K}$ function χ, it holds that*

$$V(\xi) \geq \chi(|\mu|) \quad \textit{and} \quad V(\xi) \neq 0 \ \Rightarrow \ DV(\xi)f(\xi,\mu) < 0.$$

Then, there is a function $\alpha \in \mathcal{KL}$ so that

$$V(\xi) \geq \chi(|\mu|) \ \Rightarrow \ DV(\xi)f(\xi,\mu) \leq -\alpha(V(\xi), |\xi|)$$

for all $\xi \in \mathbb{R}^n, \mu \in \mathbb{R}^m$.

Proof. Without loss of generality, we assume that $\chi \in \mathcal{K}_\infty$. Define the set for each $s, t \geq 0$:

$$R(s,t) := \{(x,u) : \ |\xi| \leq t, \ V(\xi) \geq s, \ |\mu| \leq \chi^{-1}(V(\xi))\}.$$

These sets are compact (possibly empty) for each s and t. Note the following properties:

$$s > s' \ \Rightarrow \ R(s,t) \subseteq R(s',t),$$

$$t > t' \ \Rightarrow \ R(s,t') \subseteq R(s,t).$$

Now let

$$\alpha_0(s,t) = \min_{(\xi,\mu)\in R(s,t)} -DV(\xi)f(\xi,\mu)$$

(with the convention that $\alpha_0(s,t) = +\infty$ if $R(s,t) = \emptyset$). Then, $\alpha_0(s,t)$ is nonincreasing in t and nondecreasing in s. Moreover, $\alpha(s,t) > 0$ whenever $s > 0$ (by the hypothesis of the lemma). Next let

$$\widehat{\alpha}(s,t) := \min\{\alpha_0(s,t), s\}.$$

This function has the same monotonicity properties as α_0, it satisfies $\alpha_0(s,t) \geq \widehat{\alpha}(s,t)$ for all s, t, and is finite-valued. It also satisfies $\widehat{\alpha}(s,t) \neq 0$ for $s > 0$. Now pick

$$\widetilde{\alpha}(s,t) := \int_{s-1}^{s} \widehat{\alpha}(\sigma,t)\, d\sigma$$

(let $\widehat{\alpha}(s,t) := 0$ for $s < 0$). This function still has the same monotonicity properties, satisfies $\widetilde{\alpha}(s,t) > 0$ for $s > 0$, and is continuous in s. It may not be strictly increasing in s, nor need it converge to zero as $t \to 0$, so we obtain finally a $\mathcal{KL}$ function α by defining

$$\alpha(s,t) := \frac{s\widetilde{\alpha}(s,t)}{(1+s)(1+t)}.$$

This satisfies the desired properties by construction, because

$$V(\xi) \geq \chi(|\mu|) \;\Rightarrow\; DV(\xi)f(\xi,\mu) \leq -\alpha(V(\xi),|\mu|),$$

and $\alpha_0 \geq \widehat{\alpha} \geq \widetilde{\alpha} \geq \alpha$ pointwise. □

REFERENCES

[1] D. Angeli, E. D. Sontag, and Y. Wang, *A characterization of integral input to state stability*, IEEE Trans. Automat. Control, to appear.

[2] S. Battilotti, *Robust stabilization of nonlinear systems with pointwise norm bounded uncertainties: A control Lyapunov approach*, IEEE Trans. Automat. Control, to appear.

[3] P. D. Christofides and A. Teel, *Singular perturbations and input-to-state stability*, IEEE Trans. Automat. Control, 41 (1996), pp. 1645–1650.

[4] A. Isidori, *Global almost disturbance decoupling with stability for non minimum-phase single-input single-output nonlinear systems*, Systems Control Lett., 28 (1996), pp. 115–122.

[5] A. Isidori, *Nonlinear Control Systems* II, Springer-Verlag, London, UK, 1999.

[6] Z.-P. Jiang, A. Teel, and L. Praly, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1994), pp. 95–120.

[7] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.

[8] M. Krichman, E. D. Sontag, and Y. Wang, *Input-output-to-state stability*, SIAM J. Control Optim., submitted.

[9] M. Krstić and H. Deng, *Stabilization of Uncertain Nonlinear Systems*, Springer-Verlag, London, UK, 1998.

[10] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley and Sons, New York, 1995.

[11] V. Lakshmikantham, S. Leela, and A. A. Martyuk, *Practical Stability of Nonlinear Systems*, World Scientific, River Edge, NJ, 1990.

[12] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[13] W. M. Lu, *A class of globally stabilizing controllers for nonlinear systems*, Systems Control Lett., 25 (1995), pp. 13–19.

[14] L. Praly and Y. Wang, *Stabilization in spite of matched unmodelled dynamics and an equivalent definition of input-to-state stability*, Math. Control Signals Systems, 9 (1996), pp. 1–33.

[15] R. Sepulchre, M. Jankovic, and P. V. Kokotović, *Integrator forwarding: a new recursive nonlinear robust design*, Automatica J. IFAC, 33 (1997), pp. 979–984.

[16] E. D. Sontag and Y. Wang, *On characterizations of the input-to-state stability property*, Systems Control Lett., 24 (1995), pp. 351–359.

[17] E. D. Sontag and Y. Wang, *A notion of input to output stability*, in Proceedings of the European Control Conference, Brussels, Belgium, paper WE-E A2, CD-ROM file ECC958.pdf, 1997.

[18] E. D. Sontag and Y. Wang, *Output-to-state stability and detectability of nonlinear systems*, Systems Control Lett., 29 (1997), pp. 279–290.

[19] E. D. Sontag and Y. Wang, *Notions of input to output stability*, Systems Control Lett., 38 (1999), pp. 351–359.

[20] J. Tsinias, *Sufficient Lyapunovlike conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343–357.

SIAM J. Control Optim.
Vol. 39, No. 1, pp. 250–280

GENERIC WELL-POSEDNESS OF OPTIMAL CONTROL PROBLEMS WITHOUT CONVEXITY ASSUMPTIONS[*]

ALEXANDER J. ZASLAVSKI[†]

Abstract. The Tonelli existence theorem in the calculus of variations and its subsequent modifications were established for integrands f which satisfy convexity and growth conditions. In A. J. Zaslavski [*Nonlinear Anal.*, to appear], a generic existence and uniqueness result (with respect to variations of the integrand of the integral functional) without the convexity condition was established for a class of optimal control problems satisfying the Cesari growth condition. In this paper we extend the generic existence and uniqueness result in A. J. Zaslavski [*Nonlinear Anal.*, to appear], to a class of optimal control problems in which constraint maps are also subject to variations. The main result of the paper is obtained as a realization of a variational principle extending the variational principle introduced in A. D. Ioffe and A. J. Zaslavski [*SIAM J. Control Optim.*, 38 (2000), pp. 566–581].

Key words. complete metric space, generic property, integrand, optimal control problem, set-valued mapping

AMS subject classifications. 49J99, 90C31

PII. S0363012998345391

Introduction. The Tonelli existence theorem in the calculus of variations [25] and its subsequent generalizations and extensions (e.g., [5, 6, 13, 18, 22, 24]) are based on two fundamental hypotheses concerning the behavior of the integrand as a function of the last argument (derivative): one that the integrand should grow superlinearly at infinity and the other that it should be convex (or exhibit a more special convexity property in case of a multiple integral with vector-valued functions) with respect to the last variable. Moreover, certain convexity assumptions are also necessary for properties of lower semicontinuity of integral functionals which are crucial in most of the existence proofs, although there are some interesting theorems without convexity (see [5, Ch. 16] and [2, 4, 7, 20, 21]).

In [27] it was shown that the convexity condition is not needed generically, and not only for the existence but also for the uniqueness of a solution and even for well-posedness of the problem (with respect to some natural topology in the space of integrands). Instead of considering the existence of a solution for a single integrand f, we investigated it for a space of integrands and showed that a unique solution exists for most of the integrands in the space. This approach has already been successfully applied in global analysis and the theory of dynamical systems [8, 9, 23], as well as in the calculus of variations (see, for example, [1, 16, 26]. Interesting generic existence results were obtained for particular cases of variational problems [3, 19]. In [27] this approach allowed us to establish the generic existence of solutions for a large class of optimal control problems without convexity assumptions.

More precisely, in [27] we considered a class of optimal control problems (with the same system of differential equations, the same functional constraints, and the same boundary conditions) which is identified with the corresponding complete metric space of cost functions (integrands), say $\mathcal{F}$. We did not impose any convexity assumptions. These integrands are only assumed to satisfy the Cesari growth condition. The main

[*] Received by the editors October 1, 1998; accepted for publication (in revised form) March 15, 2000; published electronically August 24, 2000.
http://www.siam.org/journals/sicon/39-1/34539.html

[†] Department of Mathematics, The Technion, Haifa 32000, Israel (ajzasl@techunix.technion.ac.il).

result in [27] establishes the existence of an everywhere dense G_δ-set $\mathcal{F}' \subset \mathcal{F}$ such that for each integrand in $\mathcal{F}'$ the corresponding optimal control problem has a unique solution. At this point we do not intend to describe the topology on $\mathcal{F}$. We only note that it is rather natural and that the set $\mathcal{F}'$ has the following property:

For each $f \in \mathcal{F}$ and each positive number ϵ there exists $g \in \mathcal{F}'$ such that $|f(t,x,u) - g(t,x,u)| \le \epsilon$ for all (t,x,u).

Here t is the independent variable, x is the state variable, and u is the control variable.

The next step in this area of research was done in [14]. There we introduced a general variational principle having its prototype in the variational principle of Deville, Godefroy, and Zizler [10]. A generic existence result in the calculus of variations without convexity assumptions was then obtained as a realization of this variational principle. It was also shown in [14] that some other generic well-posedness results in optimization theory known in the literature and their modifications are obtained as a realization of this variational principle. Note that the generic existence result in [14] was established for variational problems but not for optimal control problems and that the topologies in the spaces of integrands in [27] and [14] are different.

In this paper we suggest a modification of the variational principle in [14] which can be applied to classes of optimal control problems with various topologies in the corresponding spaces of integrands. As a realization of this principle we establish a generic existence result for a class of optimal control problems in which constraint maps are also subject to variations as well as the cost functions. More precisely, we establish a generic existence result for a class of optimal control problems (with the same system of differential equations, the same boundary conditions, and without convexity assumptions) which is identified with the corresponding complete metric space of pairs (f, U) (where f is an integrand satisfying the Cesari growth condition and U is a constraint map) endowed with some natural topology. We will show that for a generic pair (f, U) the corresponding optimal control problem has a unique solution.

To understand that the generic existence result which will be established in this paper is more complicated than its prototypes in [27] and [14], we note that for the class of optimal control problems (with the same constraint map) which is identified with the corresponding space of integrands the following properties hold [27]:

- The optimal value v_f in the optimal control problem with an integrand f depends on f continuously.
- For each integrand f and each number $\delta > 0$ there exists a neighborhood $\mathcal{U}$ of f in the space of integrands such that each $(g, \delta/2)$-optimal trajectory-control pair with some $g \in \mathcal{U}$ is (f, δ)-optimal.

Here we say that a trajectory-control pair (x, u) is (g, ϵ)-optimal if the value of the integral functional with the integrand g for (x, u) does not exceed $v_g + \epsilon$.

Clearly these properties which play an important role in [27] and [14] do not have analogs when constraint maps are also subject to variations.

In the theory developed in [27], [14] and in the present paper topologies on spaces of integrands and on spaces of integrand-map pairs are of great importance. Actually one space of integrand-map pairs, say $\mathcal{A}$, considered here is a topological product of a space of integrands and a space of multivalued maps. The values of these maps are elements of the space of all nonempty convex closed subsets of a finite-dimensional Euclidean space endowed with the Hausdorff distance. In the space of multivalued maps we consider the topology of uniform convergence. For the space of integrands we consider weak and strong topologies which induce weak and strong topologies on the space $\mathcal{A}$. We will prove the existence of a set $\mathcal{A}' \subset \mathcal{A}$ which is a countable intersection of open (in the weak topology) everywhere dense (in the strong topology) sets such that for each $(f, U) \in \mathcal{A}'$ the corresponding optimal control problem has a unique solution. In fact we will establish our result for various spaces of integrands: the space of the so-called $\mathcal{L} \bigotimes \mathcal{B}$-measurable integrands, the space of lower semicontinuous integrands, and the space of continuous integrands, as well as their subspaces consisting of integrands $f(t, x, u)$ differentiable in u and subspaces consisting of integrands $f(t, x, u)$ differentiable in x and u. All these spaces are endowed with the same weak topology which is a modification of the topologies introduced in [27] and [14]. Their strong topology is always stronger than the topology of uniform convergence.

1. Definitions and the main result. In this paper we use the following notations and definitions. Let $k \geq 1$ be an integer. We denote by $\text{mes}(E)$ the Lebesgue measure of a measurable set $E \subset R^k$, by $|\cdot|$ the Euclidean norm in R^k, and by $\langle \cdot, \cdot \rangle$ the scalar product in R^k. We use the convention that $\infty - \infty = 0$. For any $f \in C^q(R^k)$ we set

$$||f||_{C^q} = ||f||_{C^q(R^k)} = \sup_{z \in R^k} \{|\partial^{|\alpha|} f(z)/\partial x_1^{\alpha_1} \dots \partial x_k^{\alpha_k}| : \tag{1.1}$$

$$\alpha_i \geq 0 \text{ is an integer, } i = 1, \dots, k, \ |\alpha| \leq q\},$$

where $|\alpha| = \sum_{i=1}^k \alpha_i$.

For each function $f : X \to [-\infty, \infty]$, where X is nonempty, we set $\inf(f) = \inf\{f(x) : x \in X\}$. For each set-valued mapping $U : X \to 2^Y \setminus \{\emptyset\}$, where X and Y are nonempty, we set

$$\text{graph}(U) = \{(x, y) \in X \times Y : y \in U(x)\}. \tag{1.2}$$

In this paper we usually consider topological spaces with two topologies where one is weaker than the other. (Note that they can coincide.) We refer to them as the weak and the strong topologies, respectively. If (X, d) is a metric space with a metric d and $Y \subset X$, then usually Y is also endowed with the metric d (unless another metric is introduced in Y). Assume that X_1 and X_2 are topological spaces and that each of them is endowed with a weak and a strong topology. Then for the product $X_1 \times X_2$ we also introduce a pair of topologies: a weak topology which is the product of the weak topologies of X_1 and X_2 and a strong topology which is the product of the strong topologies of X_1 and X_2. If $Y \subset X_1$, then we consider the topological subspace Y with the relative weak and strong topologies (unless other topologies are introduced). If (X_i, d_i), $i = 1, 2$, are metric spaces with the metrics d_1 and d_2, respectively, then the space $X_1 \times X_2$ is endowed with the metric d defined by

$$d((x_1, x_2), (y_1, y_2)) = d_1(x_1, y_1) + d_2(x_2, y_2), \quad (x_i, y_i) \in X \times Y, \quad i = 1, 2.$$

Let $m, n, N \geq 1$ be integers. In this paper we assume that Ω is a fixed bounded domain in R^m, $H(t, x, u)$ is a fixed continuous function defined on $\Omega \times R^n \times R^N$ with values in R^{mn} such that $H(t, x, u) = (H_i)_{i=1}^n$ and $H_i = (H_{ij})_{j=1}^m$, $i = 1, \dots, n$, B_1 and B_2 are fixed nonempty closed subsets of R^n, and $\theta^* = (\theta_i^*)_{i=1}^n \in (W^{1,1}(\Omega))^n$ is also fixed. Here

$$W^{1,1}(\Omega) = \{u \in L^1(\Omega) : \partial u/\partial x_j \in L^1(\Omega), \ j = 1, \dots, m\}$$

and $W_0^{1,1}(\Omega)$ is the closure of $C_0^\infty(\Omega)$ in $W^{1,1}(\Omega)$, where $C_0^\infty(\Omega)$ is the space of smooth functions $u: \ \Omega \to R^1$ with compact support in Ω.

If $m = 1$, then we assume that $\Omega = (T_1, T_2)$, where T_1 and T_2 are fixed real numbers for which $T_1 < T_2$.

For a function $u = (u_1, \dots, u_n)$, where $u_i \in W^{1,1}(\Omega)$, $i = 1, \dots, n$, we set

$$\nabla u_i = (\partial u_i/\partial x_j)_{j=1}^m, \quad i = 1, \dots, n, \quad \nabla u = (\nabla u_i)_{i=1}^n.$$

Define set-valued mappings $\tilde{A} : \Omega \to 2^{R^n} \setminus \{\emptyset\}$ and $\tilde{U} : \Omega \times R^n \to 2^{R^N} \setminus \{\emptyset\}$ by

$$\tilde{A}(t) = R^n, \quad t \in \Omega, \quad \tilde{U}(t,x) = R^N, \quad (t,x) \in \Omega \times R^n. \tag{1.3}$$

For each $A : \Omega \to 2^{R^n} \setminus \{\emptyset\}$ and each $U : \text{graph}(A) \to 2^{R^N} \setminus \{\emptyset\}$ for which graph(U) is a closed subset of the space $\Omega \times R^n \times R^N$ with the product topology, we denote by $X(A, U)$ the set of all pairs of functions (x, u), where $x = (x_1, \dots, x_n) \in (W^{1,1}(\Omega))^n$, $u = (u_1, \dots, u_N) : \ \Omega \to R^N$ is measurable, and the following relations hold:

(1.4a)
$$x(t) \in A(t), \quad t \in \Omega \text{ almost everywhere (a.e.)}, \quad u(t) \in U(t, x(t)), \quad t \in \Omega \text{ a.e.},$$

$$\nabla x(t) = H(t, x(t), u(t)), \qquad t \in \Omega \text{ a.e.}, \tag{1.4b}$$

$$\text{if } m = 1, \quad \text{ then } x(T_i) \in B_i, \quad i = 1, 2, \tag{1.4c}$$

$$\text{if } m > 1, \quad \text{ then } x - \theta^* \in (W_0^{1,1}(\Omega))^n. \tag{1.4d}$$

Note that in the definition of the space $X(A, U)$ we use the boundary condition (1.4c) in the case $m = 1$, while in the case $m > 1$ we use the boundary condition (1.4d). Both of them are common in the literature. We do this to provide a unified treatment for both cases. Note that we prove our main result in the case $m = 1$ for a class of Bolza problems (with the same boundary condition (1.4c)), while in the case $m > 1$ it will be established for a class of Lagrange problems (with the same boundary condition (1.4d)).

To be more precise, we have to define elements of $X(A, U)$ as classes of pairs equivalent in the sense that (x_1, u_1) and (x_2, u_2) are equivalent if and only if $x_2(t) = x_1(t)$, $u_2(t) = u_1(t)$, $t \in \Omega$ a.e. If $m = 1$, then by an appropriate choice of representatives, $W^{1,1}(T_1, T_2)$ can be identified with the set of absolutely continuous functions $x : [T_1, T_2] \to R^1$, and we will henceforth assume that this has been done.

Let $A : \Omega \to 2^{R^n} \setminus \{\emptyset\}$, $U : \text{graph}(A) \to 2^{R^N} \setminus \{\emptyset\}$, and let graph$(U)$ be a closed subset of the space $\Omega \times R^n \times R^N$ with the product topology.

For the set $X(A, U)$ defined above we consider the uniformity which is determined by the following base:

$$E_X(\epsilon) = \{((x_1, u_1), (x_2, u_2)) \in X(A, U) \times X(A, U) : \tag{1.5}$$

$$\text{mes}\{t \in \Omega : \ |x_1(t) - x_2(t)| + |u_1(t) - u_2(t)| \geq \epsilon\} \leq \epsilon\},$$

where $\epsilon > 0$. It is easy to see that the uniform space $X(A, U)$ is metrizable (by a metric ρ) (see [15]). In the space $X(A, U)$ we consider the topology induced by the metric ρ.

Next we define spaces of integrands associated with the maps A and U. By $\mathcal{M}(A,U)$ we denote the set of all functions f : graph$(U) \to R^1 \cup \{\infty\}$ with the following properties:

(i) f is measurable with respect to the σ-algebra generated by products of Lebesgue measurable subsets of Ω and Borel subsets of $R^n \times R^N$;

(ii) $f(t,\cdot,\cdot)$ is lower semicontinuous for almost every $t \in \Omega$;

(iii) for each $\epsilon > 0$ there exists an integrable scalar function $\psi_\epsilon(t) \geq 0,\ t \in \Omega$, such that $|H(t,x,u)| \leq \psi_\epsilon(t) + \epsilon f(t,x,u)$ for all $(t,x,u) \in$ graph (U).

The growth condition in (iii) was proposed by Cesari (see [5]), and its equivalents and modifications are rather common in the literature. Due to the property (i) for every $f \in \mathcal{M}(A,U)$ and every $(x,u) \in X(A,U)$ the function $f(t,x(t),u(t))$, $t \in \Omega$, is measurable.

Denote by $\mathcal{M}^l(A,U)$ (respectively, $\mathcal{M}^c(A,U)$) the set of all lower semicontinuous (respectively, finite-valued continuous) functions f : graph$(U) \to R^1 \cup \{\infty\}$ in $\mathcal{M}(A,U)$. Now we equip the set $\mathcal{M}(A,U)$ with the strong and weak topologies. For the space $\mathcal{M}(A,U)$ we consider the uniformity determined by the following base:

$$E_{\mathcal{M}}(\epsilon) = \{(f,g) \in \mathcal{M}(A,U) \times \mathcal{M}(A,U) : \tag{1.6}$$

$$|f(t,x,u) - g(t,x,u)| \leq \epsilon,\ (t,x,u) \in \text{ graph}(U)\},$$

where $\epsilon > 0$. It is easy to see that the uniform space $\mathcal{M}(A,U)$ with this uniformity is metrizable (by a metric $d_{\mathcal{M}}$) and complete. This uniformity generates in $\mathcal{M}(A,U)$ the strong topology. Clearly $\mathcal{M}^l(A,U)$ and $\mathcal{M}^c(A,U)$ are closed subsets of $\mathcal{M}(A,U)$ with this topology.

For each $\epsilon > 0$ we set

$$E_{\mathcal{M}w}(\epsilon) = \{(f,g) \in \mathcal{M}(A,U) \times \mathcal{M}(A,U) : \text{ there exists a nonnegative} \tag{1.7}$$

$$\phi \in L^1(\Omega) \text{ such that } \int_\Omega \phi(t)dt \leq 1, \text{ and for almost every } t \in \Omega,$$

$$|f(t,x,u) - g(t,x,u)| < \epsilon + \epsilon \max\{|f(t,x,u)|, |g(t,x,u)|\} + \epsilon\phi(t)$$

$$\text{for each } x \in A(t) \text{ and each } u \in U(t,x)\}.$$

Using the following simple lemma we can easily show that for the set $\mathcal{M}(A,U)$ there exists the uniformity which is determined by the base $E_{\mathcal{M}w}(\epsilon)$, $\epsilon > 0$. This uniformity induces in $\mathcal{M}(A,U)$ the weak topology.

Lemma 1.1. *Let $a, b \in R^1$, $\epsilon \in (0,1)$, $\Delta \geq 0$, and*

$$|a - b| < (1+\Delta)\epsilon + \epsilon \max\{|a|, |b|\}.$$

Then

$$|a - b| < (1+\Delta)(\epsilon + \epsilon^2(1-\epsilon)^{-1}) + \epsilon(1-\epsilon)^{-1}\min\{|a|, |b|\}.$$

Denote by $C_l(B_1 \times B_2)$ the set of all lower semicontinuous functions $\xi : B_1 \times B_2 \to R^1 \cup \{\infty\}$ bounded from below. We also equip the set $C_l(B_1 \times B_2)$ with strong and weak topologies. For the set $C_l(B_1 \times B_2)$ we consider the uniformity determined by the following base:

(1.8)
$$E_c(\epsilon) = \{(\xi, h) \in C_l(B_1 \times B_2) \times C_l(B_1 \times B_2) : \ |\xi(z) - h(z)| \le \epsilon, \ z \in B_1 \times B_2\},$$

where $\epsilon > 0$. It is easy to see that the uniform space $C_l(B_1 \times B_2)$ is metrizable (by a metric d_c) and complete. This metric induces in $C_l(B_1 \times B_2)$ the strong topology. We do not write down the explicit expressions for the metrics $d_{\mathcal{M}}$ and d_c because we are not going to use them in what follows.

For any $\epsilon > 0$ we set

$$E_{cw}(\epsilon) = \{(\xi, h) \in C_l(B_1 \times B_2) \times C_l(B_1 \times B_2) : \ |\xi(z) - h(z)| \tag{1.9}$$

$$< \epsilon + \epsilon \max\{|\xi(z)|, |h(z)|\}, \ z \in B_1 \times B_2\},$$

where $\epsilon > 0$. By using Lemma 1.1 we can easily show that for the set $C_l(B_1 \times B_2)$ there exists a uniformity which is determined by the base $E_{cw}(\epsilon)$, $\epsilon > 0$. This uniformity induces in $C_l(B_1 \times B_2)$ the weak topology. Denote by $C(B_1 \times B_2)$ the set of all finite-valued continuous functions h in $C_l(B_1 \times B_2)$. Clearly it is a closed subset of $C_l(B_1 \times B_2)$ with the weak topology.

In the case $m > 1$ for each $f \in \mathcal{M}(A, U)$ we define $I^{(f)} : \ X(A, U) \to R^1 \cup \{\infty\}$ by

$$I^{(f)}(x, u) = \int_{\Omega} f(t, x(t), u(t)) dt, \qquad (x, u) \in X(A, U). \tag{1.10}$$

In the case $m = 1$ for each $f \in \mathcal{M}(A, U)$ and each $\xi \in C_l(B_1 \times B_2)$ we define $I^{(f,\xi)} : \ X(A, U) \to R^1 \cup \{\infty\}$ by

$$I^{(f,\xi)}(x, u) = \int_{T_1}^{T_2} f(t, x(t), u(t)) dt + \xi(x(T_1), x(T_2)), \quad (x, u) \in X(A, U). \tag{1.11}$$

We will show (see Propositions 4.1 and 4.2) that in both cases (1.10) and (1.11) define lower semicontinuous functionals on $X(A, U)$.

From now on in this section we consider a fixed set-valued mapping $A : \Omega \to 2^{R^n} \setminus \{\emptyset\}$ for which graph(A) is a closed subset of the space $\Omega \times R^n$ with the product topology. Denote by $\tilde{U}_A$ the restriction of $\tilde{U}$ (see (1.3)) to the graph(A). Namely,

$$\tilde{U}_A : \ \text{graph}(A) \to 2^{R^N}, \quad \tilde{U}(t, x) = R^N, \quad (t, x) \in \ \text{graph}(A). \tag{1.12}$$

We consider functionals $I^{(f,\xi)}$ with $(f, \xi) \in \mathcal{M}(A, \tilde{U}_A) \times C_l(B_1 \times B_2)$ (in the case $m = 1$) and functionals $I^{(f)}$ with $f \in \mathcal{M}(A, \tilde{U}_A)$ (in the case $m > 1$) defined on the space $X(A, \tilde{U}_A)$ (see (1.4)). As we have already noted in the introduction our main result will be established for several classes of optimal control problems with different corresponding spaces of the integrands which are subsets of the space $\mathcal{M}(A, \tilde{U}_A)$. The subspaces of lower semicontinuous and continuous integrands ($\mathcal{M}^l(A, \tilde{U}_A)$ and $\mathcal{M}^c(A, \tilde{U}_A)$) have already been defined. Now we define subspaces of $\mathcal{M}(A, \tilde{U}_A)$ which consist of integrands differentiable with respect to the control variable u.

Let $k \geq 1$ be an integer. Denote by $\mathcal{M}_k(A, \tilde{U}_A)$ the set of all finite-valued $f \in \mathcal{M}(A, \tilde{U}_A)$ such that for each $(t, x) \in \text{graph}(A)$ the function $f(t, x, \cdot) \in C^k(R^N)$. We consider the topological subspace $\mathcal{M}_k(A, \tilde{U}_A) \subset \mathcal{M}(A, \tilde{U}_A)$ with the relative weak topology. The strong topology on $\mathcal{M}_k(A, \tilde{U}_A)$ is induced by the uniformity which is determined by the following base:

$$(1.13) \quad E_{\mathcal{M}k}(\epsilon) = \{(f, g) \in \mathcal{M}_k(A, \tilde{U}_A) \times \mathcal{M}_k(A, \tilde{U}_A) : \ |f(t, x, u) - g(t, x, u)| \leq \epsilon$$

$$\text{for all } (t, x, u) \in \ \text{graph}(A) \times R^N \text{ and}$$

$$||f(t, x, \cdot) - g(t, x, \cdot)||_{C^k(R^N)} \leq \epsilon \text{ for all } (t, x) \in \text{graph}(A)\},$$

where $\epsilon > 0$. It is easy to see that the space $\mathcal{M}_k(A, \tilde{U}_A)$ with this uniformity is metrizable (by a metric $d_{\mathcal{M},k}$) and complete. Define

$$(1.14) \quad \mathcal{M}_k^l(A, \tilde{U}_A) = \mathcal{M}_k(A, \tilde{U}_A) \cap \mathcal{M}^l(A, \tilde{U}_A), \ \mathcal{M}_k^c(A, \tilde{U}_A) = \mathcal{M}_k(A, \tilde{U}_A) \cap \mathcal{M}^c(A, \tilde{U}_A).$$

Clearly $\mathcal{M}_k^l(A, \tilde{U}_A)$ and $\mathcal{M}_k^c(A, \tilde{U}_A)$ are closed sets in $\mathcal{M}_k(A, \tilde{U}_A)$ with the strong topology.

Finally, we define subspaces of $\mathcal{M}(\tilde{A}, \tilde{U})$ which consist of integrands differentiable with respect to the state variable x and the control variable u. Denote by $\mathcal{M}_k^*(\tilde{A}, \tilde{U})$ the set of all $f : \Omega \times R^n \times R^N \to R^1$ in $\mathcal{M}(\tilde{A}, \tilde{U})$ (see (1.3)) such that for each $t \in \Omega$ the function $f(t, \cdot, \cdot) \in C^k(R^n \times R^N)$. We consider the topological subspace $\mathcal{M}_k^*(\tilde{A}, \tilde{U}) \subset \mathcal{M}(\tilde{A}, \tilde{U})$ with the relative weak topology. The strong topology in $\mathcal{M}_k^*(\tilde{A}, \tilde{U})$ is induced by the uniformity which is determined by the following base:

$$(1.15) \qquad E^*_{\mathcal{M}k}(\epsilon) = \{(f, g) \in \mathcal{M}_k^*(\tilde{A}, \tilde{U}) \times \mathcal{M}_k^*(\tilde{A}, \tilde{U}) :$$

$$|f(t, x, u) - g(t, x, u)| \leq \epsilon \text{ for all } (t, x, u) \in \Omega \times R^n \times R^N \text{ and}$$

$$||f(t, \cdot, \cdot) - g(t, \cdot, \cdot)||_{C^k(R^{n+N})} \leq \epsilon \text{ for all } t \in \Omega\},$$

where $\epsilon > 0$. It is easy to see that the space $\mathcal{M}_k^*(\tilde{A}, \tilde{U})$ with this uniformity is metrizable (by a metric $d^*_{\mathcal{M},k}$) and complete. Define

$$(1.16) \quad \mathcal{M}_k^{*l}(\tilde{A}, \tilde{U}) = \mathcal{M}_k^*(\tilde{A}, \tilde{U}) \cap \mathcal{M}^l(\tilde{A}, \tilde{U}), \ \mathcal{M}_k^{*c}(\tilde{A}, \tilde{U}) = \mathcal{M}_k^*(\tilde{A}, \tilde{U}) \cap \mathcal{M}^c(\tilde{A}, \tilde{U}).$$

Clearly $\mathcal{M}_k^{*l}(\tilde{A}, \tilde{U})$ and $\mathcal{M}_k^{*c}(\tilde{A}, \tilde{U})$ are closed sets in $\mathcal{M}_k^*(\tilde{A}, \tilde{U})$ with the strong topology.

Thus we have defined all the spaces of integrands for which we will prove our main result. Now we will define a space of constraint maps $\mathcal{P}_A$. Denote by $S(R^N)$ the set of all nonempty convex closed subsets of R^N. For each $x \in R^N$ and each $E \subset R^N$, set $d_H(x, E) = \inf_{y \in E} |x - y|$. For each pair of sets $C_1, C_2 \subset R^N$,

$$d_H(C_1, C_2) = \max \left\{ \sup_{y \in C_1} d_H(y, C_2), \ \sup_{x \in C_2} d_H(x, C_1) \right\}$$

is the Hausdorff distance between C_1 and C_2. For the space $S(R^N)$ we consider the uniformity determined by the following base:

$$E_{R^N}(\epsilon) = \{(C_1, C_2) \in S(R^N) \times S(R^N) : \ d_H(C_1, C_2) \le \epsilon\}, \tag{1.17}$$

where $\epsilon > 0$. It is well known that the space $S(R^N)$ with this uniformity is metrizable and complete. Denote by $\mathcal{P}_A$ the set of all set-valued mappings $U : \text{graph}(A) \to S(R^N)$ such that graph(U) is a closed subset of the space graph$(A) \times R^N$ with the product topology. For the space $\mathcal{P}_A$ we consider the uniformity determined by the following base:

$$\begin{aligned} E_{\mathcal{P}_A}(\epsilon) = \{(U_1, U_2) \in \mathcal{P}_A \times \mathcal{P}_A : \ & d_H(U_1(t,x), U_2(t,x)) \le \epsilon \\ & \text{for all } (t,x) \in \text{graph}(A)\}, \end{aligned} \tag{1.18}$$

where $\epsilon > 0$. It is easy to see that the space $\mathcal{P}_A$ with this uniformity is metrizable and complete.

We consider the space $X(A, \tilde{U}_A)$ with the metric ρ (see (1.5)). For each $U \in \mathcal{P}_A$ define

$$S_U = X(A, U) = \{(x, u) \in X(A, \tilde{U}_A) : \ u(t) \in U(t, x(t)), \ t \in \Omega \text{ a.e.}\}. \tag{1.19}$$

In the case $m = 1$ for each $U \in \mathcal{P}_A$ and each $(f, \xi) \in \mathcal{M}(A, \tilde{U}_A) \times C_l(B_1 \times B_2)$ we consider the optimal control problem

$$I^{(f,\xi)}(x, u) \to \min, \qquad (x, u) \in X(A, U),$$

and in the case $m > 1$ for each $U \in \mathcal{P}_A$ and each $f \in \mathcal{M}(A, \tilde{U}_A)$ we consider the optimal control problem

$$I^{(f)}(x, u) \to \min, \qquad (x, u) \in X(A, U).$$

We will state our main result, Theorem 1.1, in such a manner that it will be applicable to the Bolza problem in case $m = 1$ and to the Lagrange problem in case $m > 1$, and also applicable for all the spaces of integrands defined above.

To meet this goal we set $\mathcal{A}_2 = \mathcal{P}_A$ and define a space $\mathcal{A}_1$ as follows.

$$\mathcal{A}_1 = \mathcal{A}_{11} \times \mathcal{A}_{12} \quad \text{if } m = 1 \quad \text{and} \quad \mathcal{A}_1 = \mathcal{A}_{11} \quad \text{if } m > 1,$$

where $\mathcal{A}_{12}$ is either $C_l(B_1 \times B_2)$ or $C(B_1 \times B_2)$ or a singleton $\{\xi\} \subset C_l(B_1 \times B_2)$, and $\mathcal{A}_{11}$ is one of the following spaces:

$$\mathcal{M}(A, \tilde{U}_A); \ \mathcal{M}^l(A, \tilde{U}_A); \ \mathcal{M}^c(A, \tilde{U}_A);$$

$$\mathcal{M}_k(A, \tilde{U}_A); \ \mathcal{M}_k^l(A, \tilde{U}_A); \ \mathcal{M}_k^c(A, \tilde{U}_A) \ \text{ (here } k \ge 1 \text{ is an integer);}$$

$$\mathcal{M}_k^*(\tilde{A}, \tilde{U}); \ \mathcal{M}_k^{*l}(\tilde{A}, \tilde{U}); \ \mathcal{M}_k^{*c}(\tilde{A}, \tilde{U}) \text{ (here } k \ge 1 \text{ is an integer and } A = \tilde{A}).$$

For each $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ we define $J_a : X(A, \tilde{U}_A) \to R^1 \cup \{\infty\}$ by

$$J_a(x, u) = I^{(a_1)}(x, u), \quad (x, u) \in S_{a_2}, \quad J_a(x, u) = \infty, \quad (x, u) \in X(A, \tilde{U}_A) \setminus S_{a_2}.$$

We will show that J_a is lower semicontinuous for all $a \in \mathcal{A}_1 \times A_2$. Denote by $\mathcal{A}$ the closure of the set $\{a \in \mathcal{A}_1 \times \mathcal{A}_2 : \inf(J_a) < \infty\}$ in the space $\mathcal{A}_1 \times A_2$ with the strong topology. We assume that $\mathcal{A}$ is nonempty. The following theorem is the main result of this paper.

Theorem 1.1. *There exists an everywhere dense (in the strong topology) set $\mathcal{B} \subset \mathcal{A}$ which is a countable intersection of open (in the weak topology) subsets of $\mathcal{A}$ such that for any $a \in \mathcal{B}$ the following assertions hold:*

(1) $\inf(J_a)$ *is finite and attained at a unique pair* $(\bar{x}, \bar{u}) \in X(A, \tilde{U}_A)$.

(2) *For each $\epsilon > 0$ there are a neighborhood $\mathcal{V}$ of a in $\mathcal{A}$ with the weak topology and $\delta > 0$ such that for each $b \in \mathcal{V}$, $\inf(J_b)$ is finite, and if $(z, w) \in X(A, \tilde{U}_A)$ satisfies $J_b(z, w) \le \inf(J_b) + \delta$, then $\rho((\bar{x}, \bar{u}), (z, w)) \le \epsilon$ and $|J_b(z, w) - J_a(\bar{x}, \bar{u})| \le \epsilon$.*

2. Generic variational principle. We will obtain our main result as a realization of a variational principle which will be introduced in this section. This variational principle is a modification of the variational principle in [14].

We consider a metric space (X, ρ) which is called the domain space and a complete metric space $(\mathcal{A}, d)$ which is called the data space. We always consider the set X with the topology generated by the metric ρ. For the space $\mathcal{A}$ we consider the topology generated by the metric d. This topology will be called the strong topology. As mentioned in section 1 in addition to the strong topology we also consider a weaker topology on $\mathcal{A}$ which is not necessarily Hausdorff. This topology will be called the weak topology. (Note that these topologies can coincide.) We assume that with every $a \in \mathcal{A}$ a lower semicontinuous function f_a on X is associated with values in $\bar{R} = [-\infty, \infty]$. In our study we use the following basic hypotheses about the functions.

(H1) For any $a \in \mathcal{A}$, any $\epsilon > 0$, and any $\gamma > 0$ there exist a nonempty open set $\mathcal{W}$ in $\mathcal{A}$ with the weak topology, $x \in X$, $\alpha \in R^1$, and $\eta > 0$ such that

$$\mathcal{W} \cap \{b \in \mathcal{A} : \ d(a, b) < \epsilon\} \ne \emptyset,$$

and for any $b \in \mathcal{W}$

(i) $\inf(f_b)$ is finite;

(ii) if $z \in X$ is such that $f_b(z) \le \inf(f_b) + \eta$, then $\rho(z, x) \le \gamma$ and $|f_b(z) - \alpha| \le \gamma$.

(H2) If $a \in \mathcal{A}$, $\inf(f_a)$ is finite, $\{x_n\}_{n=1}^{\infty} \subset X$ is a Cauchy sequence, and the sequence $\{f_a(x_n)\}_{n=1}^{\infty}$ is bounded, then the sequence $\{x_n\}_{n=1}^{\infty}$ converges in X.

We will show (see Theorem 2.1) that if (H1) and (H2) hold, then for a generic $a \in \mathcal{A}$ the minimization problem $f_a(x) \to \min$, $x \in X$, has a unique solution. This result generalizes the variational principle in [14, Theorem 2.2] which was obtained for the complete domain space (X, ρ). Note that if (X, ρ) is complete, the weak and strong topologies on $\mathcal{A}$ coincide, and for any $a \in \mathcal{A}$ the function f_a is not identically ∞, then the variational principles in [14] and in this section are equivalent.

For the classes of optimal control problems considered in this paper the domain space is usually the space $X(A, \tilde{U}_A)$ with the metric ρ (see (1.5)) which is not complete. Since the variational principle in [14] was established only for complete domain spaces it cannot be applied to these classes of optimal control problems. Fortunately, instead of the completeness assumption we can use (H2), and this hypothesis holds for spaces of integrands (integrand-map pairs) which satisfy the Cesari growth condition.

Theorem 2.1. *Assume that* (H1) *and* (H2) *hold. Then there exists an everywhere dense (in the strong topology) set $\mathcal{B} \subset \mathcal{A}$ which is a countable intersection of open (in the weak topology) subsets of $\mathcal{A}$ such that for any $a \in \mathcal{B}$ the following assertions hold:*

(1) $\inf(f_a)$ *is finite and attained at a unique point* $\bar{x} \in X$.

(2) *For each* $\epsilon > 0$ *there are a neighborhood* $\mathcal{V}$ *of* a *in* $\mathcal{A}$ *with the weak topology and* $\delta > 0$ *such that for each* $b \in \mathcal{V}$, $\inf(f_b)$ *is finite and if* $z \in X$ *satisfies* $f_b(z) \leq \inf(f_b) + \delta$, *then* $\rho(\bar{x}, z) \leq \epsilon$ *and* $|f_b(z) - f_a(\bar{x})| \leq \epsilon$.

Following the tradition, we can summarize the theorem by saying that under the assumptions (H1) and (H2) the minimization problem for f_a on (X, ρ) *is generically strongly well posed with respect to* $\mathcal{A}$.

Proof. Let $a \in \mathcal{A}$. By (H1) for any natural $n = 1, 2, \ldots$ there are a nonempty open set $\mathcal{U}(a, n)$ in $\mathcal{A}$ with the weak topology, $x(a, n) \in X$, $\alpha(a, n) \in R^1$, and $\eta(a, n) > 0$ such that

$$\mathcal{U}(a, n) \cap \{b \in \mathcal{A} :\ d(a, b) < 1/n\} \neq \emptyset,$$

and for any $b \in \mathcal{U}(a, n)$, $\inf(f_b)$ is finite and if $z \in X$ satisfies $f_b(z) \leq \inf(f_b) + \eta(a, n)$, then

$$\rho(z, x(a, n)) \leq 1/n, \qquad |f_b(z) - \alpha(a, n)| \leq 1/n.$$

Define $\mathcal{B}_n = \cup\{\mathcal{U}(a, m) :\ a \in \mathcal{A},\ m \geq n\}$ for $n = 1, 2, \ldots$. Clearly, for each integer $n \geq 1$ the set $\mathcal{B}_n$ is open in the weak topology and everywhere dense in the strong topology. Set $\mathcal{B} = \cap_{n=1}^{\infty} \mathcal{B}_n$. Since for each integer $n \geq 1$ the set $\mathcal{B}_n$ is also open in the strong topology generated by the complete metric d we conclude that $\mathcal{B}$ is everywhere dense in the strong topology.

Let $b \in \mathcal{B}$. Evidently $\inf(f_b)$ is finite. There are a sequence $\{a_n\}_{n=1}^{\infty} \subset \mathcal{A}$ and a strictly increasing sequence of natural numbers $\{k_n\}_{n=1}^{\infty}$ such that $b \in \mathcal{U}(a_n, k_n)$, $n = 1, 2, \ldots$. Assume that $\{z_n\}_{n=1}^{\infty} \subset X$ and $\lim_{n\to\infty} f_b(z_n) = \inf(f_b)$.

Let $m \geq 1$ be an integer. Clearly for all large enough n the inequality $f_b(z_n) < \inf(f_b) + \eta(a_m, k_m)$ is true and it follows from the definition of $\mathcal{U}(a_m, k_m)$ that

$$\rho(z_n, x(a_m, k_m)) \leq k_m^{-1},\ |f_b(z_n) - \alpha(a_m, k_m)| \leq k_m^{-1} \tag{2.1}$$

for all large enough n. Since m is an arbitrary natural number we conclude that $\{z_n\}_{n=1}^{\infty} \subset X$ is a Cauchy sequence. By (H2) there is an $\bar{x} = \lim_{n\to\infty} z_n$. As f_b is lower semicontinuous, we have $f_b(\bar{x}) = \inf(f_b)$. Clearly f_b does not have another minimizer, for otherwise we would be able to construct a nonconvergent sequence $\{z_n\}_{n=1}^{\infty}$. This proves the first part of the theorem. We further note that by (2.1)

$$\rho(\bar{x}, x(a_m, k_m)) \leq k_m^{-1}, \quad |f_b(\bar{x}) - \alpha(a_m, k_m)| \leq k_m^{-1}, \quad m = 1, 2, \ldots. \tag{2.2}$$

We turn now to the second assertion. Let $\epsilon > 0$. Choose a natural number m for which $4k_m^{-1} < \epsilon$. Let $a \in \mathcal{U}(a_m, k_m)$. Clearly $\inf(f_a)$ is finite. Let $z \in X$ and $f_a(z) \leq \inf(f_a) + \eta(a_m, k_m)$. By the definition of $\mathcal{U}(a_m, k_m)$,

$$\rho(z, x(a_m, k_m)) \leq k_m^{-1}, \qquad |f_a(z) - \alpha(a_m, k_m)| \leq k_m^{-1}.$$

Together with (2.2) this implies that

$$\rho(z, \bar{x}) \leq 2k_m^{-1}, \quad |f_b(\bar{x}) - f_a(z)| \leq 2k_m^{-1} < \epsilon.$$

The second assertion is proved. □

3. Concretization of the hypothesis (H1). The proof of our main result consists in verifying that the hypotheses (H1) and (H2) hold for the space of integrand-map pairs introduced in section 1. Hypothesis (H2) will follow from Proposition 4.2, which will be proved in section 4. The verification of (H1) is more complicated. Recall that our space of integrand-map pairs is a product of the space of integrands and the space of maps. Therefore we should seek the set $\mathcal{W}$ (see (H1)) in the form $\mathcal{V} \times \mathcal{U}$, where $\mathcal{V}$ is an open set in the space of integrands and $\mathcal{U}$ is an open set in the space of maps. To simplify the verification of (H1) in this section we introduce new assumptions (A1)–(A4) and show that they imply (H1) (see Proposition 3.1). Using (A1)–(A4) we can construct the set $\mathcal{W} = \mathcal{V} \times \mathcal{U}$ step by step, roughly speaking. Namely, using (A4) we construct the set $\mathcal{U}$, using (A3) we find an integrand $\bar{a}_1$, and then using (A2) we construct the set $\mathcal{V}$, which is an open neighborhood of $\bar{a}_1$. Thus to verify (H1) we need to show that the assumptions (A1)–(A4) are valid. In fact this approach allows us to simplify the problem because each of (A2)–(A4) concerns either the space of integrands or the space of maps while it is not difficult to verify (A1).

Let (X, ρ) be a metric space with the topology generated by the metric ρ and let $(\mathcal{A}_1, d_1)$, $(\mathcal{A}_2, d_2)$ be metric spaces. For the space $\mathcal{A}_i$ $(i = 1, 2)$ we consider the topology generated by the metric d_i. This topology is called the strong topology. In addition to the strong topology we consider a weak topology on $\mathcal{A}_i$, $i = 1, 2$.

Assume that with every $a \in \mathcal{A}_1$ a lower semicontinuous function $\phi_a : X \to R^1 \cup \{\infty\}$ is associated and with every $a \in \mathcal{A}_2$ a set $S_a \subset X$ is associated. For each $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ define $f_a : X \to R^1 \cup \{\infty\}$ by

$$f_a(x) = \phi_{a_1}(x) \quad \text{for all } x \in S_{a_2}, \quad f_a(x) = \infty \quad \text{for all } x \in X \setminus S_{a_2}. \tag{3.1}$$

Denote by $\mathcal{A}$ the closure of the set $\{a \in \mathcal{A}_1 \times \mathcal{A}_2 : \inf(f_a) < \infty\}$ in the space $\mathcal{A}_1 \times \mathcal{A}_2$ with the strong topology. We assume that $\mathcal{A}$ is nonempty.

In this paper we use the following hypotheses:

(A1) For each $a_1 \in \mathcal{A}_1$, $\inf(\phi_{a_1}) > -\infty$ and for each $a \in \mathcal{A}_1 \times \mathcal{A}_2$ the function f_a is lower semicontinuous.

(A2) For each $a \in \mathcal{A}_1$ and each $D, \epsilon > 0$ there is a neighborhood $\mathcal{U}$ of a in $\mathcal{A}_1$ with the weak topology such that for each $b \in \mathcal{U}$ and each $x \in X$ satisfying $\min\{\phi_a(x), \phi_b(x)\} \le D$ the relation $|\phi_a(x) - \phi_b(x)| \le \epsilon$ holds.

(A3) For each $\gamma \in (0, 1)$ there exist positive numbers $\epsilon(\gamma)$ and $\delta(\gamma)$ such that $\epsilon(\gamma), \delta(\gamma) \to 0$ as $\gamma \to 0$ and the following property holds.

For each $\gamma \in (0, 1)$, each $a \in \mathcal{A}_1$, each nonempty set $Y \subset X$, and each $\bar{x} \in Y$ for which

$$\phi_a(\bar{x}) \le \inf\{\phi_a(z) : z \in Y\} + \delta(\gamma) < \infty, \tag{3.2}$$

there is an $\bar{a} \in \mathcal{A}_1$ such that the following conditions hold:

$$d_1(a, \bar{a}) \le \epsilon(\gamma), \quad \phi_{\bar{a}}(z) \ge \phi_a(z), \quad z \in X, \quad \phi_{\bar{a}}(\bar{x}) \le \phi_a(\bar{x}) + \delta(\gamma); \tag{3.3}$$

for each $y \in Y$ satisfying

$$\phi_{\bar{a}}(y) \le \inf\{\phi_{\bar{a}}(z) : z \in Y\} + 2\delta(\gamma) \tag{3.4}$$

the inequality $\rho(y, \bar{x}) \le \gamma$ is valid.

(A4) For each $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ satisfying $\inf(f_a) < \infty$ and each $\epsilon, \delta > 0$ there exist $\bar{a}_2 \in \mathcal{A}_2$, $\bar{x} \in S_{\bar{a}_2}$, and an open set $\mathcal{U}$ in $\mathcal{A}_2$ with the weak topology such that

$$d_2(a_2, \bar{a}_2) < \epsilon, \qquad \mathcal{U} \cap \{b \in \mathcal{A}_2 : \ d_2(b, a_2) < \epsilon\} \neq \emptyset, \tag{3.5}$$

$$\phi_{a_1}(\bar{x}) \leq \inf\{\phi_{a_1}(z) : \ z \in S_{\bar{a}_2}\} + \delta < \infty, \tag{3.6}$$

and

$$\bar{x} \in S_b \subset S_{\bar{a}_2} \quad \text{for all } b \in \mathcal{U}. \tag{3.7}$$

Assume that (A3) holds. We show that the numbers $\epsilon(\gamma)$ and $\delta(\gamma)$ can be chosen such that $0 < \delta(\gamma) \leq \epsilon(\gamma) \leq \gamma$.

Let $\epsilon(\gamma)$ and $\delta(\gamma)$, $\gamma \in (0,1)$, be as guaranteed by (A3). Assume that $\gamma \in (0,1)$. Since $\lim_{t\to 0} \epsilon(t) = 0$ and $\lim_{t \to 0} \delta(t) = 0$ there exist $\gamma_1 \in (0, \gamma)$ and $\gamma_0 \in (0, \gamma_1)$ such that $\epsilon(\gamma_1) < \gamma$ and $\epsilon(\gamma_0), \delta(\gamma_0) < \epsilon(\gamma_1)$. Set $\bar{\epsilon}(\gamma) = \epsilon(\gamma_1)$ and $\bar{\delta}(\gamma) = \delta(\gamma_0)$. Clearly $\bar{\delta}(\gamma) < \bar{\epsilon}(\gamma) < \gamma$.

Assume that $a \in \mathcal{A}_1$, Y is a nonempty subset of X, and $\bar{x} \in Y$ satisfies $\phi_a(\bar{x}) \leq \inf\{\phi_a(z) : \ z \in Y\} + \bar{\delta}(\gamma) < \infty$. By (A3) and the equality $\bar{\delta}(\gamma) = \delta(\gamma_0)$ there exists $\bar{a} \in \mathcal{A}_1$ such that the following conditions hold:

$$d_1(a, \bar{a}) \leq \epsilon(\gamma_0) < \epsilon(\gamma_1) = \bar{\epsilon}(\gamma), \quad \phi_{\bar{a}}(z) \geq \phi_a(z), \quad z \in X,$$

$$\phi_{\bar{a}}(\bar{x}) \leq \phi_a(\bar{x}) + \delta(\gamma_0) = \phi_a(\bar{x}) + \bar{\delta}(\gamma);$$

for each $y \in Y$ satisfying

$$\phi_{\bar{a}}(y) \leq \inf\{\phi_{\bar{a}}(z) : \ z \in Y\} + 2\delta(\gamma_0)$$

the inequality $\rho(y, \bar{x}) \leq \gamma_0 \leq \gamma$ is valid. Therefore (A3) holds with $\epsilon(\gamma) = \bar{\epsilon}(\gamma)$ and $\delta(\gamma) = \bar{\delta}(\gamma)$.

PROPOSITION 3.1. *Assume that* (A1)–(A4) *hold. Then* (H1) *holds for the space* $\mathcal{A}$.

Proof. Let $a = (a_1, a_2) \in \mathcal{A}$ and let $\epsilon, \gamma > 0$. We may assume that $\inf(f_a) < \infty$. Choose a positive number

$$\gamma_0 < 8^{-1} \min\{1, \epsilon, \gamma\}. \tag{3.8}$$

Let $\epsilon(\gamma_0)$, $\delta(\gamma_0) > 0$ be as guaranteed by (A3) (namely, (A3) is true with $\gamma = \gamma_0$, $\epsilon(\gamma) = \epsilon(\gamma_0)$, $\delta(\gamma) = \delta(\gamma_0)$). Choose

$$\delta_1 \in (0, 4^{-1}\delta(\gamma_0)). \tag{3.9}$$

By (A4) there are $\bar{a}_2 \in \mathcal{A}_2$, $\bar{x} \in S_{\bar{a}_2}$, and an open nonempty set $\mathcal{U}$ in $\mathcal{A}_2$ with the weak topology such that (3.7) holds,

$$d_2(a_2, \bar{a}_2) < \epsilon(\gamma_0), \qquad \mathcal{U} \cap \{b \in \mathcal{A}_2 : \ d_2(b, a_2) < \epsilon(\gamma_0)\} \neq \emptyset, \tag{3.10}$$

and

$$\phi_{a_1}(\bar{x}) \leq \inf\{\phi_{a_1}(z) : \ z \in S_{\bar{a}_2}\} + \delta_1 < \infty. \tag{3.11}$$

It follows from the definition of $\epsilon(\gamma_0)$ and $\delta(\gamma_0)$, (A3) (with $a_1 = a$ and $Y = S_{\bar{a}_2}$), and (3.11) that there is an $\bar{a}_1 \in \mathcal{A}_1$ such that

$$d_1(a_1, \bar{a}_1) \le \epsilon(\gamma_0), \quad \phi_{\bar{a}_1}(z) \ge \phi_{a_1}(z), \quad z \in X, \tag{3.12}$$

$$\phi_{\bar{a}_1}(\bar{x}) \le \phi_{a_1}(\bar{x}) + \delta(\gamma_0),$$

and the following property holds.

(Pi) For each $y \in S_{\bar{a}_2}$ satisfying

$$\phi_{\bar{a}_1}(y) \le \inf\{\phi_{\bar{a}_1}(z) : \ z \in S_{\bar{a}_2}\} + 2\delta(\gamma_0) \tag{3.13}$$

the relation $\rho(y, \bar{x}) \le \gamma_0$ is valid.

Let $b \in \mathcal{U}$. Then by the definition of $\mathcal{U}$, (3.7), and (3.11),

$$\bar{x} \in S_b \subset S_{\bar{a}_2}, \qquad \inf\{\phi_{a_1}(z) : \ z \in S_b\} \le \phi_{a_1}(\bar{x}) < \infty. \tag{3.14}$$

We will show that the following property holds.

(Pii) If $y \in S_b$ satisfies

$$\phi_{\bar{a}_1}(y) \le \inf\{\phi_{\bar{a}_1}(z) : \ z \in S_b\} + \delta_1, \tag{3.15}$$

then

$$\rho(y, \bar{x}) \le \gamma_0 \quad \text{and} \quad |\phi_{\bar{a}_1}(y) - \phi_{\bar{a}_1}(\bar{x})| \le \delta_1 + \delta(\gamma_0). \tag{3.16}$$

It follows from (3.11), (3.14), and (3.12) that

$$\phi_{a_1}(\bar{x}) - \delta_1 \le \inf\{\phi_{a_1}(z) : \ z \in S_{\bar{a}_2}\} \le \inf\{\phi_{a_1}(z) : \ z \in S_b\} \tag{3.17}$$

$$\le \inf\{\phi_{\bar{a}_1}(z) : \ z \in S_b\} \le \phi_{\bar{a}_1}(\bar{x}) \le \phi_{a_1}(\bar{x}) + \delta(\gamma_0)$$

$$\le \inf\{\phi_{a_1}(z) : \ z \in S_{\bar{a}_2}\} + \delta_1 + \delta(\gamma_0).$$

Assume that $y \in S_b$ and (3.15) is true. It follows from (3.14), (3.15), (3.17), (3.12), and (3.9) that

$$y \in S_{\bar{a}_2}, \ \phi_{\bar{a}_1}(y) \le \inf\{\phi_{a_1}(z) : \ z \in S_{\bar{a}_2}\} + \delta(\gamma_0) + 2\delta_1$$

$$< \inf\{\phi_{\bar{a}_1}(z) : \ z \in S_{\bar{a}_2}\} + 2\delta(\gamma_0).$$

By these relations and property (Pi), $\rho(y, \bar{x}) \le \gamma_0$. Relations (3.15), (3.17), (3.11), (3.14), and (3.12) imply that

$$|\phi_{\bar{a}_1}(y) - \phi_{\bar{a}_1}(\bar{x})| \le \delta_1 + \delta(\gamma_0). \tag{3.18}$$

Thus, (3.16) is valid. Therefore we have shown that for each $b \in \mathcal{U}$ relation (3.14) and property (Pii) hold. Choose a number

$$D > |\inf(\phi_{\bar{a}_1})| + 1 + |\phi_{\bar{a}_1}(\bar{x})|. \tag{3.19}$$

By (A2) there exists an open neighborhood $\mathcal{V}$ of $\bar{a}_1$ in $\mathcal{A}_1$ with the weak topology such that the following property holds.

(Piii) For each $b \in \mathcal{V}$ and each $x \in X$ for which $\min\{\phi_b(x), \phi_{\bar{a}_1}(x)\} \le D + 2$ the relation $|\phi_{\bar{a}_1}(x) - \phi_b(x)| \le 4^{-1}\delta_1$ is true.

Property (Piii) and (3.19) imply that for each $b \in \mathcal{V}$,

$$|\phi_b(\bar{x}) - \phi_{\bar{a}_1}(\bar{x})| \le 4^{-1}\delta_1, \qquad \inf(\phi_b) \le \phi_b(\bar{x}) \le D. \tag{3.20}$$

Now we will show that (H1) is true with the open set $\mathcal{W} = \mathcal{V} \times \mathcal{U}$, $x = \bar{x}$, $\alpha = \phi_{\bar{a}_1}(\bar{x})$, and $\eta = 4^{-1}\delta_1$.

Assume that $b = (b_1, b_2) \in \mathcal{V} \times \mathcal{U}$. By (3.20) and (3.14)

$$\bar{x} \in S_{b_2}, \qquad \inf(f_b) = \inf\{\phi_{b_1}(z) : \ z \in S_{b_2}\} \le \phi_{b_1}(\bar{x}) < \infty. \tag{3.21}$$

Assume now that $z \in X$ and $f_b(z) \le \inf(f_b) + 4^{-1}\delta_1$. Then

$$z \in S_{b_2}, \qquad \phi_{b_1}(z) \le \inf\{\phi_{b_1}(y) : \ y \in S_{b_2}\} + 4^{-1}\delta_1. \tag{3.22}$$

By (3.21), (3.20), and (3.19),

$$\inf\{\phi_{b_1}(y) : \ y \in S_{b_2}\} \le \phi_{b_1}(\bar{x}) \le D, \qquad \inf\{\phi_{\bar{a}_1}(y) : \ y \in S_{b_2}\} \le \phi_{\bar{a}_1}(\bar{x}) \le D.$$

These inequalities imply that

$$\inf\{\phi_{b_1}(y) : \ y \in S_{b_2}\} = \inf\{\phi_{b_1}(y) : \ y \in S_{b_2} \text{ and } \phi_{b_1}(y) \le D + 1\}$$

and

$$\inf\{\phi_{\bar{a}_1}(y) : \ y \in S_{b_2}\} = \inf\{\phi_{\bar{a}_1}(y) : \ y \in S_{b_2} \text{ and } \phi_{\bar{a}_1}(y) \le D + 1\}.$$

It follows from these two relations and property (Piii) that

$$|\inf\{\phi_{b_1}(y) : \ y \in S_{b_2}\} - \inf\{\phi_{\bar{a}_1}(y) : \ y \in S_{b_2}\}| \le 4^{-1}\delta_1. \tag{3.23}$$

Relations (3.23), (3.22), (3.21), (3.19), and property (Piii) imply that

$$|\phi_{\bar{a}_1}(z) - \phi_{b_1}(z)| \le 4^{-1}\delta_1, \tag{3.24}$$

$$\phi_{\bar{a}_1}(z) \le \inf\{\phi_{\bar{a}_1}(y) : \ y \in S_{b_2}\} + \delta_1. \tag{3.25}$$

It follows from (3.25), (3.22), and property (Pii) that

$$\rho(z, \bar{x}) \le \gamma_0 \quad \text{and} \quad |\phi_{\bar{a}_1}(z) - \phi_{\bar{a}_1}(\bar{x})| \le \delta_1 + \delta(\gamma_0).$$

Together with (3.24), (3.9), and the definition of $\delta(\gamma_0)$ this implies that

$$|\phi_{b_1}(z) - \phi_{\bar{a}_1}(\bar{x})| \le 2\delta(\gamma_0) \le 2\gamma_0 < \gamma.$$

This completes the proof of the proposition. □

Remark 3.1. In the proof of Proposition 3.1 for any $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$ satisfying $\inf(f_a) < \infty$ and any $\epsilon > 0$ we constructed an open set $\mathcal{V}$ in $\mathcal{A}_1$ with the weak topology and an open set $\mathcal{U}$ in $\mathcal{A}_2$ with the weak topology which satisfy

$$\mathcal{V} \cap \{b \in \mathcal{A}_1 : \ d_1(b, a_1) < \epsilon\} \neq \emptyset \quad \text{and} \quad \mathcal{U} \cap \{b \in \mathcal{A}_2 : \ d_2(b, a_2) < \epsilon\} \neq \emptyset$$

and such that $\inf(f_b) < \infty$ for each $b = (b_1, b_2) \in \mathcal{V} \times \mathcal{U}$. This implies that there exists an open set $\mathcal{F}$ in $\mathcal{A}_1 \times \mathcal{A}_2$ with the weak topology such that $\inf(f_a) < \infty$ for all $a \in \mathcal{F}$ and $\mathcal{A}$ is the closure of $\mathcal{F}$ in the space $\mathcal{A}_1 \times \mathcal{A}_2$ with the strong topology.

4. Preliminary results for hypotheses (A2) and (H2). Assume that $A : \Omega \to 2^{R^n} \setminus \{\emptyset\}$, $U : \mathrm{graph}(A) \to 2^{R^N} \setminus \{\emptyset\}$, and that $\mathrm{graph}(U)$ is a closed subset of the space $\Omega \times R^n \times R^N$ with the product topology. Consider the spaces $X(A,U)$, $\mathcal{M}(A,U)$, and $C_l(B_1 \times B_2)$ introduced in section 1.

PROPOSITION 4.1. *Let $f \in \mathcal{M}(A,U)$, $(x,u) \in X(A,U)$, and $\{(x_i,u_i)\}_{i=1}^{\infty} \subset X(A,U)$, and let $\rho((x_i,u_i),(x,u)) \to 0$ as $i \to \infty$. Then*

$$\int_{\Omega} f(t,x(t),u(t))dt \le \liminf_{i\to\infty} \int_{\Omega} f(t,x_i(t),u_i(t))dt.$$

Proof. We may assume that there is a finite $\lim_{i\to\infty} \int_{\Omega} f(t,x_i(t),u_i(t))dt$. There is a subsequence $\{(x_{i_k},u_{i_k})\}_{k=1}^{\infty}$ such that

$$(x_{i_k}(t),u_{i_k}(t)) \to (x(t),u(t)) \quad \text{as } k \to \infty, \quad t \in \Omega \text{ a.e.}$$

(see [12, p. 68]). By property (ii) (see the definition of $\mathcal{M}(A,U)$) for almost every $t \in \Omega$,

$$\liminf_{k\to\infty} f(t,x_{i_k}(t),u_{i_k}(t)) \ge f(t,x(t),u(t)).$$

The proposition now follows from property (iii) (see the definition of $\mathcal{M}(A,U)$) and Fatou's lemma. □

The following proposition is an auxiliary result for the hypothesis (H2).

PROPOSITION 4.2. *Assume that $f \in \mathcal{M}(A,U)$, $\{(x_i,u_i)\}_{i=1}^{\infty} \subset X(A,U)$ is a Cauchy sequence, and the sequence $\{\int_{\Omega} f(t,x_i(t),u_i(t))dt\}_{i=1}^{\infty}$ is bounded. Then there is $(x_*,u_*) \in X(A,U)$ such that (x_i,u_i) converges to (x_*,u_*) as $i \to \infty$ in $X(A,U)$, and moreover, if $m = 1$, then $x_i(t) \to x_*(t)$ as $i \to \infty$ uniformly on $[T_1,T_2]$.*

Proof. To prove the proposition it is sufficient to show that there exists a subsequence $\{(x_{i_k},u_{i_k})\}_{k=1}^{\infty}$ and $(x_*,u_*) \in X(A,U)$ such that $(x_{i_k},u_{i_k}) \to (x_*,u_*)$ as $k \to \infty$ in $X(A,U)$ and if $m = 1$, then $x_{i_k}(t) \to x_*(t)$ as $k \to \infty$ uniformly on $[T_1,T_2]$. (In the case $m = 1$ this implies that each subsequence of $\{x_i\}_{i=1}^{\infty}$ has a subsequence which converges to x_* uniformly on $[T_1,T_2]$. This proves that $\{x_i\}_{i=1}^{\infty}$ converges to x_* uniformly on $[T_1,T_2]$.)

Since $\{(x_i,u_i)\}_{i=1}^{\infty}$ is a Cauchy sequence there is a strictly increasing sequence of natural numbers $\{i_k\}_{k=1}^{\infty}$ and a sequence of measurable sets $D_k \subset \Omega$, $k = 1,2,\dots$, such that for all $k = 1,2,\dots$,

$$\mathrm{mes}(D_k) \le 2^{-k}, \qquad |x_{i_{k+1}}(t) - x_{i_k}(t)| \le 2^{-k}, \tag{4.1}$$

$$|u_{i_{k+1}}(t) - u_{i_k}(t)| \le 2^{-k}, \qquad t \in \Omega \setminus D_k.$$

Set $C_k = \cup_{i=k}^{\infty} D_i$, $k = 1,2,\dots$. By (4.1) there exist measurable functions $u_* : \Omega \to R^N$ and $x_* : \Omega \to R^n$ such that

$$\lim_{k\to\infty} x_{i_k}(t) = x_*(t), \quad \lim_{k\to\infty} u_{i_k}(t) = u_*(t), \quad t \in \Omega \setminus \cap_{k=1}^{\infty} C_k. \tag{4.2}$$

Since the function $f(t,\cdot,\cdot)$ is lower semicontinuous for $t \in \Omega$ a.e. (see the definition of $\mathcal{M}(A,U)$, property (ii)) it follows from (4.2) that

$$f(t,x_*(t),u_*(t)) \le \liminf_{k\to\infty} f(t,x_{i_k}(t),u_{i_k}(t)), \qquad t \in \Omega \text{ a.e.} \tag{4.3}$$

Clearly the function $f(t, x_*(t), u_*(t))$, $t \in \Omega$, is measurable. By (4.3), Fatou's lemma, and property (iii), $\int_\Omega f(t, x_*(t), u_*(t))dt$ is finite. It follows from property (iii) and the boundedness of the sequence $\{\int_\Omega f(t, x_i(t), u_i(t))dt\}_{i=1}^\infty$ that the family of functions

$$\mathcal{E} = \{|H(t, x_*(t), u_*(t))|,\ t \in \Omega,\ |H(t, x_{i_k}(t), u_{i_k}(t))|,\ t \in \Omega,\ k = 1, 2, \dots\}$$

is uniformly integrable [11, p. 74]. Namely, for each $\epsilon > 0$ there exists $\delta > 0$ such that for each measurable set $e \subset \Omega$ satisfying $\text{mes}(e) \le \delta$ the following relations hold:

$$\int_e |H(t, x_*(t), u_*(t))|dt \le \epsilon, \quad \int_e |H(t, x_{i_k}(t), u_{i_k}(t))|dt \le \epsilon, \quad k = 1, 2, \dots.$$

It follows from this property, the continuity of H, (4.1), (4.2), and Egorov's theorem that for each measurable set $e \subset \Omega$,

$$\text{(4.4)} \qquad \int_e H(t, x_{i_k}(t), u_{i_k}(t))dt \to \int_e H(t, x_*(t), u_*(t))dt \quad \text{as } k \to \infty.$$

Now we consider the case with $m = 1$. Since the set $\mathcal{E}$ is uniformly integrable it follows from (1.4b), (4.2), and Ascoli's compactness theorem that a subsequence of the sequence $\{x_{i_k}\}_{k=1}^\infty$ converges to a continuous function $y : [T_1, T_2] \to R^n$ uniformly on $[T_1, T_2]$. By (4.2) we may assume that $x_*(t) = y(t)$, $t \in [T_1, T_2]$ a.e. Thus $x_* : \Omega \to R^n$ is continuous and some subsequence of $\{x_{i_k}\}_{k=1}^\infty$ converges to x_* uniformly on $[T_1, T_2]$. Together with (4.4) this implies that $(x_*, u_*) \in X(A, U)$. Since $\text{mes}(\cap_{k=1}^\infty C_k) = 0$ (see (4.1)) it follows from (4.2) that $(x_{i_k}, u_{i_k}) \to (x_*, u_*)$ as $k \to \infty$ in $X(A, U)$. Therefore the proposition is true in the case with $m = 1$.

We turn now to the case with $m > 1$. Since the set $\mathcal{E}$ is uniformly integrable it is easy to verify that

$$\text{(4.5)} \qquad H(\cdot, x_*(\cdot), u_*(\cdot)) \in L^1(\Omega), \quad H(\cdot, x_{i_k}(\cdot), u_{i_k}(\cdot)) \in L^1(\Omega), \quad k = 1, 2, \dots,$$

$$H(\cdot, x_{i_k}(\cdot), u_{i_k}(\cdot)) \to H(\cdot, x_*(\cdot), u_*(\cdot)) \quad \text{as } k \to \infty \quad \text{in } L^1(\Omega).$$

Note that $x_{i_k} - \theta^* \in (W_0^{1,1}(\Omega))^n$, $k = 1, 2, \dots$ (see (1.4)). By [28, Thm. 2.4.1] there is a constant $c > 0$ such that $||h||_{L^1(\Omega)} \le c||\nabla h||_{L^1(\Omega)}$ for all $h \in W_0^{1,1}(\Omega)$. Together with (4.5) and (4.2) this implies that $x_{i_k} \to x_*$ as $k \to \infty$ in $L^1(\Omega; R^n)$, $x_* \in (W^{1,1}(\Omega))^n$, $\nabla x_* = H(\cdot, x_*(\cdot), u_*(\cdot))$, and $(x_*, u_*) \in X(A, U)$. Analogously to the previous case we obtain that $(x_{i_k}, u_{i_k}) \to (x_*, u_*)$ as $k \to \infty$ in $X(A, U)$. Thus in the case $m > 1$ the proposition is proved. □

PROPOSITION 4.3. *Let $h \in C_l(B_1 \times B_2)$ and $\epsilon, D > 0$. Then there exists a neighborhood $\mathcal{V}$ of h in $C_l(B_1 \times B_2)$ with the weak topology such that for each $\xi \in \mathcal{V}$ and each $x \in B_1 \times B_2$ which satisfies* $\min\{\xi(x), h(x)\} \le D$ *the relation* $|\xi(x) - h(x)| \le \epsilon$ *holds.*

Proof. There is a $c_0 > 0$ such that $h(x) \ge -c_0$ for all $x \in B_1 \times B_2$. Choose a positive number $\epsilon_1 < 1$ for which

$$\epsilon_1 + \epsilon_1(1 - \epsilon_1)^{-1}(2 + D + c_0) < \epsilon$$

and define $\mathcal{V} = \{\xi \in C_l(B_1 \times B_2) :\ (\xi, h) \in E_{cw}(\epsilon_1)\}$ (see (1.9)). Assume that $\xi \in \mathcal{V}$, $x \in B_1 \times B_2$, and $\min\{\xi(x), h(x)\} \le D$. It follows from the definition of $\mathcal{V}$ and ϵ_1, (1.9), and Lemma 1.1 that $\xi(x), h(x)$ are finite and

$$|\xi(x)-h(x)|<\epsilon_1+\epsilon_1^2(1-\epsilon_1)^{-1}+\epsilon_1(1-\epsilon_1)^{-1}\min\{|\xi(x)|,|h(x)|\}$$

$$<\epsilon_1+\epsilon_1^2(1-\epsilon_1)^{-1}+\epsilon_1(1-\epsilon_1)^{-1}(D+c_0)<\epsilon.$$

The proposition is proved. □

Corollary 4.1. *Let $h\in C_l(B_1\times B_2)$ and $\epsilon>0$. Then there is a neighborhood $\mathcal{V}$ of h in $C_l(B_1\times B_2)$ with the weak topology such that for each $\xi\in\mathcal{V}$ the inequality $|\inf(\xi)-\inf(h)|\le\epsilon$ holds.*

Proof. We may assume that $\inf(h)$ is finite and $\epsilon<1$. By Proposition 4.3 there exists a neighborhood $\mathcal{V}$ of h in $C_l(B_1\times B_2)$ with the weak topology such that for each $\xi\in\mathcal{V}$ and each $x\in B_1\times B_2$ which satisfies $\min\{\xi(x),h(x)\}\le\inf(h)+2$ the relation $|\xi(x)-h(x)|\le 2^{-1}\epsilon$ holds.

Assume that $\xi\in\mathcal{V}$. It follows from the definition of $\mathcal{V}$ that for each $x\in B_1\times B_2$ satisfying $h(x)\le\inf(h)+2$ the relation $|\xi(x)-h(x)|\le 2^{-1}\epsilon$ is true. Choose $y\in X$ such that $h(y)\le\inf(h)+2^{-1}\epsilon$. Then

$$\inf(\xi)\le\xi(y)\le h(y)+2^{-1}\epsilon\le\inf(h)+\epsilon\le\inf(h)+1.$$

It follows from this inequality and the definition of $\mathcal{V}$ that for each $x\in B_1\times B_2$ satisfying $\xi(x)\le\inf(\xi)+1$ the relation $|\xi(x)-h(x)|\le 2^{-1}\epsilon$ holds. Choose $z\in X$ such that $\xi(z)\le\inf(\xi)+2^{-1}\epsilon$. Then

$$\inf(h)\le h(z)\le\xi(z)+2^{-1}\epsilon\le\inf(\xi)+\epsilon.$$

The corollary is proved. □

The following proposition is an auxiliary result for the assumption (A2).

Proposition 4.4. *Let $f\in\mathcal{M}(A,U)$ and $\epsilon\in(0,1)$, $D>0$. Then there exists a neighborhood $\mathcal{V}$ of f in $\mathcal{M}(A,U)$ with the weak topology such that for each $g\in\mathcal{V}$ and each $(x,u)\in X(A,U)$ satisfying*

$$\min\left\{\int_\Omega f(t,x(t),u(t))dt,\ \int_\Omega g(t,x(t),u(t))dt\right\}\le D \tag{4.6}$$

the following relation holds:

$$\left|\int_\Omega f(t,x(t),u(t))dt-\int_\Omega g(t,x(t),u(t))dt\right|\le\epsilon. \tag{4.7}$$

Proof. There is an integrable function $\phi_0(t)\ge 0$, $t\in\Omega$, such that

$$f(t,x,u)\ge-\phi_0(t)\quad\text{for all } (t,x,u)\in\ \text{graph}(U). \tag{4.8}$$

Choose a positive number ϵ_1 for which

$$\epsilon_1\left(2\text{mes}(\Omega)+2+\int_\Omega\phi_0(t)dt+D\right)<\epsilon \tag{4.9}$$

and a positive number ϵ_0 that satisfies

$$\epsilon_0+\epsilon_0(1-\epsilon_0)^{-1}<4^{-1}\epsilon_1. \tag{4.10}$$

Define

$$\mathcal{V} = \{g \in \mathcal{M}(A,U) : (g,f) \in E_{\mathcal{M}w}(\epsilon_0)\} \quad \text{(see (1.7))}. \tag{4.11}$$

Assume that $g \in \mathcal{V}$, $(x,u) \in X(A,U)$, and (4.6) is valid. By (4.11) and (1.7) there is a nonnegative function $\phi \in L^1(\Omega)$ such that $\int_\Omega \phi(t)dt \le 1$, and for almost every $t \in \Omega$ the inequality

$$|f(t,y,v) - g(t,y,v)| < \epsilon_0 + \epsilon_0\phi(t) + \epsilon_0 \max\{|f(t,y,v)|, |g(t,y,v)|\}$$

is true for each $y \in A(t)$ and each $v \in U(t,y)$. It follows from this inequality, Lemma 1.1, and (4.10) that for almost every $t \in \Omega$ the relation

$$\begin{aligned} |f(t,y,v) - g(t,y,v)| &< \epsilon_0 + \epsilon_0^2(1-\epsilon_0)^{-1} + \phi(t)(\epsilon_0^2(1-\epsilon_0)^{-1} + \epsilon_0) \\ &\quad + \epsilon_0(1-\epsilon_0)^{-1}\min\{|f(t,y,v)|, |g(t,y,v)|\} \\ &< 4^{-1}\epsilon_1 + 4^{-1}\epsilon_1\phi(t) + 4^{-1}\epsilon_1 \min\{|f(t,y,v)|, |g(t,y,v)|\} \end{aligned} \tag{4.12}$$

is valid for each $y \in A(t)$ and each $v \in U(t,y)$. Relations (4.12) and (4.8) imply that for almost every $t \in \Omega$ the inequality

$$\begin{aligned} g(t,y,v) &\ge f(t,y,v) - 4^{-1}\epsilon_1 - 4^{-1}\epsilon_1\phi(t) - 4^{-1}\epsilon_1|f(t,y,v)| \\ &\ge -4^{-1}\epsilon_1\phi(t) - 2\phi_0(t) - 4^{-1}\epsilon_1 \end{aligned} \tag{4.13}$$

holds for each $y \in A(t)$ and each $v \in U(t,y)$. Set

$$\lambda(t) = \min\{f(t,x(t),u(t)), g(t,x(t),u(t))\}, \qquad t \in \Omega. \tag{4.14}$$

It follows from (4.12), (4.8), (4.13), and (4.14) that for almost every $t \in \Omega$,

$$\begin{aligned} &|f(t,x(t),u(t)) - g(t,x(t),u(t))| < 4^{-1}\epsilon_1 + 4^{-1}\epsilon_1\phi(t) \\ &+ 4^{-1}\epsilon_1 \min\{f(t,x(t),u(t)) + 2\phi_0(t),\ g(t,x(t),u(t)) + \phi(t) + 4\phi_0(t) + 2\} \\ &\le 4^{-1}\epsilon_1 + 4^{-1}\epsilon_1\phi(t) + 4^{-1}\epsilon_1(\phi(t) + 4\phi_0(t) + 2) + 4^{-1}\epsilon_1\lambda(t). \end{aligned}$$

By this relation, (4.6), and (4.9),

$$\begin{aligned} \int_\Omega |f(t,x(t),u(t)) - g(t,x(t),u(t))|dt &\le 4^{-1}\epsilon_1 \mathrm{mes}(\Omega) + 4^{-1}\epsilon_1 \int_\Omega \phi(t)dt \\ &+ 4^{-1}\epsilon_1 \int_\Omega \phi(t)dt + \epsilon_1 \int_\Omega \phi_0(t)dt + \epsilon_1\mathrm{mes}(\Omega) + 4^{-1}\epsilon_1 D < \epsilon. \end{aligned}$$

This completes the proof of the proposition. □

Analogously to the proof of Corollary 4.1 we can show that Proposition 4.4 implies the following corollary.

Corollary 4.2. *Let $f \in \mathcal{M}(A,U)$ and $\epsilon > 0$. Then there exists a neighborhood $\mathcal{V}$ of f in $\mathcal{M}(A,U)$ with the weak topology such that for all $g \in \mathcal{V}$,*

$$\left|\inf\left\{\int_\Omega f(t,x(t),u(t))dt :\ (x,u) \in X(A,U)\right\} - \inf\left\{\int_\Omega g(t,x(t),u(t))dt :\right.\right.$$

$$\left.\left.(x,u) \in X(A,U)\right\}\right| < \epsilon.$$

Proposition 4.5. *Let $m = 1$, $f \in \mathcal{M}(A,U)$, $h \in C_l(B_1 \times B_2)$, and $\epsilon \in (0,1)$, $D > 0$. Then there exist a neighborhood $\mathcal{U}$ of f in $\mathcal{M}(A,U)$ with the weak topology and a neighborhood $\mathcal{V}$ of h in $C_l(B_1 \times B_2)$ with the weak topology such that for each $(\xi, g) \in \mathcal{V} \times \mathcal{U}$ and each $(x,u) \in X(A,U)$ which satisfies*

$$\min\{I^{(f,h)}(x,u),\ I^{(g,\xi)}(x,u)\} \le D \tag{4.15}$$

the following relations are valid:

$$|h(x(T_1),x(T_2)) - \xi(x(T_1),x(T_2))| \le \epsilon, \tag{4.16}$$

$$\left|\int_{T_1}^{T_2}[f(t,x(t),u(t)) - g(t,x(t),u(t))]dt\right| \le \epsilon. \tag{4.17}$$

Proof. We may assume that $\inf(h)$ and

$$\inf\left\{\int_{T_1}^{T_2} f(t,x(t),u(t))dt :\ (x,u) \in X(A,U)\right\}$$

are finite. Choose a number

$$c_0 > 4 + |\inf(h)| + \left|\inf\left\{\int_{T_1}^{T_2} f(t,x(t),u(t))dt :\ (x,u) \in X(A,U)\right\}\right|.$$

By Corollaries 4.1 and 4.2 there exists a neighborhood $\mathcal{V}_1$ of $h \in C_l(B_1 \times B_2)$ with the weak topology such that

$$|\inf(\xi)| < c_0 \quad \text{for all } \xi \in \mathcal{V}_1 \tag{4.18}$$

and a neighborhood $\mathcal{U}_1$ of f in $\mathcal{M}(A,U)$ with the weak topology such that

$$\left|\inf\left\{\int_{T_1}^{T_2} g(t,x(t),u(t))dt :\ (x,u) \in X(A,U)\right\}\right| < c_0 \quad \text{for all } g \in \mathcal{U}_1. \tag{4.19}$$

By Proposition 4.3 there exists a neighborhood $\mathcal{V}$ of h in $C_l(B_1 \times B_2)$ with the weak topology such that $\mathcal{V} \subset \mathcal{V}_1$ and that for each $\xi \in \mathcal{V}$ and each $z \in B_1 \times B_2$ which satisfies $\min\{\xi(z), h(z)\} \le D + c_0 + 2$ the relation $|\xi(z) - h(z)| \le \epsilon$ holds. By Proposition 4.4 there exists a neighborhood $\mathcal{U}$ of f in $\mathcal{M}(A,U)$ with the weak topology such that $\mathcal{U} \subset \mathcal{U}_1$ and that for each $g \in \mathcal{U}$ and each $(x,u) \in X(A,U)$ satisfying

$$\min\left\{\int_{T_1}^{T_2} f(t,x(t),u(t))dt,\ \int_{T_1}^{T_2} g(t,x(t),u(t))dt\right\} \le D + c_0 + 2$$

the inequality (4.17) holds.

Now assume that $(\xi, g) \in \mathcal{V} \times \mathcal{U}$ and $(x, u) \in X(A, U)$ satisfies (4.15). It follows from (4.15), (4.18), and (4.19) that

$$\min\{\xi(x(T_1), x(T_2)), h(x(T_1), x(T_2))\} - c_0 \le \min\{I^{(f,h)}(x,u),\ I^{(g,\xi)}(x,u)\} \le D$$

and

$$\min\left\{\int_{T_1}^{T_2} f(t, x(t), u(t))dt,\ \int_{T_1}^{T_2} g(t, x(t), u(t))dt\right\} - c_0$$

$$\le \min\{I^{(f,h)}(x,u),\ I^{(g,\xi)}(x,u)\} \le D.$$

By these inequalities and the definition of $\mathcal{U}$ and $\mathcal{V}$, the inequalities (4.16) and (4.17) are valid. The proposition is proved. □

5. Preliminary lemma for hypothesis (A3). Fix a number $d_0 \in (0,1)$. There is a C^∞-function $\phi_0 : R^1 \to [0,1]$ such that $\phi_0(t) = 1$ if $|t| \le d_0$, $1 > \phi_0(t) > 0$ if $d_0 < |t| < 1$, and $\phi_0(t) = 0$ if $|t| \ge 1$. Define a C^∞-function $\bar{\phi} : R^1 \to R^1$ by $\bar{\phi}(x) = \int_0^x \phi_0(t)dt$, $x \in R^1$. Clearly $\bar{\phi}$ is monotone increasing, $\bar{\phi}(x) = x$ if $|x| \le d_0$, and

$$\bar{\phi}(x) = \bar{\phi}(1) \quad \text{if } x \ge 1, \quad \bar{\phi}(x) = \bar{\phi}(-1) \quad \text{if } x \le -1, \tag{5.1}$$

$$d_0 = \bar{\phi}(d_0) \le \bar{\phi}(x) \le \bar{\phi}(1) \le 1 \qquad \text{for all } x \in (d_0, 1). \tag{5.2}$$

Now we define a set $\mathcal{L} \subset C_l(B_1 \times B_2)$. In the case $m = 1$ we set $\mathcal{L} = C_l(B_1 \times B_2)$ and in the case $m > 1$ denote by $\mathcal{L}$ a singleton $\{0\}$ where 0 is a function in $C_l(B_1 \times B_2)$ which is identically zero. In the case $m > 1$ for each $(f, \xi) \in \mathcal{M}(A, U) \times \mathcal{L}$ and each $(x, u) \in X(A, U)$ we set

$$I^{(f,\xi)}(x,u) = I^{(f)}(x,u) \tag{5.3}$$

(see (1.10) and (1.11)). For each measurable set $E \subset R^m$, each measurable set $E_0 \subset E$, and each $h \in L^1(E)$ we set

$$||h||_{L^1(E_0)} = \int_{E_0} |h(t)|dt. \tag{5.4}$$

Fix an integer $k \ge 1$. It is easy to verify that all partial derivatives of the functions $(x, y) \to \bar{\phi}(|x-y|^2)$, $(x, y) \in R^q \times R^q$ with $q = n, N$ up to the order k are bounded (by some $\bar{d} > 0$).

For each $\gamma \in (0,1)$ choose $\epsilon_0(\gamma) \in (0, \gamma)$ such that

$$E_X(8\epsilon_0(\gamma)) \subset \{((x_1, u_1), (x_2, u_2)) \in X(A,U) \times X(A,U) :\ \rho((x_1, u_1), (x_2, u_2)) \le \gamma\} \tag{5.5}$$

(see (1.5)) and

$$\epsilon_0(\gamma) < 4^{-1}\gamma(\bar{d} + 2)^{-1} \tag{5.6}$$

and choose

$$\epsilon_1(\gamma) \in (0, d_0\epsilon_0(\gamma)), \tag{5.7}$$

$$\delta(\gamma) \in (0, 16^{-1}\epsilon_1(\gamma)^4). \tag{5.8}$$

LEMMA 5.1. *Let $\gamma \in (0,1)$, $f \in \mathcal{M}(A,U)$, $\xi \in \mathcal{L}$, and let $Y \subset X(A,U)$, $(\bar{x},\bar{u}) \in Y$,*

$$I^{(f,\xi)}(\bar{x},\bar{u}) \le \inf\{I^{(f,\xi)}(x,u) : (x,u) \in Y\} + \delta(\gamma) < \infty. \tag{5.9}$$

Then there is a $g : R^m \times R^n \times R^N \to R^1$ in $C^k(R^{m+n+N})$ which satisfies

$$0 \le g(t,x,u) \le \gamma \quad \textit{for all } (t,x,u) \in R^m \times R^n \times R^N, \tag{5.10}$$

$$||g(t,\cdot,\cdot)||_{C^k(R^n \times R^N)} \le \gamma \quad \textit{for all } t \in R^m$$

such that for a function $\bar{f} \in \mathcal{M}(A,U)$ defined by

$$\bar{f}(t,x,u) = f(t,x,u) + g(t,x,u),\ (t,x,u) \in \textit{graph}(U), \tag{5.11}$$

the following properties hold:

$$I^{(\bar{f},\xi)}(\bar{x},\bar{u}) \le I^{(f,\xi)}(\bar{x},\bar{u}) + \delta(\gamma); \tag{5.12}$$

for each $(y,v) \in Y$ satisfying

$$I^{(\bar{f},\xi)}(y,v) \le \inf\{I^{(\bar{f},\xi)}(z,w) : (z,w) \in Y\} + 2\delta(\gamma) \tag{5.13}$$

the relation $\rho((y,v),(\bar{x},\bar{u})) \le \gamma$ is valid.

Moreover the function g is the sum of two functions, one of them depends only on (t,x) while the other depend only on (t,u).

Proof. Choose a positive number ϵ_2 for which

$$e_2 < (\text{mes}(\Omega) + 1)^{-1} 8^{-1} \delta(\gamma) d_0 (\bar{d} + 1)^{-1}. \tag{5.14}$$

There is a measurable set $E_0 \subset \Omega$ such that

$$\text{mes}(\Omega \setminus E_0) < 2^{-1}\epsilon_2 \tag{5.15}$$

and the functions $\bar{x}$ and $\bar{u}$ are bounded on E_0. There exist sequences of functions $\{\bar{x}_i\}_{i=1}^{\infty} \in C^{\infty}(R^m; R^n)$ and $\{\bar{u}_i\}_{i=1}^{\infty} \subset C^{\infty}(R^m; R^N)$ such that

$$||\bar{u}_i - \bar{u}||_{L^1(E_0)},\ ||\bar{x}_i - \bar{x}||_{L^1(E_0)} \to 0 \quad \text{as } i \to \infty \tag{5.16}$$

[17, p. 13]. We may assume without loss of generality that $\bar{u}_i(t) \to \bar{u}(t)$, $\bar{x}_i(t) \to \bar{x}(t)$ as $i \to \infty$, $t \in E_0$ a.e. By Egorov's theorem there is a measurable set $E_1 \subset E_0$ such that

$$\text{mes}(E_0 \setminus E_1) < 2^{-1}\epsilon_2 \tag{5.17}$$

and

$$\bar{u}_i(t) \to \bar{u}(t) \quad \text{and} \quad \bar{x}_i(t) \to \bar{x}(t) \quad \text{uniformly in } E_1 \text{ as } i \to \infty. \tag{5.18}$$

There is an integer $s \geq 1$ such that

$$\max\{|\bar{u}_s(t) - \bar{u}(t)|,\ |\bar{x}_s(t) - \bar{x}(t)|\} \leq 4^{-1}\epsilon_2(\text{mes}(\Omega) + 1)^{-1}, \qquad t \in E_1. \tag{5.19}$$

Define a function $g : R^m \times R^n \times R^N \to R^1$ by

(5.20)

$$g(t, x, u) = \epsilon_0(\gamma)\bar{\phi}(|x - \bar{x}_s(t)|^2) + \epsilon_0(\gamma)\bar{\phi}(|u - \bar{u}_s(t)|^2), \qquad (t, x, u) \in R^m \times R^n \times R^N.$$

Clearly $g \in C^\infty(R^m \times R^n \times R^N)$. Define

$$\bar{f}(t, x, u) = f(t, x, u) + g(t, x, u), \qquad (t, x, u) \in \text{ graph}(U). \tag{5.21}$$

Evidently $\bar{f} \in \mathcal{M}(A, U)$. It follows from (5.20), the definition of $\bar{d}$, (5.1), (5.2), and (5.6) that (5.10) is true. We will show that (5.12) is true. By (5.21), (5.20), (5.19), (5.1), and (5.2),

$$I^{(\bar{f},\xi)}(\bar{x}, \bar{u}) = I^{(f,\xi)}(\bar{x}, \bar{u}) + \epsilon_0(\gamma)\int_\Omega \bar{\phi}(|\bar{x}(t) - \bar{x}_s(t)|^2)dt$$

$$+ \epsilon_0(\gamma)\int_\Omega \bar{\phi}(|\bar{u}(t) - \bar{u}_s(t)|^2)dt = I^{(f,\xi)}(\bar{x}, \bar{u}) + \epsilon_0(\gamma)\int_{E_1} \bar{\phi}(|\bar{x}(t) - \bar{x}_s(t)|^2)\, dt$$

$$+ \epsilon_0(\gamma)\int_{\Omega \setminus E_1} \bar{\phi}(|\bar{x}(t) - \bar{x}_s(t)|^2)dt + \epsilon_0(\gamma)\int_{E_1} \bar{\phi}(|\bar{u}(t) - \bar{u}_s(t)|^2)dt$$

$$+ \epsilon_0(\gamma)\int_{\Omega \setminus E_1} \bar{\phi}(|\bar{u}(t) - \bar{u}_s(t)|^2)dt \leq I^{(f,\xi)}(\bar{x}, \bar{u})$$

$$+ 2(\text{mes}(\Omega))\epsilon_0(\gamma)\bar{\phi}((4^{-1}\epsilon_2)^2) + 2\epsilon_0(\gamma)\text{ mes }(\Omega \setminus E_1).$$

It follows from this relation, (5.14), (5.1), (5.2), (5.15), and (5.17) that

$$I^{(\bar{f},\xi)}(\bar{x}, \bar{u}) \leq I^{(f,\xi)}(\bar{x}, \bar{u}) + 2\text{mes}(\Omega)\epsilon_0(\gamma)(4^{-1}\epsilon_2)^2 + 2\epsilon_0(\gamma)\epsilon_2$$

$$\leq I^{(f,\xi)}(\bar{x}, \bar{u}) + 4\epsilon_0(\gamma)\epsilon_2 \leq I^{(f,\xi)}(\bar{x}, \bar{u}) + \delta(\gamma).$$

Thus (5.12) is valid. Now assume that $(y, v) \in Y$ satisfies (5.13). It follows from (5.13), (5.21), (5.20), and (5.9) that

$$I^{(f,\xi)}(y, v) + \epsilon_0(\gamma)\int_\Omega \bar{\phi}(|\bar{x}_s(t) - y(t)|^2)dt + \epsilon_0(\gamma)\int_\Omega \bar{\phi}(|v(t) - \bar{u}_s(t)|^2)dt$$

$$= I^{(\bar{f},\xi)}(y, v) \leq 2\delta(\gamma) + I^{(\bar{f},\xi)}(\bar{x}, \bar{u}) \leq 3\delta(\gamma) + I^{(f,\xi)}(\bar{x}, \bar{u}) \leq I^{(f,\xi)}(y, v) + 4\delta(\gamma).$$

This implies that

$$\int_\Omega \bar\phi(|\bar x_s(t)-y(t)|^2)dt+\int_\Omega \bar\phi(|\bar u_s(t)-v(t)|^2)dt \le 4\delta(\gamma)(\epsilon_0(\gamma))^{-1}. \tag{5.22}$$

Set

$$E_2=\{t\in\Omega:\ |y(t)-\bar x_s(t)|\ge 2^{-1}\epsilon_1(\gamma)\},\ E_3=\{t\in\Omega:\ |v(t)-\bar u_s(t)|\ge 2^{-1}\epsilon_1(\gamma)\}. \tag{5.23}$$

Then by (5.23), (5.22), (5.6), (5.7), (5.1), (5.2), and (5.8)

$$\operatorname{mes}(E_2)+\operatorname{mes}(E_3)\le 4\epsilon_1(\gamma)^{-2}\left[\int_{E_2}\bar\phi(|\bar x_s(t)-y(t)|^2)dt+\int_{E_3}\bar\phi(|\bar u_s(t)-v(t)|^2)dt\right] \tag{5.24}$$

$$\le 16\epsilon_1(\gamma)^{-2}\delta(\gamma)(\epsilon_0(\gamma))^{-1}<\epsilon_1(\gamma).$$

It follows from (5.24), (5.23), (5.19), (5.14), (5.15), and (5.17) that

$$\operatorname{mes}\{t\in\Omega:\ |y(t)-\bar x(t)|\ge\epsilon_1(\gamma)\}\le\ \operatorname{mes}(\Omega\setminus E_1)$$

$$+\operatorname{mes}\{t\in\Omega:\ |y(t)-\bar x_s(t)|\ge 2^{-1}\epsilon_1(\gamma)\}\le\epsilon_2+\epsilon_1(\gamma)\le 2\epsilon_1(\gamma)$$

and

$$\operatorname{mes}\{t\in\Omega:\ |v(t)-\bar u(t)|\ge\epsilon_1(\gamma)\}\le\operatorname{mes}(\Omega\setminus E_1)$$

$$+\operatorname{mes}\{t\in\Omega:\ |v(t)-\bar u_s(t)|\ge 2^{-1}\epsilon_1(\gamma)\}\le\epsilon_2+\epsilon_1(\gamma)\le 2\epsilon_1(\gamma).$$

These relations and (5.5) imply that $((y,v),(\bar x,\bar u))\in E_X(4\epsilon_1(\gamma))$, $\rho((y,v),(\bar x,\bar u))\le\gamma$. This completes the proof of the lemma. □

6. An auxiliary result. Let $p\ge 1$ be an integer and let $e_1=(1,0,\dots 0)$,..., $e_p=(0,\dots,0,1)$ be the standard basis in R^p. For each set $E\subset R^p$ denote by $\operatorname{conv}(E)$ its convex hull.

PROPOSITION 6.1. *Let a finite set* $E=\{h_{ij}:\ i=1,2,\dots,p,\ j=1,2\}\subset R^p$ *satisfy*

$$|h_{i1}-e_i|,\ |h_{i2}+e_i|\le(2p)^{-1},\qquad i=1,\dots,p.$$

Then the relation $0\in\operatorname{conv}(E)$ *holds.*

Proof. Let us assume the converse. Then $0\notin\operatorname{conv}(E)$ and there is $\xi=(\xi_1,\dots,\xi_p)\in R^p\setminus\{0\}$ such that $\inf\{\langle g,\xi\rangle:\ g\in\operatorname{conv}(E)\}>0$. We may assume that $|\xi_1|\ge|\xi_i|$, $i=1,\dots,p$. There are two cases: $\xi_1>0$; $\xi_1<0$. Consider the case with $\xi_1>0$. Then $0<\langle\xi,h_{12}\rangle=\langle\xi,-e_1\rangle+\langle\xi,h_{12}+e_1\rangle\le-\xi_1+(2p)^{-1}p|\xi_1|<0$, a contradiction. Analogously we obtain a contradiction in the second case. The proposition is proved. □

7. Auxiliary lemma for hypothesis (A4). Assume that $A : \Omega \to 2^{R^n} \setminus \{\emptyset\}$ and graph(A) is a closed subset of the space $\Omega \times R^n$ with the product topology. Let $e_1 = (1, 0, \dots, 0)$, $e_2 = (0, 1, \dots, 0), \dots, e_N = (0, 0, \dots, 1)$ be a standard basis in R^N. Now we define a set $\mathcal{L} \subset C_l(B_1 \times B_2)$. In the case $m = 1$ we set $\mathcal{L} = C_l(B_1 \times B_2)$ and in the case $m > 1$ we denote by $\mathcal{L}$ a singleton $\{0\}$ where 0 is a function in $C_l(B_1 \times B_2)$, which is identically zero. In the case $m > 1$ for each $(f, \xi) \in \mathcal{M}(A, \tilde{U}_A) \times \mathcal{L}$ and each $(x, u) \in X(A, \tilde{U}_A)$ we set

$$I^{(f,\xi)}(x, u) = I^{(f)}(x, u)$$

(see (1.10), (1.11), and (1.12)).

LEMMA 7.1. *Let $f \in \mathcal{M}(A, \tilde{U}_A)$, $\xi \in \mathcal{L}$, $U \in \mathcal{P}_A$,*

$$\{(x, u) \in X(A, U) : \ I^{(f,\xi)}(x, u) < \infty\} \neq \emptyset, \tag{7.1}$$

and let $\epsilon, \delta > 0$. Then there are $U_* \in \mathcal{P}_A$, $(\bar{x}, \bar{u}) \in X(A, U_*)$, and an open set $\mathcal{W}$ in $\mathcal{P}_A$ such that

$$(U_*, U) \in E_{\mathcal{P}_A}(\epsilon), \qquad \mathcal{W} \cap \{V \in \mathcal{P}_A : \ (U, V) \in E_{\mathcal{P}_A}(\epsilon)\} \neq \emptyset, \tag{7.2}$$

$$I^{(f,\xi)}(\bar{x}, \bar{u}) \leq \inf\{I^{(f,\xi)}(x, u) : \ (x, u) \in X(A, U_*)\} + \delta < \infty, \tag{7.3}$$

and for all $V \in \mathcal{W}$,

$$(\bar{x}, \bar{u}) \in X(A, V) \subset X(A, U_*). \tag{7.4}$$

Proof. For each $r \in [0, 1]$ define $U_r \in \mathcal{P}_A$ by

$$U_r(t, x) = \{u \in R^N : \ d_H(u, U(t, x)) \leq r\}, \qquad (t, x) \in \text{graph}(A), \tag{7.5}$$

and define

$$\mu(r) = \inf\{I^{(f,\xi)}(x, u) : \ (x, u) \in X(A, U_r)\}. \tag{7.6}$$

Clearly $\mu(r)$ is finite for all $r \in [0, 1]$ and the function μ is monotone decreasing. There is an $r_0 \in (0, 8^{-1}\epsilon)$ such that μ is continuous at r_0. Choose $r_1 \in (0, r_0)$ such that

$$|\mu(r_1) - \mu(r_0)| < 16^{-1}\delta. \tag{7.7}$$

There is

$$(\bar{x}, \bar{u}) \in X(A, U_{r_1}) \tag{7.8}$$

such that

$$I^{(f,\xi)}(\bar{x}, \bar{u}) \leq \mu(r_1) + 16^{-1}\delta. \tag{7.9}$$

Relations (7.7) and (7.9) imply that

$$I^{(f,\xi)}(\bar{x}, \bar{u}) \leq \mu(r_0) + 8^{-1}\delta. \tag{7.10}$$

Set

$$r_2 = 2^{-1}(r_0 + r_1). \tag{7.11}$$

Clearly

$$(U_{r_i}, U) \in E_{\mathcal{P}_A}(\epsilon), \qquad i = 0, 1, 2. \tag{7.12}$$

Choose a positive number γ for which

$$\gamma < \min\{4^{-1}\delta,\ (16N)^{-1}(r_0 - r_1)\} \tag{7.13}$$

and define

$$\mathcal{W} = \{V \in \mathcal{P}_A :\ (U_{r_2}, V) \in E_{\mathcal{P}_A}(\gamma)\}, \qquad U_* = U_{r_0}. \tag{7.14}$$

It follows from (7.12), (7.14), (7.10), and (7.6) that (7.2) and (7.3) are true.

Assume that $V \in \mathcal{W}$. Then by (7.14), (7.13), and (7.11), for each $(t, x) \in \operatorname{graph}(A)$,

$$V(t, x) \subset \{z \in R^N :\ d_H(z, U_{r_2}(t, x)) \le \gamma\}$$

$$\subset \{z \in R^N :\ d_H(z, U(t, x)) \le r_0\} = U_{r_0}(t, x).$$

Therefore

$$X(A, V) \subset X(A, U_{r_0}). \tag{7.15}$$

We will show that $(\bar{x}, \bar{u}) \in X(A, V)$. It is sufficient to show that

$$\bar{u}(t) \in V(t, \bar{x}(t)) \quad \text{for almost every } t \in \Omega. \tag{7.16}$$

By (7.8) for almost every $t \in \Omega$,

$$\bar{u}(t) \in U_{r_1}(t, \bar{x}(t)). \tag{7.17}$$

Assume that $t \in \Omega$ and (7.17) is true. By (7.17), (7.11), (7.5), and (7.14) for $i = 1, \dots, N$,

$$\bar{u}(t) + 2^{-1}(r_0 - r_1)e_i, \qquad \bar{u}(t) - 2^{-1}(r_0 - r_1)e_i \in U_{r_2}(t, \bar{x}(t)),$$

and there are $z_{i1}, z_{i2} \in R^N$ such that

(7.18)
$$\bar{u}(t) + z_{i1},\ \bar{u}(t) + z_{i2} \in V(t, \bar{x}(t)),\ |z_{i1} - 2^{-1}(r_0 - r_1)e_i|,\ |z_{i2} + 2^{-1}(r_0 - r_1)e_i| \le \gamma.$$

Since the set $V(t, \bar{x}(t))$ is convex it follows from (7.18), (7.13), and Proposition 6.1 that

$$0 \in \operatorname{conv}\{z_{ij} :\ i = 1, \dots, N,\ j = 1, 2\}, \qquad \bar{u}(t) \in V(t, \bar{x}(t)).$$

This implies that $(\bar{x}, \bar{u}) \in X(A, V)$. The lemma is proved. □

8. Proof of Theorem 1.1 and its extensions.

Proof of Theorem 1.1. By Propositions 4.1 and 4.2 (A1) holds and J_a is lower semicontinuous for all $a \in \mathcal{A}_1 \times \mathcal{A}_2$. By Theorem 2.1 we need to verify that (H1) and (H2) are valid. (H2) follows from Proposition 4.2. Therefore it is sufficient to show that (H1) holds. By Proposition 3.1 it is sufficient to show that (A2), (A3), and (A4) are valid. Hypothesis (A2) follows from Propositions 4.4 and 4.5. By Lemma 5.1, (A3) holds. Hypothesis (A4) follows from Lemma 7.1. This completes the proof of the theorem. □

As we mentioned in section 1 we proved Theorem 1.1 in such a manner that it is applicable for all the spaces of integrands introduced there. All the spaces of integrands are subspaces of $\mathcal{M}(A, U)$. Since (H2), (A1), (A2), and (A4) hold for the class of optimal control problems with the space of integrands $\mathcal{M}(A, U)$, they are also valid for all its subclasses considered here. On the other hand (A3) follows from Lemma 5.1, which establishes that $f + g$ and f belong to the same subspaces of integrands. This implies that (A3) holds for all classes of optimal control problems introduced in section 1.

As seen from the proof of Lemma 5.1 the perturbation g of the integrand f is chosen as the sum of two functions, one of them depends only on (t, x) while the other depend only on (t, u). Therefore Theorem 1.1 can be easily extended to subclasses of the classes of optimal control problems introduced in section 1 in which integrands are sums of two finite-valued functions, one of them, depending only on (t, x), is defined on graph(A), while the other, depending only on (t, u), is defined on $\Omega \times R^N$.

Finally in this section we present the extension of the generic existence and uniqueness result established in [27] for the space of lower semicontinuous integrands f : graph$(U) \to R^1$. Our generalization holds for all the spaces of integrands defined in section 1, and it is obtained as a realization of the generic variational principle established in section 2.

Assume that $A : \Omega \to 2^{R^n} \setminus \{\emptyset\}$, U : graph$(A) \to 2^{R^N} \setminus \{\emptyset\}$, and graph$(U)$ is a closed subset of $\Omega \times R^n \times R^N$ with the product topology. We consider the metric space $X(A, U)$ with the metric ρ (see (1.5)).

Now we define $\mathcal{A}_1$ as follows:

$$\mathcal{A}_1 = \mathcal{A}_{11} \times \mathcal{A}_{12} \quad \text{if } m = 1 \quad \text{and} \quad \mathcal{A}_1 = \mathcal{A}_{11} \quad \text{if } m > 1,$$

where $\mathcal{A}_{12}$ is either $C_l(B_1 \times B_2)$ or $C(B_1 \times B_2)$ or a singleton $\{\xi\} \subset C_l(B_1 \times B_2)$, and $\mathcal{A}_{11}$ is one of the following spaces:

$$\mathcal{M}(A, U);\ \mathcal{M}^l(A, U);\ \mathcal{M}^c(A, U);$$

$$\mathcal{M}_k(A, \tilde{U}_A);\ \mathcal{M}_k^l(A, \tilde{U}_A);\ \mathcal{M}_k^c(A, \tilde{U}_A) \text{ (here } k \geq 1 \text{ is an integer, } U = \tilde{U}_A,$$

$$\text{and graph}(A) \text{ is a closed subset of the space } \Omega \times R^n \text{ with the product topology);}$$

$$\mathcal{M}_k^*(\tilde{A}, \tilde{U});\ \mathcal{M}_k^{*l}(\tilde{A}, \tilde{U});\ \mathcal{M}_k^{*c}(\tilde{A}, \tilde{U}) \text{ (here } k \geq 1 \text{ is an integer and } A = \tilde{A},\ U = \tilde{U}).$$

Denote by $\mathcal{A}$ the closure of the set $\{a \in \mathcal{A}_1 : \inf(I^{(a)}) < \infty\}$ in the space $\mathcal{A}_1$ with the strong topology. We assume that $\mathcal{A}$ is nonempty. The following result is proved analogously to Theorem 1.1.

THEOREM 8.1. *The minimization problem for $I^{(a)}$ on $(X(A, U), \rho)$ is generically strongly well posed with respect to $\mathcal{A}$.*

9. Generic existence and uniqueness of solutions for variational problems without convexity assumptions. We use the notation and definitions introduced in section 1. Assume that $n = N$, $H(t, x, u) = u$, $(t, x, u) \in \Omega \times R^n \times R^n$, and B_1 and B_2 are singletons. Let $A : \Omega \to 2^{R^n} \setminus \{\emptyset\}$, $U : \mathrm{graph}(A) \to 2^{R^n} \setminus \{\emptyset\}$ and let $\mathrm{graph}(U)$ be a closed subset of the space $\Omega \times R^n \times R^n$ with the product topology. If $(x, u) \in X(A, U)$, then $u = \nabla x$ and (x, u) is identified with $x \in (W^{1,1}(\Omega))^n$. In what follows we omit the notation u in describing the elements of $X(A, U)$. For the set $X(A, U)$ we consider the metric ρ introduced in section 1 (see (1.5)) and the metric ρ_s defined by

$$\rho_s(x, y) = ||x - y||_{W^{1,1}(\Omega)} \quad \text{for all } x, y \in X(A, U).$$

Clearly $(X(A, U), \rho_s)$ is a complete metric space and its uniform structure is stronger than the uniformity that generates the metric ρ. Finally for the set $X(A, U)$ we consider the third uniformity which is determined by the following base:

$$E_{Xw}(\epsilon) = \{(x_1, x_2) \in X(A, U) \times X(A, U) : \tag{9.1}$$

$$\mathrm{mes}\{t \in \Omega : \ |\nabla x_1(t) - \nabla x_2(t)| \geq \epsilon\} \leq \epsilon\},$$

where $\epsilon > 0$. (Note that if $x, y \in X(A, U)$ and $\nabla x = \nabla y$, then $x = y$ [28, Theorem 2.4.1].) It is easy to see that this uniform structure is metrizable (by a metric ρ_w) and weaker than the uniformity which generates the metric ρ.

For variational problems considered in this section we can obtain strong versions of Theorems 1.1 and 8.1. These strong versions establish generic strong well-posedness of the minimization problem on the space (X, ρ_s), while in Theorems 1.1 and 8.1 it is obtained on (X, ρ). They are derived from Theorems 1.1 and 8.1, Proposition 4.4, and the following proposition.

PROPOSITION 9.1. *Let $f \in \mathcal{M}(A, U)$,*

$$c_0 > \inf\left\{\int_\Omega f(t, x(t), \nabla x(t))dt : \ x \in X(A, U)\right\} \tag{9.2}$$

and let

$$Y = \left\{x \in X(A, U) : \ \int_\Omega f(t, x(t), \nabla x(t))dt \leq c_0\right\}. \tag{9.3}$$

Then for each $\epsilon > 0$ there exists $\delta > 0$ such that if $x_1, x_2 \in Y$ and $(x_1, x_2) \in E_{Xw}(\delta)$, then $\rho_s(x_1, x_2) \leq \epsilon$.

Proof. Let $\epsilon > 0$. In the case $m > 1$ by [28, Theorem 2.4.1] there exists a constant $c > 0$ such that $||h||_{L^1(\Omega)} \leq c||\nabla h||_{L^1(\Omega)}$ for all $h \in W_0^{1,1}(\Omega)$. In the case $m = 1$ set $c = 1$. Choose a positive number

$$\Delta < (32(c + 1)(\mathrm{mes}(\Omega) + 1))^{-1}\epsilon.$$

By the property (iii) (see the definition of $\mathcal{M}(A, U)$) and (9.3), the family of functions $\{|\nabla x(\cdot)| : \ x \in Y\}$ is uniformly integrable. Therefore there exists $\gamma \in (0, \Delta)$ such that for each $x \in Y$ and each measurable set $e \subset \Omega$ satisfying $\mathrm{mes}(e) \leq \gamma$ the inequality $\int_e |\nabla x(t)|dt \leq \Delta$ holds. Choose a positive number $\delta < (8c + 8)^{-1}(\mathrm{mes}(\Omega) + 1)^{-2}\gamma$.

Assume that $x_1, x_2 \in Y$ and $(x_1, x_2) \in E_{Xw}(\delta)$. There exists a measurable set $e \subset \Omega$ such that $\mathrm{mes}(e) \le \delta$ and $|\nabla x_1(t) - \nabla x_2(t)| \le \delta$, $t \in \Omega \setminus e$. It follows from these inequalities and the definition of γ and δ that

$$\int_e |\nabla x_i(t)|dt \le \Delta, \qquad i = 1, 2, \int_\Omega |\nabla x_1(t) - \nabla x_2(t)|dt \tag{9.4}$$

$$\le \int_e |\nabla x_1(t) - \nabla x_2(t)|dt + \int_{\Omega\setminus e} |\nabla x_1(t) - \nabla x_2(t)|dt \le 2\Delta + \delta \mathrm{mes}(\Omega).$$

In the case $m = 1$ we have

$$|x_1(t) - x_2(t)| \le \int_\Omega |\nabla x_1(s) - \nabla x_2(s)|ds, \qquad t \in \Omega,$$

and by (9.4) and the definition of δ and Δ,

$$\rho_s(x_1, x_2) = ||x_1 - x_2||_{W^{1,1}(\Omega)} \le (\mathrm{mes}(\Omega) + 1)||\nabla x_1 - \nabla x_2||_{L^1(\Omega)}$$

$$\le (\mathrm{mes}(\Omega) + 1)(2\Delta + \delta \mathrm{mes}(\Omega)) < \epsilon.$$

In the case $m > 1$ it follows from (1.4), the definition of c, (9.4), and the definition of δ, Δ that

$$\rho_s(x_1, x_2) = ||x_1 - x_2||_{L^1(\Omega)} + ||\nabla x_1 - \nabla x_2||_{L^1(\Omega)}$$

$$\le (c + 1)||\nabla x_1 - \nabla x_2||_{L^1(\Omega)} \le (c + 1)(2\Delta + \delta \mathrm{mes}(\Omega)) < \epsilon.$$

This completes the proof of the proposition. □

Proposition 9.1 and the completeness of the space $(X(A, U), \rho_s)$ imply the following result.

PROPOSITION 9.2. *Assume that* $f \in \mathcal{M}(A, U)$, $\{x_i\}_{i=1}^\infty$ *is a Cauchy sequence in the space* $X(A, U)$ *with the metric* ρ_w, *and the sequence* $\{\int_\Omega f(t, x_i(t), \nabla x_i(t))dt\}_{i=1}^\infty$ *is bounded. Then there is* $x_* \in X(A, U)$ *such that* $\rho_s(x_i, x_*) \to 0$ *as* $i \to \infty$ *and, moreover, if* $m = 1$, *then* $x_i(t) \to x_*(t)$ *as* $i \to \infty$ *uniformly on* $[T_1, T_2]$.

From now on we consider a fixed set-valued mapping $A : \Omega \to 2^{R^n} \setminus \{\emptyset\}$ for which graph(A) is a closed subset of the space $\Omega \times R^n$ with the product topology and a set-valued mapping $\tilde{U}_A :$ graph$(A) \to 2^{R^n} \setminus \{\emptyset\}$, where $\tilde{U}_A(t, x) = R^n$, $(t, x) \in$ graph(A). For each $f \in \mathcal{M}(A, \tilde{U}_A)$ we define $I^{(f)} : X(A, \tilde{U}_A) \to R^1 \cup \{\infty\}$ by

$$I^{(f)}(x) = \int_\Omega f(t, x(t), \nabla x(t))dt, \qquad x \in X(A, \tilde{U}_A).$$

Consider the space of set-valued mappings $\mathcal{A}_2$ and the space of integrands $\mathcal{A}_{11}$ defined in section 1. Denote by $\mathcal{A}_0$ the set of all functions $f \in \mathcal{A}_{11}$ which do not depend on x. Clearly $\mathcal{A}_0$ is a closed subset of $\mathcal{A}_{11}$ with the strong topology. We consider the topological subspace $\mathcal{A}_0 \subset \mathcal{A}_{11}$ with the relative weak and strong topologies.

Let a function $F :$ graph$(A) \times R^n \to R^1 \cup \{\infty\}$ have the following properties:

F is measurable with respect to the σ-algebra generated by products of Lebesgue measurable subsets of Ω and Borel subsets of $R^n \times R^n$; $F(t, \cdot, \cdot)$ is lower semicontinuous for almost every $t \in \Omega$; there exists an integrable scalar function $\psi_F(t) \geq 0,\ t \in \Omega$, such that $F(t, x, u) \geq \psi_F(t)$ for all $(t, x, u) \in$ graph $(A) \times R^n$.

Clearly, for each $g \in \mathcal{M}(A, \tilde{U}_A)$, $g + F \in \mathcal{M}(A, \tilde{U}_A)$.

For each $a = (a_1, a_2) \in \mathcal{A}_0 \times \mathcal{A}_2$ we define $J_a : X(A, \tilde{U}_A) \to R^1 \cup \{\infty\}$ by

$$J_a(x) = I^{(a_1+F)}(x), \quad x \in S_{a_2}, \quad J_a(x) = \infty, \quad x \in X(A, \tilde{U}_A) \setminus S_{a_2}.$$

Here $S_{a_2} = X(A, a_2)$ (see (1.19)). Denote by $\mathcal{A}$ the closure of the set $\{a \in \mathcal{A}_0 \times \mathcal{A}_2 : \inf(J_a) < \infty\}$ in the space $\mathcal{A}_0 \times \mathcal{A}_2$ with the strong topology. We assume that $\mathcal{A}$ is nonempty.

THEOREM 9.1. *The minimization problem for J_a on $(X(A, \tilde{U}_A), \rho_s)$ is generically strongly well posed with respect to $\mathcal{A}$.*

Proof. We will show that the following assertion holds: The minimization problem for J_a on $(X(A, \tilde{U}_A), \rho_w)$ is generically strongly well posed with respect to $\mathcal{A}$.

This assertion is proved analogously to Theorem 1.1. Note that Propositions 4.1 and 9.2 imply the lower semicontinuity of J_a for all $a \in \mathcal{A}_0 \times \mathcal{A}_2$, (H2) follows from Proposition 9.2, and (A3) is derived from a modification of Lemma 5.1. In this modification the perturbation $g = g(t, x, u)$ does not depend on x and in the last line of the statement of Lemma 5.1 ρ is substituted by ρ_w. The proof of this modification is analogous to the proof of Lemma 5.1. In the relation (5.5), ρ is substituted by ρ_w and E_X is substituted by E_{Xw}, and in (5.20) g is defined by

$$g(t, x, u) = \epsilon_0(\gamma)\bar{\phi}(|u - \bar{u}_s(t)|^2), \qquad (t, x, u) \in R^m \times R^n \times R^n.$$

Thus there exists an everywhere dense (in the strong topology) set $\mathcal{B} \subset \mathcal{A}$ which is a countable intersection of open (in the weak topology) subsets of $\mathcal{A}$ such that for any $a \in \mathcal{B}$ the assertions (1) and (2) of Theorem 2.1 hold with $(X, \rho) = (X(A, \tilde{U}_A), \rho_w)$ and $f_b = J_b$, $b \in \mathcal{A}$.

Let $a = (a_1, a_2) \in \mathcal{B}$. By the assertion (1) of Theorem 2.1 $\inf(J_a)$ is finite and attained at a unique element $\bar{x} \in X(A, \tilde{U}_A)$. In order to complete the proof of the theorem it is sufficient to show that the assertion (2) of Theorem 2.1 holds with $(X, \rho) = (X(A, \tilde{U}_A), \rho_s)$ and $f_b = J_b$, $b \in \mathcal{A}$.

By Proposition 4.4 there exists an open (in the weak topology) neighborhood $\mathcal{V}_1$ of a_1 in $\mathcal{A}_0$ such that for each $b \in \mathcal{V}_1$ and each $x \in X(A, \tilde{U}_A)$ satisfying $I^{(b+F)}(x) \leq \inf(J_a) + 1$ the following relation holds:

$$I^{(a_1+F)}(x) \leq I^{(b+F)}(x) + 1 \leq \inf(J_a) + 2. \tag{9.5}$$

Let $\epsilon \in (0, 1)$. It follows from Proposition 9.1 that there exists $\epsilon_0 \in (0, \epsilon)$ such that for each $x_1, x_2 \in X(A, \tilde{U}_A)$ satisfying

$$I^{(a_1+F)}(x_i) \leq \inf(J_a) + 2, \quad i = 1, 2, \qquad \text{and } \rho_w(x_1, x_2) \leq \epsilon_0 \tag{9.6}$$

the relation $\rho_s(x_1, x_2) \leq \epsilon$ holds.

By the assertion (2) of Theorem 2.1 which holds for the space $(X(A, \tilde{U}_A), \rho_w)$, there are a neighborhood $\mathcal{V}$ of a in $\mathcal{A}$ with the weak topology and $\delta > 0$ such that for each $b \in \mathcal{V}$, $\inf(J_b)$ is finite, and if $z \in X(A, \tilde{U}_A)$ satisfies

$$J_b(z) \leq \inf(J_b) + \delta, \tag{9.7}$$

then

$$\rho_w(\bar{x}, z) \leq \epsilon_0 \quad \text{and} \quad |J_b(z) - J_a(\bar{x})| \leq \epsilon_0. \tag{9.8}$$

We may assume that

$$\mathcal{V} \subset \mathcal{V}_1 \times \mathcal{A}_2. \tag{9.9}$$

Now assume that $b = (b_1, b_2) \in \mathcal{V}$, and $z \in X(A, \tilde{U}_A)$ satisfies (9.7). Then (9.8) holds. By (9.8), (9.9), and the definition of $\mathcal{V}_1$ (see (9.5))

$$I^{(a_1+F)}(z) \leq \inf(J_a) + 2.$$

It follows from this inequality, (9.8), and the definition of ϵ_0 (see (9.6)) that the relation $\rho_s(\bar{x}, z) \leq \epsilon$ holds. Thus the assertion (2) of Theorem 2.1 holds with $(X, \rho) = (X(A, \tilde{U}_A), \rho_s)$ and $f_b = J_b$, $b \in \mathcal{A}$. This completes the proof of the theorem. □

Note that for the class of variational problems considered here we can also prove an analog of Theorem 8.1 in which only integrands are subject to variations.

Example. Let us consider the scalar variational problem

$$\int_0^1 f(t, x'(t))dt \to \min, \qquad x(0) = x(1) = 0,$$

where $f(t, u) = 0$ for $u \in \{-1, 1\}$ and $f(t, u) = \infty$ otherwise. Clearly the functions x_* and $\tilde{x}$ defined by

$$x_*(t) = t,\ t \in [0, 2^{-1}], \quad x_*(t) = 1 - t, \quad t \in (1/2, 1], \tag{9.10}$$

$$\tilde{x}(t) = -x_*(t), \qquad t \in [0, 1]$$

are solutions of the problem. Define a continuous function $\phi : R^2 \to R^1$ by

$$\phi(t, u) = \min\{1, |u - 1|(1/2 - t)\}, \qquad (t, u) \in (-\infty, 1/2] \times R^1,$$

$$\phi(t, u) = \min\{1, |u + 1|(t - 1/2)\}, \qquad (t, u) \in (1/2, \infty) \times R^1,$$

and for each $r \in (0, 1)$ define a function f_r by

$$f_r(t, u) = f(t, u) + r\phi(t, u), \qquad (t, u) \in [0, 1] \times R^1.$$

It is easy to see that $f_r \to f$ as $r \to 0$, and for each $r \in (0, 1)$ the variational problem

$$\int_0^1 f_r(t, x'(t))dt \to \min, \qquad x(0) = x(1) = 0$$

has a unique solution x_* defined by (9.10).

Acknowledgment. The author is very grateful to the referees for their helpful comments and suggestions.

REFERENCES

[1] J.M. Ball and N.S. Nadirashvili, *Universal singular sets for one-dimensional variational problems*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 429–438.

[2] A. Cellina and G. Colombo, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Linéare, 7 (1990), pp. 97–106.

[3] A. Cellina and C. Mariconda, *The existence question in the calculus of variations: A density result*, Proc. Amer. Math. Soc., 120 (1994), pp. 1145–1150.

[4] A. Cellina and S. Zagatti, *An existence result in a problem of the vectorial case of calculus of variations*, SIAM J. Control Optim., 33 (1995), pp. 960–970.

[5] L. Cesari, *Optimization—Theory and Applications*, Springer-Verlag, New York, 1983.

[6] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[7] G. Crasta and A. Malusa, *Existence results for noncoercive variational problems*, SIAM J. Control Optim., 34 (1996), pp. 2064–2076.

[8] F.S. De Blasi and J. Myjak, *Sur la convergence des approximations successives pour les contractions non linéaires dans un espace de Banach*, C. R. Acad. Sci. Paris Sér. A-B, 283 (1976), pp. 185–187.

[9] F.S. De Blasi and J. Myjak, *Generic flows generated by continuous vector fields in Banach spaces*, Adv. in Math., 50 (1983), pp. 266–280.

[10] R. Deville, R. Godefroy, and V. Zizler, *Smoothness and Renorming in Banach Spaces*, Longman Scientific and Technical, Harlow, UK, 1993.

[11] J. Diestel and J.J. Uhl, *Vector Measures*, Amer. Math. Soc., Providence, RI, 1977.

[12] J.L. Doob, *Measure Theory*, Springer-Verlag, New York, 1994.

[13] A.D. Ioffe and V.M. Tikhomirov, *Theory of Extremal Problems*, North-Holland, New York, 1979.

[14] A.D. Ioffe and A.J. Zaslavski, *Variational principles and well-posedness in optimization and calculus of variations*, SIAM J. Control Optim., 38 (2000), pp. 566–581.

[15] J.L. Kelley, *General Topology*, Van Nostrand, Princeton, NJ, 1955.

[16] M. Marcus and A.J. Zaslavski, *The structure of extremals of a class of second order variational problems*, Ann. Inst. H. Poincaré, Anal. Non Linéare, 16 (1999), pp. 593–629.

[17] V.G. Maz'ja, *Sobolev Spaces*, Springer-Verlag, Berlin, 1985.

[18] E.J. McShane, *Existence theorem for the ordinary problem of the calculus of variations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 3 (1934), pp. 181–211.

[19] M.D.P. Monteiro Marques and A. Ornelas, *Genericity and existence of a minimum for scalar integral functionals*, J. Optim. Theory Appl., 86 (1995), pp. 421–431.

[20] B.S. Mordukhovich, *Approximation Methods in Optimization and Control*, Nauka, Moscow, 1988.

[21] B.S. Mordukhovich, *Existence theorems in nonconvex optimal control*, in Calculus of Variations and Optimal Control, CRC Press, Boca Raton, FL, 1999, pp. 175–197.

[22] Ch. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, Heidelberg, New York, 1967.

[23] S. Reich and A.J. Zaslavski, *Convergence of generic infinite products of nonexpansive and uniformly continuous operators*, Nonlinear Anal., 36 (1999), pp. 1049–1065.

[24] R.T. Rockafellar, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.

[25] L. Tonelli, *Fondamenti di Calcolo delle Variazioni*, Zanicelli, Bolonia, 1921–1923.

[26] A.J. Zaslavski, *Dynamic properties of optimal solutions of variational problems*, Nonlinear Anal., 27 (1996), pp. 895–932.

[27] A.J. Zaslavski, *Existence of solutions of optimal control problems without convexity assumptions*, Nonlinear Anal., to appear.

[28] W.P. Zeimer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 281–305

EXTENDED HAMILTON–JACOBI CHARACTERIZATION OF VALUE FUNCTIONS IN OPTIMAL CONTROL[*]

GRANT N. GALBRAITH[†]

Abstract. This paper examines generalized solutions to the Hamilton–Jacobi equation. The Hamiltonian $H(t, x, p)$ is assumed convex in p but is not constrained to have linear growth in this variable. This corresponds to a certain class of generalized Bolza problems, related to optimal control. Lower semicontinuous solutions are considered and it is shown that there is a unique solution, the so-called value function of the underlying Bolza problem. In proving the main result we use recent improved necessary optimality conditions. Viability is also used in a new way, in connection to differential inclusions with unbounded images.

Key words. Hamilton–Jacobi equation, optimal control, Bolza problem, nonsmooth analysis, viability theory, viscosity solution

AMS subject classifications. Primary, 49K24, 49J52; Secondary, 49L25, 35B37

PII. S0363012998347882

1. Introduction.

The classical Cauchy problem for the Hamilton–Jacobi equation is the partial differential equation with initial condition

$$\begin{aligned} u_t(t,x) + H(t,x,u_x(t,x)) &= 0, \qquad (t,x) \in (0,\infty) \times \mathbb{R}^n, \\ u(0,x) &= \varphi(x), \quad x \in \mathbb{R}^n. \end{aligned} \tag{1.1}$$

If the Hamiltonian $H(t, x, p)$ is convex with respect to p, there are connections between solutions to (1.1) and optimization problems involving a function dual to the Hamiltonian. This function L, called the Lagrangian, is derived from H using the Legendre–Fenchel transform as follows:

$$L(t,x,v) = \sup_{p \in \mathbb{R}^n} \{ \langle\, p, v \,\rangle - H(t,x,p) \}. \tag{1.2}$$

Here $\langle\, p, v \,\rangle$ denotes the inner product of p and v. It is possible for L to assume the value $+\infty$, and it is this important feature which allows constraints to be incorporated directly into L in the form of "infinite penalties." The *generalized problem of Bolza* we parameterize here in (τ, ξ) as

$$(\mathcal{P}_{\tau,\xi}) \qquad \text{minimize} \;\; \Gamma(x) := \varphi(x(0)) + \int_0^\tau L\big(t, x(t), \dot{x}(t)\big) dt,$$

where we are minimizing Γ over $\mathcal{A}_n^1[0,\tau]$ (the space of all absolutely continuous arcs $x : [0,\tau] \to \mathbb{R}^n$) with $x(\tau) = \xi$. This has a simple appearance, and yet a wide range of problems involving optimal control, differential inclusions, and constraints can be expressed in this form. See Loewen [20] or the introductory chapter of Clarke [8] for a discussion on the equivalences between these various formulations.

An extended-real-valued function is called *proper* if it never takes on the value $-\infty$ and is not identically $+\infty$. When $H(t, x, \cdot\,)$ is proper, lower semicontinuous (lsc),

[*] Received by the editors November 20, 1998; accepted for publication (in revised form) March 28, 2000; published electronically August 24, 2000.
http://www.siam.org/journals/sicon/39-1/34788.html

[†] Department of Mathematics, University of California, Davis, CA, 95616-8633 (gng@math.ucdavis.edu).

and convex for each (t, x), then $L(t, x, \cdot\,)$ also has these properties. Furthermore, we can then retrieve H from L by performing the Legendre–Fenchel transform a second time:

$$H(t, x, p) = \sup_{v \in \mathbb{R}^n} \{ \langle\, p, v \,\rangle - L(t, x, v) \}. \tag{1.3}$$

Thus we have a one-to-one correspondence between Hamiltonians and Lagrangians in this convex case, and every equation of the form (1.1) can be paired with a problem of the form $(\mathcal{P}_{\tau,\xi})$.

Definition 1.1. *The value function* $V : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ *is defined from* Γ *as follows:*

$$V(\tau, \xi) = \begin{cases} \inf\{\, \Gamma(x) \mid x \in \mathcal{A}_n^1[0, \tau],\ x(\tau) = \xi \} & \text{if } \tau > 0, \\ \varphi(\xi) & \text{if } \tau = 0, \\ +\infty & \text{if } \tau < 0. \end{cases} \tag{1.4}$$

When the value function is real-valued and differentiable, it is known to satisfy (1.1) in the classical sense, and yet in many situations the value function is extended-real-valued and merely lsc. Our goal is to provide a way to interpret (1.1) for extended-real-valued, lsc functions u in such a way that the value function V will be the unique solution. This is not the first such attempt, as the discussion below will reveal; however, we achieve this uniqueness result over a much larger class of Hamiltonians than had been previously attained.

The first of these earlier results dates back to the early 1980s when Crandall and Lions [10] introduced *viscosity solutions*, with Crandall, Evans, and Lions giving a simplified approach in [11]. Viscosity solutions attracted a lot of attention, and over subsequent years a sizable literature developed from many authors dealing with, among other issues, existence and uniqueness of solutions.

In section 2 we will go into more detail about the classes of Hamiltonians and solutions addressed by viscosity theory, but for now we will mention briefly that they typically require that the solution be uniformly continuous and bounded and that the Hamiltonian satisfy some sort of uniform continuity condition. For an up-to-date account of the subject, see the recent book of Bardi and Capuzzo-Dolcetta [4].

Another class of continuous generalized solutions, named *minimax solutions*, was developed by Subbotin, initially in [29], with a more detailed approach in [31]. For a large class of Hamiltonians, minimax solutions and viscosity solutions are equivalent.

As mentioned earlier, even continuity may fail for the value function, and thus there is a need for semicontinuous solutions to (1.1). Ishii [18] examined issues regarding the existence of possibly discontinuous solutions, which were defined using upper semicontinuous (usc) and lsc envelopes. Barles and Perthame [5] used Ishii's notion of solution and obtained an early uniqueness result. In [7], Barron and Jensen extended viscosity solutions to certain semicontinuous functions for Hamiltonians that are convex in p and provided a uniqueness result. The paper by Barles [6] provides some extensions and an informal discussion of Barron and Jensen's ideas.

Frankowska [16] also considered lsc solutions and presented a uniqueness result for the Hamilton–Jacobi equation arising from the Mayer problem in optimal control. This corresponds to considering $(\mathcal{P}_{\tau,\xi})$ with the Lagrangian $L(t, x, \cdot\,)$ as the indicator of a compact set for each (t, x). The indicator function $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ of a set

$C \subset \mathbb{R}^n$ is defined as

$$\Psi_C(v) = \begin{cases} 0 & \text{if } v \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{1.5}$$

The Mayer problem then has $L(t,x,v) = \Psi_{F(t,x)}(v)$ for some set-valued mapping F with compact, convex images. This gives rise to a Hamiltonian which is finite and positively homogeneous in p. An important feature of [16], which inspired the approach taken here, is Frankowska's use of viability theory.

This paper will present a result on the existence and uniqueness of solutions to (1.1) with a solution concept similar to those in [7] and [16]. As in [16], viability plays an important role in deriving our main result; however, we advance by working with unbounded differential inclusions. The differential inclusions we wish to employ are epigraphical set-valued mappings, whose images by their very nature cannot be bounded.

Typically, viability properties of a set K are determined through properties of the tangent and normal cones to K. Our viability approach differs in that we use normal cone properties of the *reachable set* of K. This allows us to deal with the epigraphical differential inclusion directly, without having to resort to truncations of the epigraph. Ultimately this allows us greater range in our choice of Lagrangians and Hamiltonians. In particular, we can go beyond cases where L has the form $L(t,x,v) = L_1(t,x,v) + \Psi_{F(t,x)}(v)$ for some Lipschitz function L_1. (This case can be reduced to a Mayer problem through truncations of the epigraphs $L(t,x,\,\cdot\,)$.) More generally speaking, we are not restricting ourselves to Hamiltonians $H(t,x,p)$ which have linear growth in p (this corresponds to the essential domain of $L(t,x,v)$ being bounded in the variable v for each (t,x)). Instead we have the extremely mild growth condition (A1) on the Hamiltonian, given in the next section.

The proof of our main theorem also relies on the improved necessary optimality conditions achieved in the papers [21], [22], and [23] of Loewen and Rockafellar. The assumption (A2) used in the main theorem is designed precisely to allow the application of these necessary conditions, which are given in section 5.

Because nonsmooth analysis will be used here, we will need the notion of a subgradient. We use the notation adopted in Rockafellar and Wets [27]. The symbol $\overline{\mathbb{R}}$ is used to denote $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$. For $v \in \mathbb{R}^n$ and an lsc function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$,

(i) v is a *regular subgradient of f at x*, written $v \in \widehat{\partial} f(x)$, if

$$\liminf_{y \to x,\ y \neq x} \frac{f(y) - f(x) - \langle v, y - x \rangle}{|y - x|} \geq 0;$$

(ii) v is a (general) *subgradient of f at x*, written $v \in \partial f(x)$, if there are sequences $x^\nu \to x$ with $f(x^\nu) \to f(x)$ and $v^\nu \in \widehat{\partial} f(x^\nu)$ with $v^\nu \to v$;

(iii) v is a *horizon subgradient of f at x*, written $v \in \partial^\infty f(x)$, if there are sequences $x^\nu \to x$ with $f(x^\nu) \to f(x)$ and $v^\nu \in \widehat{\partial} f(x^\nu)$ with $\lambda^\nu v^\nu \to v$ for some sequence $\lambda^\nu \searrow 0$.

There is another way to define regular subgradients, equivalent to the above, which is more in line with the ideas used for viscosity solutions. We can say a vector v belongs to $\widehat{\partial} f(x)$ if and only if on some neighborhood W of x, there exists a function g continuously differentiable on W with $g(x) = f(x)$, $\nabla g(x) = v$, and $g(x') < f(x')$ for all $x' \in W$ with $x' \neq x$. (See Prop. 8.5 in [27] for details.)

Given a closed set K and a point $x \in K$, we can define the general normal cone $N_K(x)$ and the regular normal cone $\widehat{N}_K(x)$ from the corresponding subgradients as follows:

$$N_K(x) = \partial \Psi_K(x), \tag{1.6}$$

$$\widehat{N}_K(x) = \widehat{\partial} \Psi_K(x). \tag{1.7}$$

We use the notation $\operatorname{dom} f$ for the *effective domain of f*, which is the set $\{x : f(x) < \infty\}$. We can now specify what we mean by a generalized solution to (1.1).

Definition 1.2. *A function $u : \mathbb{R} \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is a solution to* (1.1) *if it satisfies the following:*

(a) *u is proper and lsc, with* $\operatorname{dom} u \subset [0, \infty) \times \mathbb{R}^n$*;*

(b) *$u(0, \xi) = \varphi(\xi)$ for all $\xi \in \mathbb{R}^n$;*

(c) *For every $(\tau, \xi) \in \operatorname{dom} u$, every $(\sigma, \eta) \in \partial u(\tau, \xi)$ satisfies*

$$\begin{cases} \sigma + H(\tau, \xi, \eta) \leq 0 & \textit{if } \tau = 0, \\ \sigma + H(\tau, \xi, \eta) = 0 & \textit{if } \tau > 0. \end{cases} \tag{1.8}$$

Note the use of general subgradients in the definition. This differs from the papers [7] and [16], where the emphasis is on regular subgradients. The choice of using general subgradients in our definition of solution has been made because of the richer calculus rules these subgradients enjoy and due to their emphasis in [27]. Definition 1.2 aside, the general subgradient is a better choice later in the paper when characterizing sub-Lipschitz and pseudo-Lipschitz properties of set-valued mappings (see Mordukhovich [24]). Thus we make it our subgradient of choice. Note, however, that if H is continuous, we can replace $\partial u(\tau, \xi)$ with $\widehat{\partial} u(\tau, \xi)$ in the above, and the resulting definition of solution will be equivalent to Definition 1.2. As will be seen, the continuity of H is a property that holds in the main theorem below.

Due to the viability approach we will be taking, let us introduce some notation and terminology associated with set-valued mappings. We write $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ to denote a mapping which associates a subset $F(x)$ of $\mathbb{R}^m$ to each point $x \in \mathbb{R}^n$. The graph of F, denoted $\operatorname{gph} F$, is the set of points $\{(x, y) \mid y \in F(x)\}$ in $\mathbb{R}^n \times \mathbb{R}^m$.

Of particular importance are the epigraphical mappings. These are set-valued mappings $E_f : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m+1}$, defined from a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ as $E_f(x) = \operatorname{epi} f(x, \cdot\,)$, where

$$\operatorname{epi} f(x, \cdot\,) := \{(v, \alpha) \mid v \in \operatorname{dom} f(x, \cdot\,),\ \alpha \geq f(x, v)\}. \tag{1.9}$$

Thus the graph of E_f corresponds to the epigraph of f, and the sets $E_f(x)$ are closed if and only if $f(x, \cdot\,)$ is lsc for each x.

Another way in which our Theorem 2.2 improves upon previous results is that we are able to relax the Lipschitz behavior of the Lagrangian set-valued mapping E_L. Typically, one defines Lipschitz continuity with respect to the Hausdorff metric. However, epigraphical mappings are unbounded, and placing a Lipschitz assumption on E_L proves to be rather restrictive.

The closed unit ball is denoted by $\mathbb{B}$. To say $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *Lipschitz* on some open set W means there exists a constant $k > 0$, a Lipschitz constant, so that

$$F(x') \subset F(x) + k|x' - x|\mathbb{B} \quad \text{for all } x', x \in W. \tag{1.10}$$

The notion of *sub-Lipschitz* continuity relaxes the condition in (1.10) by allowing a truncation on the left-hand side of this inclusion. The Lipschitz constant k is then allowed to grow with the size of the corresponding truncation. Specifically, F is sub-Lipschitz on some open set W if there exists a positive ρ_0 so that for each $\rho > \rho_0$, there exists a $k > 0$, a ρ-Lipschitz constant, so that

$$F(x') \cap \rho\mathbb{B} \subset F(x) + k|x' - x|\mathbb{B} \quad \text{for all } x', x \in W. \tag{1.11}$$

Mordukhovich (section 5 of [24]) shows how one can characterize Lipschitz and sub-Lipschitz properties of a set-valued mapping F through properties of the normal cones $N_{\operatorname{gph} F}(x, v)$ and the associated *coderivative mapping* $D^*F(x|v) : \mathbb{R}^m \to \mathbb{R}^n$ defined as

$$D^*F(x|v)(y) = \{w : (w, -y) \in N_{\operatorname{gph} F}(x, v)\}. \tag{1.12}$$

Moreover, he shows how these coderivatives can lead to bounds for the Lipschitz and ρ-Lipschitz constants (see also section 9.F of [27]). For an epigraphical set-valued mapping E_f, the coderivative can be expressed using the general and horizon subgradients of f.

Our uniqueness result applies to cases where the epigraphical mapping of the Lagrangian has a special kind of sub-Lipschitz behavior. This Lipschitz condition is given in a subgradient form on the Hamiltonian as assumption (A2) in the next section. This subgradient expression is placing bounds on a certain coderivative mapping.

Throughout this paper, H will be seen to depend on t for $t \in \mathbb{R}$, while (1.8) implies that the behavior of $H(t, x, p)$ for $t < 0$ has no effect on our notion of solution. The reason for defining H on a larger domain than is necessary is to avoid complications that may arise with relative neighborhoods of points (t, x). This shouldn't present a problem if we have a Hamiltonian defined only for $t \geq 0$. By setting $H(t, x, p) = H(0, x, p)$ for $t < 0$, we can easily extend the domain of H without affecting (1.8). We can extend L in the same manner.

2. Statement of the main result.

The main result of this paper is Theorem 2.2, which depends upon the assumptions (A) and (A0)–(A2) given below. After a comparison with other uniqueness results in the literature, the latter half of this section gives some implications of (A1) and (A2) as well as connections between the Hamiltonian and the Lagrangian under these assumptions.

When we say that a function u exhibits *linear growth*, we mean there exists a constant $k > 0$ so that $|u(x)| \leq k(1 + |x|)$ for all x. It will be useful to define the class of functions which have only "half" linear growth, in that no restrictions are placed on how the positive values of u may behave.

Definition 2.1. *A function $u : \mathbb{R}^n \to \overline{\mathbb{R}}$ satisfies the lower linear growth (LLG) condition if there exists $k > 0$ such that*

$$u(x) \geq -k(1 + |x|) \quad \textit{for all} \quad x \in \mathbb{R}^n. \tag{2.1}$$

A function $u : \mathbb{R} \times \mathbb{R}^n \to \overline{\mathbb{R}}$ satisfies the uniform lower linear growth (ULLG) condition if for each $T > 0$, there exists $k > 0$ such that

$$u(t, x) \geq -k(1 + |x|) \quad \textit{for all } (t, x) \in [-T, T] \times \mathbb{R}^n. \tag{2.2}$$

These mild conditions are effectively ruling out countercoercivity in the state variable, but not necessarily in t. A function f is *coercive* if it is bounded from below and $\liminf_{|x|\to\infty} f(x)/|x| = +\infty$, and f is *countercoercive* if $\liminf_{|x|\to\infty} f(x)/|x| = -\infty$. For example, the function $u(t,x) = -t^2$ satisfies the ULLG condition, whereas $u(t,x) = -|x|^2$ does not.

Basic assumption. We have a Hamiltonian $H : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

(A) *$H(t,x,\,\cdot\,)$ is proper, lsc, and convex for each $(t,x) \in \mathbb{R} \times \mathbb{R}^n$.*

Initial cost assumption. We have an extended-real-valued function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$.

(A0) *The function φ is proper, lsc, and satisfies the LLG condition* (2.1).

Hamiltonian assumptions. We have a Hamiltonian $H : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

(A1) *There exists a convex, nondecreasing function $\mu : [0,\infty) \to \mathbb{R}$ and positive constants α and β so that*

$$H(t,x,p) \le \mu(|p|) + (|t| + |x|)(\beta + \alpha|p|) \quad \textit{for all } (t,x,p) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n.$$

(A2) *H is lsc on $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$, and for each $(\bar{t},\bar{x}) \in \mathbb{R} \times \mathbb{R}^n$, there exists a neighborhood W of $(\bar{t},\bar{x})$ and a positive constant k so that at every point $(t,x,p) \in W \times \mathbb{R}^n$, every $(w_1,w_2,v) \in \partial H(t,x,p)$ satisfies*

$$|(w_1,w_2)| \le k(1 + |v| + |L(t,x,v)|)(1 + |p| + |H(t,x,p)|). \tag{2.3}$$

Assumption (A1) is a mild growth condition on H that is directly related to the *stronger growth condition* introduced by Rockafellar [25] to help guarantee the existence of optimal arcs in generalized Bolza problems. Note that taken together, (A) and (A1) imply that for each (t,x), H is finite and convex in p, which forces H to be continuous in p (see Corollary 2.36 in [27]).

Assumption (A2) is a special kind of sub-Lipschitz behavior of the epigraphical mapping $E_H : \mathbb{R} \times \mathbb{R}^n \rightrightarrows \mathbb{R}^{n+1}$, where E_H is defined as

$$E_H(t,x) = \operatorname{epi} H(t,x,\,\cdot\,). \tag{2.4}$$

See Proposition 2.6 below for a connection with the epigraphical mapping E_L of the Lagrangian.

The main theorem below shows that the value function is the unique solution to (1.1) among the class of functions satisfying the ULLG condition.

THEOREM 2.2 (existence and uniqueness). *Under* (A) *and* (A0)–(A2)*, the value function V satisfies the ULLG condition and is a solution to the Cauchy problem* (1.1) *for the Hamilton–Jacobi equation, in the sense given by Definition* 1.2.

Furthermore, if a function u is a solution to (1.1) *as given by Definition* 1.2*, and u satisfies the ULLG condition, then u must be the value function V.*

The proof of the theorem is contained in section 6.

Let us now compare our uniqueness result of Theorem 2.2 to others in the literature. We begin by mentioning that most of the uniqueness (or, comparison) results for viscosity solutions place no convexity assumptions on the Hamiltonian. However, convexity of $H(t,x,p)$ with respect to p will be present in any optimization context, due to the definition of the Hamiltonian as a maximization of functions affine in p.

Viscosity solutions are defined as functions satisfying a pair of inequalities. Originally, the definition was given in terms of smooth support functions, but we can equivalently define them using regular subgradients. A usc function u is termed a *viscosity subsolution* if

$$w + H(t,x,p) \le 0 \quad \text{whenever} \quad (w,p) \in -\widehat{\partial}(-u)(t,x),$$

whereas an lsc function u is a *viscosity supersolution* if

$$w + H(t,x,p) \geq 0 \quad \text{whenever} \quad (w,p) \in \widehat{\partial} u(t,x).$$

A viscosity solution is a continuous function which is simultaneously a subsolution and a supersolution.

Initially in [10], [11], the uniqueness results for viscosity solutions were rather restrictive as they gave uniqueness only over the class of bounded, uniformly continuous functions. Furthermore, the Hamiltonian was also required to have uniform continuity properties in all its variables.

These restrictions were relaxed somewhat in later papers. In [12] the boundedness assumption on the viscosity solution was removed, but the solution was still required to be uniformly continuous. The uniqueness of solutions was extended to the class of all continuous functions in [13], but the Hamiltonian was required to have a certain uniform continuity property and exhibit linear growth in p. The assumptions on the Hamiltonian were subsequently relaxed in [14] and [19]; however, the class of solutions had to be restricted to those with linear growth in the state variable and uniformly continuous initial condition.

For lsc solutions, the uniqueness result of [16] uses hypotheses similar to those in [31] and which are more general than those appearing in [7] or [6]. One assumption is the Lipschitz behavior in x of the mapping $\operatorname{epi} L(t,x,\cdot)$. Another assumption in [16] is that the Hamiltonian be positively homogeneous in p. This is not quite as restrictive as it first seems. For example, through a change of variables, any Bolza problem that has L of the form $L(t,x,v) = L_0(t,x,v) + \Psi_{F(t,x)}(v)$, with L_0 locally Lipschitz in v and F a locally Lipschitz compact set-valued mapping, is covered with such H. Still, this corresponds to a class of Hamiltonians H which, among other properties, must exhibit linear growth in p.

By way of a different change of variables, Subbotin (section 5 of [30]) shows how one can convert a Hamiltonian H into one which is positively homogeneous. Given $H : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, define $h : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ via

$$h(t,x,r,p) = |r| H(t,x,p/|r|) \quad \text{if } r \neq 0. \tag{2.5}$$

For $r = 0$, take limits in the above expression so that h is continuous in (r,p). Then h is positively homogeneous in (r,p). (Although [30] doesn't require convexity, note that in general, h will not be convex in (r,p) unless we restrict $r \geq 0$.) In order to use this technique to gain uniqueness under the original H, the new Hamiltonian h must satisfy certain continuity conditions (for now, let us fix some (t,x) and just write $H(p)$ and $h(r,p)$). These conditions (H5 in [30]) state that h must be Lipschitz continuous on the set where $|p|^2 + r^2 \leq 1$. Assume, as in our Bolza problem context, that $H(p)$ is convex and is dual to a convex Lagrangian $L(v)$. Through subgradient calculus one can show that if $(w,v) \in \partial h(r,p)$ for some (r,p) with $r > 0$, then $v \in \partial H(p/r)$ and $w = -L(v)$. In order for h to be locally Lipschitz at $(0,0)$, $H(p)$ must be globally Lipschitz and thus have linear growth. Again, this corresponds to the Lagrangian having a bounded essential domain. But there is something even more restrictive here. With a bounded essential domain, it is still possible that L may be unbounded on this set. Assume there exists a bounded sequence v^ν with $L(v^\nu)$ unbounded. Then one can find $(r^\nu, p^\nu) \to (0,0)$ with $r^\nu > 0$ and $(-L(v^\nu), v^\nu) \in \partial h(r^\nu, p^\nu)$, so that these subgradients become unbounded, and there is no hope for h to satisfy the required Lipschitz condition.

Frankowska and Plaskasz [17] use the techniques from [16] to obtain uniqueness for lsc solutions in which state constraints are present, something which we do not consider in this paper. In [17], the Bolza problem is presented in an optimal control form. If one converts this control problem into the form $(P_{\tau,\xi})$, assumptions on the dynamics $\dot{x}(t) = f(t, x(t), u(t))$ once again force the velocity to be bounded and the Hamiltonian to have linear growth in p.

Our result covers a broader class of Hamiltonians, as there is absolutely no restriction on the growth of H other than the mild condition (A1), and in this sense Theorem 2.2 improves on previous results. Also, we can now deal with cases where the Hamiltonian does possess linear growth but arises from a Lagrangian that is unbounded in v on its essential domain (as in the discussion on [30] above). Moreover, the assumption (A2) introduces a sub-Lipschitz behavior for epigraphical set-valued mappings of a much greater scope than that of Lipschitz continuity. We hope to devote a separate paper to provide examples and an analysis of the assumption (A2), including the case of mappings which have a fully convex graph.

Rockafellar and Wolenski [28] provide an analysis of the value function and Hamilton–Jacobi theory in an autonomous, fully convex Lagrangian case. They do not present a uniqueness result, but rather they give regularity properties of the value function, develop a method of characteristics, and examine connections to a dual Bolza problem.

Let us note that (A1) and (A2) force the same restrictions on H in t and x jointly. This is not the case in [16]. If we take the case where L is of the form $\Psi_{F(t,x)}(v)$, then our assumptions would require that F be a locally Lipschitz set-valued mapping in (t, x) with compact, convex images. The assumptions in [16] would require, generally speaking, that F be locally Lipschitz in x and only continuous in t. Thus there are certain Hamiltonians not satisfying our assumptions, to which the results of Frankowska's paper can be applied.

The remainder of this section contains some lemmas and propositions giving implications of the hypotheses (A1) and (A2). These will be useful in later sections and in the proof of Theorem 2.2.

LEMMA 2.3. *Assume* (A) *and* (A2) *hold. Then for any* (t, x, p), *if* $(w_1, w_2, 0) \in \partial^\infty H(t, x, p)$, *then* $(w_1, w_2) = (0, 0)$.

Proof. Fix $(t, x, p) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$ and take W and k as given by (A2) and $(w_1, w_2, v) \in \partial^\infty H(t, x, p)$. This means there exist sequences $(t^\nu, x^\nu, p^\nu) \to (t, x, p)$ with $H(t^\nu, x^\nu, p^\nu) \to H(t, x, p)$ and $\lambda^\nu(w_1^\nu, w_2^\nu, v^\nu) \to (w_1, w_2, v)$ with $(w_1^\nu, w_2^\nu, v^\nu) \in \widehat{\partial} H(t^\nu, x^\nu, p^\nu)$ and $\lambda^\nu \searrow 0$. Assume that $(t^\nu, x^\nu) \in W$ for each ν. It follows (see Corollary 10.9 of [27]) that $v^\nu \in \widehat{\partial}_p H(t^\nu, x^\nu, p^\nu)$, and convex analysis tells us that $H(t^\nu, x^\nu, p^\nu) + L(t^\nu, x^\nu, v^\nu) = \langle p^\nu, v^\nu \rangle$ for each ν. Now using the inequality from (A2), we see that

(2.6)
$$\begin{aligned} |(w_1^\nu, w_2^\nu)| \le{}& k(1 + |v^\nu| + |L(t^\nu, x^\nu, v^\nu)|)(1 + |p^\nu| + |H(t^\nu, x^\nu, p^\nu)|) \\ \Rightarrow |\lambda^\nu(w_1^\nu, w_2^\nu)| \le{}& k(\lambda^\nu + |\lambda^\nu v^\nu| + |\lambda^\nu L(t^\nu, x^\nu, v^\nu)|)(1 + |p^\nu| + |H(t^\nu, x^\nu, p^\nu)|) \\ ={}& k(\lambda^\nu + |\lambda^\nu v^\nu| \\ &+ |\lambda^\nu(\langle v^\nu, p^\nu \rangle - H(t^\nu, x^\nu, v^\nu))|)\,(1 + |p^\nu| + |H(t^\nu, x^\nu, p^\nu)|) \\ \Rightarrow |(w_1, w_2)| \le{}& k(|v| + |\langle v, p \rangle|)(1 + |p| + |H(t, x, p)|). \end{aligned}$$

Thus if $v = 0$, this forces $(w_1, w_2) = (0, 0)$. □

The horizon subgradient condition present in Lemma 2.3 can also be expressed in terms of the coderivative D^*E_H of the set-valued mapping E_H (see (1.12)). The lemma is saying that $D^*E_H((t,x)|(p,y))(0) = \{0\}$ at every point (t,x,p,y) in the epigraph of H, from which it follows by the result of Mordukhovich [24, Thm. 5.7] that E_H has the *Aubin property* at every point of the graph of E_H (originally called the pseudo-Lipschitz property as introduced in [2]). In particular, this gives us the following proposition.

PROPOSITION 2.4. *Assume $H(t,x,\cdot)$ is finite and convex for each $(t,x) \in \mathbb{R}\times\mathbb{R}^n$ and that* (A2) *holds (this is true in particular when* (A), (A1)*, and* (A2) *hold). Then H is locally Lipschitz continuous at every $(t,x,p) \in \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n$.*

Proof. Fix (t,x,p). Showing H is locally Lipschitz at this point is equivalent to showing that $\partial^\infty H(t,x,p) = \{(0,0,0)\}$. This equivalence is proved in [27, Thm. 9.13]. Take $(w_1,w_2,v) \in \partial^\infty H(t,x,p)$. From Lemma 2.3, it suffices to show that v must be 0.

Take sequences defining (w_1,w_2,v) as in the proof of Lemma 2.3. The Aubin property at every point of the graph of the mapping E_H implies that E_H is continuous with respect to Painlevé–Kuratowski set convergence [27, Thm. 9.38] and that H has the so-called *epi-continuity property* discussed in [26]. This epi-continuity property of H leads to a result of Attouch [27, Thm. 12.35] which says that the subgradient mappings $\widehat{\partial}_p H(t^\nu,x^\nu,\,\cdot\,)$ graphically converge to $\widehat{\partial}_p H(t,x,\,\cdot\,)$. Since $H(t,x,\,\cdot\,)$ is finite and convex, $\widehat{\partial}_p H(t,x,p)$ is convex, nonempty, and bounded. This implies [27, Ex. 5.34] the existence of finite constants R and N such that $\widehat{\partial}_p H(t^\nu,x^\nu,p^\nu) \subset R\mathbb{B}$ for $\nu \geq N$. But we have that $v^\nu \in \widehat{\partial}_p H(t^\nu,x^\nu,p^\nu)$, so $|v^\nu| \leq R$ for $\nu \geq N$. Since $\lambda^\nu \searrow 0$, we must have $v = 0$. □

LEMMA 2.5. *Assume $H(t,x,\,\cdot\,)$ is finite and convex for each $(t,x) \in \mathbb{R}\times\mathbb{R}^n$ and that* (A2) *holds. Then at every $(\bar{t},\bar{x}) \in \mathbb{R}\times\mathbb{R}^n$, there exists a neighborhood W of $(\bar{t},\bar{x})$ and a positive constant k such that*

$$H(t,x,p) \geq -k(1+|p|) \quad \textit{for all } (t,x,p) \in W\times\mathbb{R}^n.$$

Proof. Take $(\bar{t},\bar{x}) \in \mathbb{R}\times\mathbb{R}^n$. Proposition 2.4 gives us a neighborhood $W_1\times W_2 \subset \mathbb{R}^{n+1}\times\mathbb{R}^n$ of $(\bar{t},\bar{x},0)$ on which H is Lipschitz. This implies the existence of $k_1 > 0$ such that $\widehat{\partial} H(t,x,0) \subset k_1\mathbb{B}$, which in turn says that $\widehat{\partial}_p H(t,x,0) \subset k_1\mathbb{B}$ for all $(t,x) \in W_1$. The Lipschitz property of H also implies that $H(t,x,0) \geq -k_2$ for some $k_2 > 0$ for $(t,x) \in W_1$. Finally, $H(t,x,\,\cdot\,)$ being a convex function gives us that

$$\begin{aligned} H(t,x,p) &\geq H(t,x,0) + \langle v,p\rangle \\ &\geq -k_2 - k_1|p| \end{aligned}$$

if $v \in \widehat{\partial}_p H(t,x,0)$ and $(t,x,p) \in W_1\times\mathbb{R}^n$. Now let $k = \max\{k_1,k_2\}$. □

PROPOSITION 2.6. *Take a dual pair H and L, with both $H(t,x,\,\cdot\,)$ and $L(t,x,\,\cdot\,)$ proper, lsc, and convex for every (t,x) where L can be derived from H via* (1.2) *and H can be derived from L via* (1.3)*. Then the following are equivalent.*

(i) (A2) *holds for H.*

(ii) *L is lsc on $\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n$, and for each $(\bar{t},\bar{x})$, there exists a neighborhood W of $(\bar{t},\bar{x})$ and a positive constant k so that at every point $(t,x,v) \in W\times\mathbb{R}^n$, every $(w_1,w_2,p) \in \partial L(t,x,v)$ satisfies the relation given by* (2.3).

Proof. Assume (i). In the proof of Proposition 2.4 we saw that H has the epi-continuity property as described in [26]. Theorem 11.34 of [27] says that then L also has the epi-continuity property, which implies that the epigraph of L is closed, and so L is lsc on $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$.

To show that the subgradient inequality holds, we would like to use Theorem 3.3 of [26], which requires checking that a condition holds on the horizon subgradients of H. This is exactly the condition shown to hold in Lemma 2.3, and so using [26, Thm. 3.3] we can conclude that, for any given (t, x, v, p),

$$\operatorname{con}\{(w_1, w_2) | (w_1, w_2, p) \in \partial L(t, x, v)\} = - \operatorname{con}\{(w_1, w_2) | (w_1, w_2, v) \in \partial H(t, x, p)\}.$$

Now choose $(\bar{t}, \bar{x})$, bounded neighborhood W, and constant k as given in (2.3). Take $(w_1, w_2, p) \in \partial L(t, x, v)$ for $(t, x, v) \in W \times \mathbb{R}^n$. Then (w_1, w_2) can be written as a finite, convex combination of points in the set $\{(-w_1', -w_2') | (w_1', w_2', v) \in \partial H(t, x, p)\}$. But this implies that

$$\begin{aligned} |(w_1, w_2)| &\le \sup\{|(-w_1', -w_2')| \bigm| (w_1', w_2', v) \in \partial H(t, x, p)\} \\ &\le k(1 + |v| + |L(t, x, v)|)(1 + |p| + |H(t, x, p)|), \end{aligned}$$

the second inequality following by assumption (i). Thus (ii) holds at $(\bar{t}, \bar{x})$ with the same constant k for our neighborhood W.

If we begin instead by assuming (ii), the entire argument presented above will go through, since at every step each result holds symmetrically in L as in H, and we can conclude that (i) holds. □

3. Viability and the value function. The concepts of viability and invariance of a differential inclusion (sometimes called weak invariance and strong invariance) are essential in the proof of Theorem 2.2. It will be seen that the epigraph of the value function is both viable and invariant with respect to a certain *unbounded* differential inclusion. The italics are meant to stress that this is an uncommon assumption. If one were to examine the main theorems on differential inclusions in texts such as [1], [3], [9], or [15], it becomes immediately apparent that set-valued mappings with compact images are the focus.

To study the viability properties of $\operatorname{epi} V$ it will be essential for us to know that this is a closed set and that solutions to the unbounded differential inclusion exist. This leads us to the following.

PROPOSITION 3.1. *Assume we have φ satisfying* (A0) *and a Lagrangian L which is lsc in all variables and that $L(t, x, \cdot\,)$ is proper and convex for each $(t, x) \in \mathbb{R} \times \mathbb{R}^n$. Further assume that* (A1) *holds for H derived from L via* (1.3). *Then the value function V for $(\mathcal{P}_{\tau,\xi})$ has the following properties.*

(a) *V is proper and lower semicontinuous on $\mathbb{R} \times \mathbb{R}^n$.*

(b) *At every $(\tau, \xi) \in \operatorname{dom} V$, there exists an optimal arc achieving the value $V(\tau, \xi)$ in $(\mathcal{P}_{\tau,\xi})$.*

Proof. Take (τ, ξ) with $V(\tau, \xi) < +\infty$ and consider the functional

$$x(\,\cdot\,) \mapsto l(x(0), x(\tau)) + \int_0^\tau L(t, x(t), \dot{x}(t))dt,$$

where $l(x(0), x(\tau)) = \varphi(x(0)) + \Psi_{\{\xi\}}(x(\tau))$. Minimizing this functional over all $x(\cdot) \in \mathcal{A}_n[0, \tau]$ is equivalent to $(\mathcal{P}_{\tau,\xi})$. The hypotheses given allow us to use "Existence Theorem 2" of [25] to conclude that the above functional attains its minimum. Thus there exists some absolutely continuous arc $x(\,\cdot\,)$ achieving the minimum in $(\mathcal{P}_{\tau,\xi})$, and (b) holds.

To see that the minimum value is not $-\infty$, note that $L(t,x,v) \geq -H(t,x,0)$ for all v, and so if x is our optimal arc, there exists a constant M with $|x(t)| \leq M$ for all $t \in [0,\tau]$, giving us

$$\begin{aligned} V(\tau,\xi) &\geq \varphi(x(0)) - \int_0^\tau H(t,x(t),0)dt \\ &\geq \varphi(x(0)) - \int_0^\tau \mu(0) + \sigma(|t| + |x(t)|)dt \\ &\geq \varphi(x(0)) - \int_0^\tau \mu(0)dt - \int_0^\tau \sigma(\tau + M)dt \\ &> -\infty. \end{aligned}$$

That V is not identically $+\infty$ follows since $V(0,\xi) = \varphi(\xi)$ and φ is proper. Thus V is a proper function.

It remains to show that V is lsc. Let $K = [a,b] \times \widetilde{K} \subset \mathbb{R} \times \mathbb{R}^n$ be a compact set with $0 \leq a < b$. For each $\tau \geq 0$ and $x(\,\cdot\,) \in \mathcal{A}_n^1[0,\tau]$, let

$$\gamma_K(\tau, x(\,\cdot\,)) = \varphi(x(0)) + \Psi_K(\tau, x(\tau)) + \int_0^\tau L(t,x(t),\dot{x}(t))dt.$$

For each $\alpha \in \mathbb{R}$, define the set $A_K^\alpha \subset \mathbb{R} \times \mathbb{R}^n$ as

$$A_K^\alpha := \{(\tau, x(\tau)) \mid \gamma_K(\tau, x(\,\cdot\,)) \leq \alpha \quad \text{for some} \quad (\tau, x(\,\cdot\,))\}.$$

These sets are related to the lower level sets of V as follows:

$$A_K^\alpha = K \cap \operatorname{lev}_{\leq\alpha} V,$$

where the notation $\operatorname{lev}_{\leq\alpha} f$ denotes the set $\{\, x \mid f(x) \leq \alpha \,\}$. So if for every compact K, A_K^α is compact for each α, this shows the lower level sets of V are closed, and V is lsc. To accomplish this, we will use the Erdmann transform to convert the above problem into one with fixed time. Let

$$L_0(t,x,v,\lambda) = \begin{cases} \Psi_{\{0\}}(v) & \text{if} \quad \lambda = 0, \\ \lambda L(t,x,v/\lambda) & \text{if} \quad \lambda > 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that since $L(t,x,\,\cdot\,)$ is proper, convex, and coercive, $L_0(t,x,\,\cdot\,,\,\cdot\,)$ will be proper, convex, and lsc. For arcs $(\theta(\,\cdot\,), \zeta(\,\cdot\,)) \in \mathcal{A}_{n+1}^1[0,1]$ consider the functional

$$\Gamma_K(\theta(\,\cdot\,), \zeta(\,\cdot\,)) = l(\theta(0), \zeta(0), \theta(1), \zeta(1)) + \int_0^1 L_K(\theta(s), \zeta(s), \dot{\zeta}(s), \dot{\theta}(s))ds,$$

where we define $l(\theta(0), \zeta(0), \theta(1), \zeta(1)) = \Psi_{\{0\}}(\theta(0)) + \varphi(\zeta(0)) + \Psi_K(\theta(1), \zeta(1))$ and $L_K(t,x,v,\lambda) = L_0(t,x,v,\lambda) + \Psi_{[a,b]}(\lambda)$. Now consider the sets

$$B_K^\alpha := \{(\theta(1), \zeta(1)) \mid \Gamma_K(\theta(\,\cdot\,), \zeta(\,\cdot\,)) \leq \alpha \text{ for some } (\theta(\,\cdot\,), \zeta(\,\cdot\,))\}.$$

The function $\Psi_{\{0\}}(\theta(0)) + \varphi(\zeta(0))$ is not countercoercive on $\mathbb{R}^{n+1}$, while $\Psi_K(\,\cdot\,,\,\cdot\,)$ is coercive, since K is compact. Furthermore, the Hamiltonian corresponding to

L_K satisfies (A1), the stronger growth condition, and we can apply the results of "Existence Theorem 2" of [25] to conclude that the level sets

$$\{(\theta(\,\cdot\,),\zeta(\,\cdot\,)) \mid \Gamma_K(\theta(\,\cdot\,),\zeta(\,\cdot\,)) \le \alpha\}$$

are compact in the norm topology of continuous arcs from $[0,1]$ to $\mathbb{R}^{n+1}$, where the norm is the usual supremum norm. Thus B_K^α must be compact.

Now we need to relate B_K^α to A_K^α. Fix α and K, and take $(\tau, x(\tau)) \in A_K^\alpha$. Let $\theta(s) = \tau s$ and $\zeta(s) = x(\tau s)$. Note that $\tau \in [a,b]$, and $\Gamma_K(\theta(\cdot),\zeta(\cdot)) = \gamma_K(\tau, x(\cdot)) \le \alpha$. Also, $(\theta(1),\zeta(1)) = (\tau, x(\tau))$, so we have $A_K^\alpha \subset B_K^\alpha$.

If we take $K = [a,b] \times \widetilde{K}$ with $a > 0$ and consider an arc for which $\Gamma_K(\theta(\cdot),\zeta(\cdot)) \le \alpha$, then necessarily $\dot{\theta}(s) \ge a > 0$ for a.e. $s \in [0,1]$ and θ is then strictly increasing and invertible. If we then let $\tau = \theta(1)$ and $x(t) = \zeta(\theta^{-1}(t))$, substituting these arcs into the functionals above, we find $\gamma_K(\tau, x(\,\cdot\,)) = \Gamma_K(\theta(\,\cdot\,),\zeta(\,\cdot\,)) \le \alpha$, and $(\tau, x(\tau)) = (\theta(1),\zeta(1))$, showing that $B_K^\alpha \subset A_K^\alpha$. Combining this with the opposite inclusion from the previous paragraph, we see that A_K^α is compact for every α and every compact set $K = [a,b] \times \widetilde{K}$ when $a > 0$. This implies that V is lsc at every (τ,ξ) with $\tau > 0$.

To see that V is also lsc at points $(0,\xi)$, consider compact $K = [0,b] \times \widetilde{K}$, and note that if $(0,\xi) = (\theta(1),\zeta(1)) \in B_K^\alpha$, then $\dot{\theta}(s) = 0$ for a.e. s, which forces $\dot{\zeta}(s) = 0$ for a.e. s, from the way L_0 is defined. Thus $\zeta(1) = \zeta(0)$ and $\Gamma_K(\theta(\cdot),\zeta(\cdot)) = \varphi(\zeta(0))$, so $\zeta(1) \in \operatorname{lev}_{\le\alpha}\varphi$. On the other hand, if we take any $\xi \in \widetilde{K} \cap \operatorname{lev}_{\le\alpha}\varphi$, the constant arc $(\theta(s),\zeta(s)) = (0,\xi)$ gives $\Gamma_K(\theta(\,\cdot\,),\zeta(\,\cdot\,)) = \varphi(\xi) \le \alpha$, so we have shown that

$$B_K^\alpha \cap (\{0\} \times \mathbb{R}^n) = \{0\} \times (\widetilde{K} \cap \operatorname{lev}_{\le\alpha}\varphi).$$

We also know from above that B_K^α is compact and that $A_K^\alpha \subset B_K^\alpha$. So take $\xi \in \operatorname{dom}\varphi$ and $\widetilde{K}$ so that ξ lies in the interior of $\widetilde{K}$. Take $b > 0$ and any sequence $(\tau^\nu,\xi^\nu) \to (0,\xi)$ with $(\tau^\nu,\xi^\nu) \in A_K^\alpha$. So in fact (τ^ν,ξ^ν) must lie in B_K^α, whose compactness forces $(0,\xi) \in B_K^\alpha$, and so from (4.1) we have $\varphi(\xi) \le \alpha$. It must be the case then that $\liminf_{(\tau^\nu,\xi^\nu)\to(0,\xi)} V(\tau^\nu,\xi^\nu) \ge \varphi(\xi) = V(0,\xi)$, and so V is lsc at every point in its effective domain. □

At this point we introduce the set-valued mapping $\widetilde{E}_L : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \rightrightarrows \mathbb{R}^{n+2}$ as

$$\widetilde{E}_L(t,x,y) := \ \{1\} \times \operatorname{epi} L(t,x,\,\cdot\,).$$

The variable $y \in \mathbb{R}$ can be thought of as an "epigraphical variable" since we will typically be considering the setting where the points $(t,x,y) \in \operatorname{epi} V$. We wish to establish that the value function is the unique function which satisfies an endpoint condition and is simultaneously viable and invariant in a certain way with respect to $\widetilde{E}_L$.

PROPOSITION 3.2. *Assume* (A) *and* (A0)–(A2). *A function* $u : \mathbb{R} \times \mathbb{R}^n \to \overline{\mathbb{R}}$ *is the value function for* $(\mathcal{P}_{\tau,\xi})$ *if and only if it satisfies the following properties.*

(P1) *u is proper and lsc on* $\mathbb{R} \times \mathbb{R}^n$ *with* $\operatorname{dom} u \subset [0,\infty) \times \mathbb{R}^n$.

(P2) $u(0,\xi) = \varphi(\xi)$ *for all* $\xi \in \mathbb{R}^n$.

(P3) *(backward viability). For all* $(\tau,\xi,\beta) \in \operatorname{epi} u$, *there exists an absolutely continuous arc* $z(\,\cdot\,) = (t(\,\cdot\,), x(\,\cdot\,), y(\,\cdot\,))$ *such that*

(i) $z(\,\cdot\,)$ *is defined on* $[0,\tau]$, (ii) $z(0) = (\tau,\xi,\beta)$, (iii) $\dot{z}(s) \in -\widetilde{E}_L(z(s))$ *for a.e.* $s \in [0,\tau]$, *and* (iv) $z(s) \in \operatorname{epi} u$ *for all* $s \in [0,\tau]$.

(P4) *(forward invariance). For all* $(\tau,\xi,\beta) \in \operatorname{epi} u$, *for all* $r > 0$, *if an absolutely continuous arc* $z(\,\cdot\,) = (t(\,\cdot\,), x(\,\cdot\,), y(\,\cdot\,))$ *satisfies*

(i) $z(\,\cdot\,)$ *is defined on* $[0,r]$*,* (ii) $z(0)=(\tau,\xi,\beta)$*, and* (iii) $\dot z(s)\in \widetilde{E}_L(z(s))$ *for a.e.* $s\in[0,r]$*, then* $z(\,\cdot\,)$ *must also satisfy* (iv) $z(s)\in \operatorname{epi} u$ *for all* $s\in[0,r]$*.*

Proof. Take the value function V. By definition of V, $\operatorname{dom} V\subset[0,\infty)\times\mathbb{R}^n$, and also (P2) holds. Proposition 3.1(a) guarantees that the rest of condition (P1) holds. To show (P3), take $(\tau,\xi,\beta)\in\operatorname{epi}V$. Let $\bar x(\,\cdot\,)$ be an optimal arc for $(\mathcal{P}_{\tau,\xi})$, its existence guaranteed by Proposition 3.1(b). By the principle of optimality, we have $V(\tau',\bar x(\tau'))=\varphi(\bar x(0))+\int_0^{\tau'}L(t,\bar x(t),\dot{\bar x}(t))dt$ for all $\tau'\in[0,\tau]$. Thus the arc

$$z(s)=\left(\tau-s,\ \bar x(\tau-s),\ \varphi(\bar x(0))+\int_0^{\tau-s}L(t,\bar x(t),\dot{\bar x}(t))\,dt+\beta-V(\tau,\xi)\right)$$

satisfies (P3) since

$$\begin{aligned}\varphi(\bar x(0))+\int_0^{\tau-s}L(t,\bar x(t),\dot{\bar x}(t))dt+\beta-V(\tau,\xi)&=V(\tau-s,\bar x(\tau-s))+(\beta-V(\tau,\xi))\\&\geq V(\tau-s,\bar x(\tau-s))\end{aligned}$$

for all $s\in[0,\tau]$, and so $z(s)\in\operatorname{epi}V$ for all s. Also, we see that

$$\dot z(s)=(-1,-\dot{\bar x}(\tau-s),-L(\tau-s,\bar x(\tau-s),\dot{\bar x}(\tau-s)))\in-\widetilde{E}_L(z(s)).$$

Assume we have an arc $z(\,\cdot\,)$ satisfying (i)–(iii) of (P4). Then for any $\tau'\in[0,r]$, we have

$$\begin{aligned}y(\tau')=\beta+\int_0^{\tau'}\dot y(s)ds&\geq\beta+\int_0^{\tau'}L(\tau+s,x(s),\dot x(s))ds\\&\geq\beta+V(\tau'+\tau,x(\tau'))-V(\tau,x(\tau))\\&\geq V(\tau'+\tau,x(\tau')),\end{aligned}$$

and thus $z(s)\in\operatorname{epi}V$ for $s\in[0,r]$.

Now assume u satisfies (P1)–(P4). Clearly $u(\tau,\xi)=V(\tau,\xi)$ if $\tau\leq0$. So fix (τ,ξ) with $\tau>0$. Let $\beta=u(\tau,\xi)$. By (P3) there exists an arc $(t(\,\cdot\,),x(\,\cdot\,),y(\,\cdot\,))$ with $(t(0),x(0),y(0))=(\tau,\xi,\beta)$ and $(t(\tau),x(\tau),y(\tau))\in\operatorname{epi}u$. But the differential inclusion that our arc satisfies means $\dot t(s)=-1$ and so $t(s)=\tau-s$ and $t(\tau)=0$. Thus $y(\tau)\geq\varphi(x(\tau))$. Now if we imagine running our arc backward via the new arc $(\bar t(s),\bar x(s),\bar y(s))=(t(\tau-s),x(\tau-s),y(\tau-s))$, we get that $\bar t(s)=s$ and $\dot{\bar y}(s)\geq L(s,\bar x(s),\dot{\bar x}(s))$ and so

$$\begin{aligned}\beta=\bar y(0)+\int_0^{\tau}\dot{\bar y}(s)ds&\geq y(\tau)+\int_0^{\tau}L(s,\bar x(s),\dot{\bar x}(s))ds\\&\geq\varphi(\bar x(0))+\int_0^{\tau}L(s,\bar x(s),\dot{\bar x}(s))ds\\&\geq V(\tau,\xi),\end{aligned}$$

showing that $u(\tau,\xi)\geq V(\tau,\xi)$.

Now using (P4), let $\bar x(\,\cdot\,)$ be a minimizing arc giving the value of $V(\tau,\xi)$. We have then $(0,\bar x(0),\varphi(\bar x(0)))\in\operatorname{epi}u$ by (P2). Let $\bar t(s)=s$ and $\bar y(s)=\varphi(\bar x(0))+\int_0^s L(w,\bar x(w),\dot{\bar x}(w))dw$. Then the arc $z(\,\cdot\,)=(\bar t(\,\cdot\,),\bar x(\,\cdot\,),\bar y(\,\cdot\,))$ satisfies (i)–(iii) of (P4). Set $r=\tau$ and so $z(\tau)\in\operatorname{epi}u$. But since $\bar y(\tau)=V(\tau,\xi)$, we must have that $u(\tau,\xi)\leq V(\tau,\xi)$. □

4. Monotonicity and the reachable set. In the previous section it was seen that the value function uniquely satisfies certain invariance and viability properties. The goal now is to characterize these properties with the monotone behavior of a certain *reachable set.*

First consider a general setup as follows. We have a closed set $K \subset \mathbb{R}^n$ and a set valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. Under the differential inclusion $\dot{x}(t) \in F(x(t))$, we can define the *forward solution set* $\mathcal{S}_K^+(t)$ and the *backward solution set* $\mathcal{S}_K^-(t)$ to be

$$\mathcal{S}_K^+(t) := \{x(\,\cdot\,) \in \mathcal{A}_n^1([0,t]) \mid x(0) \in K,\ \dot{x}(s) \in F(x(s)) \ \text{ for a.e. } \ s \in [0,t]\},$$
$$\mathcal{S}_K^-(t) := \{x(\,\cdot\,) \in \mathcal{A}_n^1([0,t]) \mid x(0) \in K,\ \dot{x}(s) \in -F(x(s)) \ \text{ for a.e. } \ s \in [0,t]\},$$

and the corresponding reachable set $\mathcal{R}_K(t)$ as

$$\mathcal{R}_K(t) := \{y \in \mathbb{R}^n \mid y = x(t) \ \text{for some} \ x(\,\cdot\,) \in \mathcal{S}_K^+(t)\}.$$

Note that we are not requiring that solutions exist for all times. It is possible that $\mathcal{R}_K(t) = \emptyset$ for some t.

Proposition 4.1. *The following are equivalent.*

(i) *The set K is invariant with respect to the forward solution set. That is, for all $t \geq 0$, every $x(\,\cdot\,) \in \mathcal{S}_K^+(t)$ satisfies $x(s) \in K$ for all $s \in [0,t]$.*

(ii) *For any t_1, t_2 with $t_2 \geq t_1 \geq 0$, we have $\mathcal{R}_K(t_2) \subset \mathcal{R}_K(t_1)$.*

Proof. It should be clear that (ii) implies (i). Assume (i) and take $t_2 \geq t_1 \geq 0$. The set K being invariant tells us that $\mathcal{R}_K(t) \subset K$ for all $t > 0$. It is also true that if $A \subset B$, then $\mathcal{R}_A(t) \subset \mathcal{R}_B(t)$. By its nature, the reachable set satisfies a certain semigroup property which tells us that

$$\mathcal{R}_K(t_2) = \mathcal{R}_{\mathcal{R}_K(t_2-t_1)}(t_1).$$

But $\mathcal{R}_K(t_2 - t_1) \subset K$, and so the right-hand side of the above is contained in the set $\mathcal{R}_K(t_1)$. □

The concept of viability is related to that of invariance. To say that the set K is viable with respect to the backward solution set means for every $t \geq 0$ and for every $x_0 \in K$, there exists an $x(\,\cdot\,) \in \mathcal{S}_K^-(t)$ such that $x(0) = x_0$ and $x(s) \in K$ for all $s \in [0,t]$.

Proposition 4.2. *The following are equivalent.*

(i) *The set K is invariant with respect to the forward solution set and viable with respect to the backward solution set.*

(ii) *$\mathcal{R}_K(t) = K$ for all $t \geq 0$.*

Proof. Assume (i) and take $t_1 \geq 0$. From Proposition 4.1 we know that $\mathcal{R}_K(t) \subset K$. Take any $x_0 \in K$. That K is viable means there is an arc $x(\,\cdot\,)$ with $\dot{x}(t) \in -F(x(t))$, $x(0) = x_0$, and $x(t) \in K$ for all t. Consider the arc $y(t) = x(t_1 - t)$. Then $y \in \mathcal{S}_K^+(t_1)$ with $y(t_1) = x_0$. So $x_0 \in \mathcal{R}_K(t_1)$ giving us that $K \subset \mathcal{R}_K(t_1)$.

If we assume (ii), Proposition 4.1 tells us that K is forward invariant. Let $x_0 \in K$. Then $x_0 \in \mathcal{R}_K(t_1)$ for any $t_1 > 0$. Thus there exists an arc $x(\,\cdot\,) \in \mathcal{S}_K^+(t_1)$ with $x(t_1) = x_0$ and $x(t) \in K$ for $t \in [0, t_1]$. Again, if we consider $y(t) = x(t_1 - t)$, we get $y \in \mathcal{S}_K^-(t_1)$ with $y(0) = x_0$, and so K is viable with respect to the backward solution set. □

Note that we require minimal hypotheses on the set K and the mapping F. In particular, F can have unbounded images.

Now we wish to use this monotonicity property in the setting of the previous section. Proposition 4.2 cannot be applied directly, however, since the time interval on which the epigraph of the value function is viable depends upon the initial point we choose, and thus the general results above should be viewed as a motivating guide for the following proposition.

Proposition 4.3. *Take the set-valued mapping* $\widetilde{E}_L(t,x,y) = \{1\} \times \operatorname{epi} L(t,x,\,\cdot\,)$ *and the corresponding differential inclusion* $\dot{z}(s) \in \widetilde{E}_L(z(s))$. *Assume* $u : \mathbb{R}\times\mathbb{R}^n \to \overline{\mathbb{R}}$ *satisfies* (P1)–(P2) *of Proposition* 3.2. *Then* u *also satisfies* (P3) *and* (P4) *if and only if*

$$\mathcal{R}_{\operatorname{epi} u}(s) = \operatorname{epi} u \cap ([s,\infty)\times\mathbb{R}^n) \quad \textit{for all } s \geq 0.$$

Proof. Assume u satisfies (P1)–(P4). The condition (P4) says that $\operatorname{epi} u$ is invariant with respect to the forward solution set of the differential inclusion $\dot{z}(s) \in \widetilde{E}_L(z(s))$. Thus, by Proposition 4.1 we have $\mathcal{R}_{\operatorname{epi} u}(s) \subset \operatorname{epi} u$. Also, any $z(\,\cdot\,) = (t(\,\cdot\,), x(\,\cdot\,), y(\,\cdot\,))$ satisfying the differential inclusion must have $\dot{t}(s) = 1$ for a.e. s and $t(0) \geq 0$, so $t(s) \geq s$. Thus $\mathcal{R}_{\operatorname{epi} u}(s) \subset \operatorname{epi} u \cap ([s,\infty)\times\mathbb{R}^n)$.

Take $(\tau,\xi,\beta) \in \operatorname{epi} u \cap ([s,\infty)\times\mathbb{R}^n)$. This is equivalent to taking $(\tau,\xi,\beta) \in \operatorname{epi} u$ with $\tau \geq s$. Now take an arc $(t(\cdot), x(\cdot), y(\cdot))$ as in (P3). Since $s \in [0,\tau]$, $z(s) \in \operatorname{epi} u$. If we now consider the arc $\bar{z}(s') = z(s-s')$, we see that $\bar{z}(0) \in \operatorname{epi} u$, $\bar{z}(s) = (\tau,\xi,\beta)$, and $\dot{\bar{z}}(s') \in \widetilde{E}_L(\bar{z}(s'))$. So $(\tau,\xi,\beta) \in \mathcal{R}_{\operatorname{epi} u}(s)$ and we have $\operatorname{epi} u \cap ([s,\infty)\times\mathbb{R}^n) \subset \mathcal{R}_{\operatorname{epi} u}(s)$.

Now take a function u satisfying (P1)–(P2) and such that for all $s \geq 0$, $\mathcal{R}_{\operatorname{epi} u}(s) = \operatorname{epi} u \cap ([s,\infty)\times\mathbb{R}^n)$. Then Proposition 4.1 immediately implies that (P4) holds.

To see that (P3) holds, take $(\tau,\xi,\beta) \in \operatorname{epi} u$. So in fact $(\tau,\xi,\beta) \in \mathcal{R}_{\operatorname{epi} u}(\tau)$, and there exists a solution to the differential inclusion, $\bar{z}(\,\cdot\,) = (\bar{t}(\,\cdot\,), \bar{x}(\,\cdot\,), \bar{y}(\,\cdot\,))$ with $(\bar{t}(0), \bar{x}(0), \bar{y}(0)) \in \operatorname{epi} u$ and $(\bar{t}(\tau), \bar{x}(\tau), \bar{y}(\tau)) = (\tau,\xi,\beta)$. Again, by letting $z(s') = \bar{z}(\tau - s')$, we get an arc satisfying (i)–(iv) of (P3). □

Note that the set-valued mapping $\widetilde{E}_L(t,x,y)$ is independent of y. Thus, if $(\tau,\xi,\beta) \in \mathcal{R}_{\operatorname{epi} u}(s)$ for some s, then in fact $(\tau,\xi,\beta') \in \mathcal{R}_{\operatorname{epi} u}(s)$ for all $\beta' \geq \beta$. This means that the graph of $\mathcal{R}_{\operatorname{epi} u}$ can be thought of as the epigraph of a function whose domain lies in $\mathbb{R}\times\mathbb{R}\times\mathbb{R}^n$. Given $u : \mathbb{R}\times\mathbb{R}^n \to \overline{\mathbb{R}}$, define this associated function $\widetilde{u} : \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n \to \overline{\mathbb{R}}$ as follows:

$$\widetilde{u}(s,\tau,\xi) = \begin{cases} u(\tau,\xi), & 0 \leq s \leq \tau, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.1}$$

In other words, $\operatorname{epi} \widetilde{u}(s,\,\cdot\,,\,\cdot\,) = \operatorname{epi} u \cap ([s,\infty)\times\mathbb{R}^n)$ for $s \geq 0$. At this point we show that the function $\widetilde{u}$ can be characterized with subgradients.

Proposition 4.4. *Assume we have proper, lsc functions* $f : \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n \to \overline{\mathbb{R}}$ *and* $u : \mathbb{R}\times\mathbb{R}^n \to \overline{\mathbb{R}}$. *Then* (i) *and* (ii) *are equivalent.*

(i) $f = \widetilde{u}$, *as defined by* (4.1).

(ii) $f(0,\,\cdot\,,\,\cdot\,) = u(\,\cdot\,,\,\cdot\,)$, *and at every* $(s,\tau,\xi) \in \operatorname{dom} f$, *every subgradient* $(\alpha,\sigma,\eta) \in \partial f(s,\tau,\xi)$ *satisfies*

$$\begin{cases} \alpha \leq 0 & \textit{if } \; 0 = s < \tau, \\ \alpha = 0 & \textit{if } \; 0 < s < \tau, \\ \alpha \geq 0 & \textit{if } \; 0 < s = \tau. \end{cases}$$

The proof of Proposition 4.4 depends on the following lemma (Proposition 8.50 in [27]).

Lemma 4.5. *For an lsc function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $\bar{x}$ where f is finite, the following are equivalent for a pair $(a, \beta) \in \mathbb{R}^n \times \mathbb{R}$.*

(i) *There exists a neighborhood W of $\bar{x}$ and $\delta > 0$ such that $\langle v, a\rangle \le \beta$ for all $v \in \partial f(x)$ when $x \in W$ and $f(x) \le f(\bar{x}) + \delta$.*

(ii) *For some neighborhood W of $\bar{x}$, $\delta > 0$, and $\varepsilon > 0$ one has*

$$\frac{f(x+ka) - f(x)}{k} \le \beta \text{ when } k \in (0, \varepsilon],\ x \in W, f(x) \le f(\bar{x}) + \delta.$$

Proof of Proposition 4.4. First we show that $\widetilde{u}$ satisfies (ii). The first condition is clearly satisfied, so we only need check that the subgradient criterion holds. First take $(0, \bar{\tau}, \bar{\xi}) \in \operatorname{dom} \widetilde{u}$ with $\bar{\tau} > 0$. Then there exists $\varepsilon > 0$ such that $(0, \bar{\tau}, \bar{\xi}) + 2\varepsilon\mathbb{B} \subset \{(s, \tau, \xi) | s < \tau\}$. Let $W = (0, \bar{\tau}, \bar{\xi}) + \varepsilon\mathbb{B}$ and let $\delta > 0$ be arbitrary. Fix $a = (1, 0, 0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$. Then if we take any $k \in (0, \varepsilon]$ and $(s, \tau, \xi) \in W$ with $\widetilde{u}(s, \tau, \xi) \neq \infty$, we have

$$\begin{aligned}(\widetilde{u}((s,\tau,\xi) + ka) - \widetilde{u}(s,\tau,\xi))/k &= (\widetilde{u}(s+k,\tau,\xi) - \widetilde{u}(s,\tau,\xi))/k \\ &= (u(\tau,\xi) - u(\tau,\xi))/k \\ &= 0.\end{aligned}$$

So (ii) of Lemma 4.5 holds with $\beta = 0$. Thus (i) of Lemma 4.5 tells us that, in particular, $\langle(\alpha, \sigma, \eta), a\rangle \le 0$ for all $(\alpha, \sigma, \eta) \in \partial\widetilde{u}(0, \bar{\tau}, \bar{\xi})$, which reduces to saying that $\alpha \le 0$.

Now take $(\bar{s}, \bar{\tau}, \bar{\xi}) \in \operatorname{dom} \widetilde{u}$ with $0 < \bar{s} < \bar{\tau}$. In this case, the same argument as above will go through, applied to both a and $-a$. Applying part (i) of Lemma 4.5 now gives the results that $\alpha \le 0$ and $\alpha \ge 0$ for any $(\alpha, \sigma, \eta) \in \partial\widetilde{u}(\bar{s}, \bar{\tau}, \bar{\xi})$. So $\alpha = 0$.

Similarly, taking $(\bar{s}, \bar{\tau}, \bar{\xi}) \in \operatorname{dom} \widetilde{u}$ with $0 < \bar{s} = \bar{\tau}$ allows us to apply the above argument with $-a$, giving the result that $\alpha \ge 0$ for any $(\alpha, \sigma, \eta) \in \partial\widetilde{u}(\bar{s}, \bar{s}, \bar{\xi})$.

Assume now we have a proper, lsc f satisfying (ii). Then at any $(s, \tau, \xi) \in \operatorname{dom} f$ with $0 \le s < \tau$, we can find a neighborhood W so that (i) of the lemma holds for $a = (1, 0, 0)$ and $\beta = 0$ (δ arbitrary). Similarly, (i) of the lemma holds for $a = (-1, 0, 0)$ and $\beta = 0$ in a neighborhood of (s, τ, ξ) with $0 < s \le \tau$. Now applying (ii) of the lemma we see that, for fixed $(\bar{\tau}, \bar{\xi})$, $f(s, \bar{\tau}, \bar{\xi})$ is monotonically decreasing for $0 \le s < \bar{\tau}$ and is monotonically increasing for $0 < s \le \bar{\tau}$. Thus $f(s, \bar{\tau}, \bar{\xi}) = c$, a constant, for $s \in (0, \bar{\tau})$.

Now by assumption, $f(0, \bar{\tau}, \bar{\xi}) = u(\bar{\tau}, \bar{\xi})$. Since f is decreasing in s at $s = 0$, we must have $c \le u(\bar{\tau}, \bar{\xi})$. But f is lsc, so $f(0, \bar{\tau}, \bar{\xi}) \le c$. So in fact $f(s, \bar{\tau}, \bar{\xi}) = u(\bar{\tau}, \bar{\xi})$ for $s \in [0, \bar{\tau})$. But again, using lower semicontinuity and monotonicity of f in s, we get $f(\bar{\tau}, \bar{\tau}, \bar{\xi}) = c = u(\bar{\tau}, \bar{\xi})$. The only points not checked in $\operatorname{dom} f$ are of the form $(0, 0, \xi)$, but by assumption we know that $f(0, 0, \xi) = u(0, \xi)$. It should be clear from the argument just given that $(s, \tau, \xi) \in \operatorname{dom} f$ if and only if $0 \le s \le \tau$ and $(\tau, \xi) \in \operatorname{dom} u$. Thus $\operatorname{dom} f = \operatorname{dom} \widetilde{u}$ and so $f = \widetilde{u}$. □

5. Necessary conditions. This section examines the necessary conditions presented in the papers [21], [22], and [23]. Some work is required to show that our assumptions (A) and (A0)–(A2) enable us to use these papers. The approach taken in that series of articles was to take an arc known to be optimal and then impose hypotheses with respect to that particular arc. The goal of this section then is to show that under our assumptions, when the need arises to examine necessary conditions for some optimal arc, we can find a neighborhood of that arc to which Loewen and Rockafellar's hypotheses can be applied.

The main result of the paper [23] followed directly from [22], after applying the Erdmann transform to place the variable time Bolza problem in the context of a fixed time problem. This transform employs the following integrand which we will denote by L_m. For each $m > 0$, define

$$L_m(t,x,v,\lambda) = \begin{cases} \lambda L(t,x,v/\lambda) & \text{if} \quad \lambda \geq m, \\ +\infty & \text{otherwise.} \end{cases} \tag{5.1}$$

Note that $L_m(t,x,\,\cdot\,,\,\cdot\,)$ is proper and convex if and only if $L(t,x,\,\cdot\,)$ is proper and convex.

Proposition 5.1. *Assume $L(t,x,\,\cdot\,)$ is proper, lsc, coercive, and convex for each (t,x), and that* (A2) *holds. Then for each $m > 0$ and every $(\bar{t},\bar{x})$, there exists a neighborhood W of $(\bar{t},\bar{x})$ and positive constant k such that for all $(t,x,v,\lambda) \in W \times \mathbb{R}^n \times \mathbb{R}$,*

$$|(w_1,w_2)| \leq k(1 + |(v,\lambda)| + |L_m(t,x,v,\lambda)|)(1 + |(p,\pi)|) \tag{5.2}$$

for all $(w_1,w_2,p,\pi) \in \partial L_m(t,x,v,\lambda)$.

Proof. In Proposition 2.6 it was shown that (A2) implies that the subgradients of L satisfy a relation given by (2.3). Using the calculus rules ([27, Thm. 10.6 and Cor. 10.9]), and the fact that we can write $L_m(t,x,v,\lambda) = \lambda L(t,x,v/\lambda) + \Psi_{[m,\infty)}(\lambda)$, we have

$$\begin{aligned} &\partial L_m(t,x,v,\lambda) \subset \{(\lambda w_1, \lambda w_2, p, r) \mid (w_1,w_2,p) \in \partial L(t,x,v/\lambda)\}, \\ &\widehat{\partial} L_m(t,x,v,\lambda) \supset \{(\lambda w_1, \lambda w_2, p, r) \mid (w_1,w_2,p) \in \widehat{\partial} L(t,x,v/\lambda)\} \\ &\qquad \text{with } \begin{cases} r = -H(t,x,p) & \text{if } \lambda > m, \\ r \leq -H(t,x,p) & \text{if } \lambda = m. \end{cases} \end{aligned} \tag{5.3}$$

Take $m > 0$. Since L is lsc, $\operatorname{epi} L_m(\cdot,\cdot,\cdot,\lambda)$ is a closed set and from the definition of L_m is varying continuously with λ for $\lambda \geq m$. So $\operatorname{epi} L_m$ is closed, implying L_m is lsc.

Take $(\bar{t},\bar{x}) \in \mathbb{R} \times \mathbb{R}^n$ and choose a neighborhood W and constant k as given by (A2). Also let k_1 be a positive constant so that $H(t,x,p) \geq -k_1(1 + |p|)$ whenever $(t,x,p) \in W \times \mathbb{R}^n$. The existence of this constant is guaranteed by Lemma 2.5. Let $(w_1,w_2,p,\pi) \in \partial L_m(t,x,v,\lambda)$ for some $(t,x,v,\lambda) \in W \times \mathbb{R}^n \times [m,\infty)$. Then (5.3) tells us that $(w_1,w_2,p,\pi) = (\lambda w_1', \lambda w_2', p, \pi)$ for some $(w_1',w_2',p) \in \partial L(t,x,v/\lambda)$ and with $\pi \leq -H(t,x,p)$. Note that we have the estimate $|H(t,x,p)| \leq \max\{k_1(1 + |p|), |\pi|\}$. So (A2) says that

$$\begin{aligned} |(w_1',w_2')| &\leq k(1 + |v/\lambda| + |L(t,x,v/\lambda)|)(1 + |p| + |H(t,x,p)|) \\ \Rightarrow |(w_1,w_2)| &\leq k(\lambda + |v| + |L_m(t,x,v,\lambda)|)(1 + |p| + |H(t,x,p)|) \\ &\leq \sqrt{2}k(1 + |(v,\lambda)| + |L_m(t,x,v,\lambda)|)(1 + |p| + k_1(1{+}|p|{+}|\pi|)) \\ &\leq \sqrt{2}k(k_1 + 1)(1 + |(v,\lambda)| + |L_m(t,x,v,\lambda)|)(1 + |p| + |\pi|) \\ &\leq 2k(k_1 + 1)(1 + |(v,\lambda)| + |L_m(t,x,v,\lambda)|)(1 + |(p,\pi)|). \qquad \square \end{aligned}$$

Lemma 5.2. *Assume we have an autonomous Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ for which, given any $M > 0$, there exists $k_1 > 0$ such that for all $(x,v) \in M\mathbb{B} \times \mathbb{R}^n$,*

$$|w| \leq k_1(1 + |v| + |L(x,v)|)(1 + |p|) \quad \text{for all} \quad (w,p) \in \partial L(x,v). \tag{5.4}$$

Let $\bar{x}(\cdot) \in \mathcal{A}_n^1([a,b])$ so that $L(\bar{x}(t), \dot{\bar{x}}(t))$ is integrable on $[a,b]$. Then there exist $\varepsilon > 0$ and integrable $\kappa : \mathbb{R} \to \mathbb{R}$ and $\delta : \mathbb{R} \to \mathbb{R}$ so that ([22, Hypothesis H5]) *holds. That is, the ratio $\kappa(t)/\delta(t)$ is essentially bounded and for a.e. $t \in [a,b]$, one has*

$$|w| \le \kappa(t)(1+|p|) \qquad \text{for all } (w,p) \in \partial L(x,v)$$

whenever $|x - \bar{x}(t)| < \varepsilon,\ \big|\,(v, L(x,v)) - (\dot{\bar{x}}(t), L(\bar{x}(t), \dot{\bar{x}}(t)))\,\big| < \delta(t)$.

Proof. Let $M = (\sup\{|\bar{x}(t)| : t \in [a,b]\} + 1)$ and let $\varepsilon < 1$. Define

$$\begin{aligned}
&\bar{L}(t) = L(\bar{x}(t), \dot{\bar{x}}(t)),\\
&\delta(t) = |\dot{\bar{x}}(t)| + |\bar{L}(t)| + 1,\\
&\kappa(t) = k_1\big[\,1 + \sqrt{2}\,\big(\,|\dot{\bar{x}}(t)| + |\bar{L}(t)| + \delta(t)\,\big)\,\big].
\end{aligned}$$

Note that δ and κ are integrable on $[a,b]$. Choose $t \in [a,b]$ so that $\delta(t)$ is finite (this holds for a.e. t). Now take any (x,v) so that $|x - \bar{x}(t)| < \varepsilon$ and $|(v, L(x,v)) - (\dot{\bar{x}}(t), \bar{L}(t))| < \delta(t)$. Thus $|x| < M$ and $|(v, L(x,v))| < |\dot{\bar{x}}(t)| + |\bar{L}(t)| + \delta(t)$, so (5.4) states that, for every $(w,p) \in \partial L(x,v)$,

$$\begin{aligned}
|w| &\le k_1\big[\,1 + (\,|v| + |L(x,v)|\,)\,\big](1+|p|)\\
&\le k_1\big[\,1 + \sqrt{2}|(v, L(x,v))|\,\big](1+|p|)\\
&\le k_1\big[\,1 + \sqrt{2}\,\big(\,|\dot{\bar{x}}(t)| + |\bar{L}(t)| + \delta(t)\,\big)\,\big](1+|p|)\\
&= \kappa(t)(1+|p|).
\end{aligned}$$

We also have that

$$\begin{aligned}
\kappa(t)/\delta(t) &= k_1[1 + \sqrt{2}\,(\,|\dot{\bar{x}}(t)| + |\bar{L}(t)| + \delta(t))]/\delta(t)\\
&\le k_1(1 + 2\sqrt{2}\delta(t))/\delta(t)\\
&\le k_1(1 + 2\sqrt{2}).
\end{aligned}$$

Thus $\kappa(t)/\delta(t)$ is essentially bounded and so ([22, Hypothesis H5]) holds. □

Ultimately, we wish to employ the main result of [23], which gives necessary conditions on an optimal arc to the following *general time* Bolza problem: Find a nondegenerate interval $[a,b]$ and arc $x \in \mathcal{A}_n^1[a,b]$ in order to

$$(\mathcal{P}_t) \qquad\qquad \text{minimize} \quad l(a, x(a), b, x(b)) + \int_a^b L(t, x(t), \dot{x}(t))\, dt.$$

The following theorem says that if our global assumptions on H are in place, then the result in [23] holds. Again, we are assuming that H and L satisfy (1.2) and (1.3).

THEOREM 5.3. *Let the arc $\bar{x}$ and interval $[\bar{a}, \bar{b}]$ provide the minimum in $(\mathcal{P}_t)$. Assume that the endpoint function $l : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and lsc; that $L(t, x, \cdot\,)$ is proper, lsc, coercive, and convex for each (t,x); and that* (A2) *holds. Then some absolutely continuous arc (h,p) taking values in $\mathbb{R} \times \mathbb{R}^n$ satisfies either the normal conditions or singular conditions below.*

Normal conditions:

(a) $(\dot{h}(t), \dot{p}(t)) \in \mathrm{co}\{(w_1, w_2) : (w_1, -w_2, \dot{\bar{x}}(t)) \in \partial H(t, \bar{x}(t), p(t))\}$
$= \mathrm{co}\{(w_1, w_2) : (-w_1, w_2, p(t)) \in\ \partial L(t, \bar{x}(t), \dot{\bar{x}}(t))\}$ *a.e.* $t \in [\bar{a}, \bar{b}]$.

(b) $h(t) = H(t, \bar{x}(t), p(t))$ *for all* $t \in [\bar{a}, \bar{b}]$.

(c) $(-h(\bar{a}), p(\bar{a}), h(\bar{b}), -p(\bar{b})) \in \partial l(\bar{a}, \bar{x}(\bar{a}), \bar{b}, \bar{x}(\bar{b}))$.

Singular conditions: One has $|(h(t), p(t))| > 0$ *for all* $t \in [\bar{a}, \bar{b}]$, *and*

(a$^\infty$) $(\dot{h}(t), \dot{p}(t)) \in \mathrm{co}\{(w_1, w_2) : (-w_1, w_2, p(t)) \in \ \partial^\infty L(t, \bar{x}(t), \dot{\bar{x}}(t))\}$ *a.e.* $t \in [\bar{a}, \bar{b}]$.
(b$^\infty$) $h(t) = \langle p(t), \dot{\bar{x}}(t)\rangle$ *a.e.* $t \in [\bar{a}, \bar{b}]$.
(c$^\infty$) $(-h(\bar{a}), p(\bar{a}), h(\bar{b}), -p(\bar{b})) \in \partial^\infty l(\bar{a}, \bar{x}(\bar{a}), \bar{b}, \bar{x}(\bar{b}))$.

Proof. Many of the details of the proof are in the paper [23]. We must convert our problem $(\mathcal{P}_t)$ into an equivalent problem over a fixed time interval and show that we can then use the necessary conditions given in [22].

Let $[\bar{a}, \bar{b}]$ and $\bar{x}(\cdot)$ be an optimal solution to $(\mathcal{P}_t)$. First we transform the problem into a fixed-time Bolza problem by using the Erdmann transform. This transform employs the Lagrangian L_m defined in (5.1), where we choose $m < \bar{b} - \bar{a}$. In particular, we consider the problem

$$(\Pi) \qquad \text{minimize} \quad l(\theta(0), \xi(0), \theta(1), \xi(1)) + \int_0^1 L_m(\theta(s), \xi(s), \dot{\xi}(s), \dot{\theta}(s))ds$$

over absolutely continuous arcs $(\theta, \xi) \in \mathcal{A}^1_{n+1}[0, 1]$. It turns out (see [23, Lemma 4.1]) that the arc $(\bar{\theta}(s), \bar{\xi}(s)) = (\bar{a} + (\bar{b} - \bar{a})s, \bar{x}(\bar{\theta}(s)))$ solves (Π). The goal is then to apply necessary conditions from [22] to $(\bar{\theta}, \bar{\xi})$ and then relate the conditions back to the original arc $\bar{x}$.

So we need to check that the hypotheses of [22] hold for the problem (Π). It is fairly straightforward to see that the first four hypotheses of that paper are satisfied. Our assumptions allow us to conclude that the subgradient inequality (5.2) of Proposition 5.1 holds. So now we can apply Lemma 5.2 to L_m and the arc $(\bar{\theta}, \bar{\xi})$. Lemma 5.2 says that ([22, Hypothesis H5]) holds. Thus the necessary conditions of [22] apply to L_m and (θ, ξ). Translating these conditions back in terms of $\bar{x}$, H, and L is done in section 4.3 of [23], giving us the results of our theorem.

One slight adjustment we have made is in condition (b). In [23] it is shown that (b) holds for a.e. t, but in our case we know that H is continuous (Proposition 2.4), which forces this condition to hold at every $t \in [\bar{a}, \bar{b}]$. □

We note that the conditions on L in this theorem are satisfied under (A), (A1), and (A2). The coercivity condition on L is equivalent to assuming that $H(t, x, \cdot)$ is finite for every (t, x). This is a basic property of the Legendre–Fenchel transform.

6. Proof of the main result. We begin this section with a proposition that will be used in the proof of the main theorem but is also of independent interest. Recall the normal cone definitions (1.6) and (1.7).

PROPOSITION 6.1. *Let* $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *be a locally sub-Lipschitz set-valued mapping (see* (1.11)*) with closed, convex images. Assume* K *is a closed set that is invariant with respect to the forward solution set (see Proposition* 4.1*). Then at each* $x \in K$ *we have*

$$\sup_{v \in F(x)} \langle v, y\rangle \leq 0 \quad \textit{for every } y \in N_K(x). \tag{6.1}$$

Proof. Take $\bar{x} \in K$ and $y \in N_K(\bar{x})$. Since F is locally sub-Lipschitz, there is a neighborhood W of $\bar{x}$ so that $d(0, F(x)) < \rho_0$ for all $x \in W$. Then for any $\rho \in (2\rho_0, \infty)$, the truncation mapping $F_\rho(x) := F(x) \cap \rho\mathbb{B}$ is locally Lipschitz continuous on W (from Theorem 9.33(a) in [27]). But if K is forward invariant under F, it also must be under F_ρ, since the solution set of the latter differential inclusion is contained in that of the former. Take a sequence (x^ν, y^ν) with $x^\nu \in K \cap W$ and $x^\nu \to \bar{x}$ so that $y^\nu \in \widehat{N}_K(x^\nu)$ and $y^\nu \to y$. Theorem 5.3.4 of [3] tells us then that $\sup_{v \in F_\rho(x^\nu)}\langle v, y^\nu\rangle \leq 0$. Then F_ρ being a bounded continuous mapping forces a similar inequality in the limit. That is,

$\sup_{v\in F_\rho(\bar{x})}\langle v, y\rangle \le 0$. But this is true regardless of how large we take ρ, so our result must hold. □

Proof of Theorem 2.2. First we will show that V satisfies the ULLG condition, as given in Definition 2.1. By (A0), there is a constant $k_1 > 0$ such that $\varphi(x) \ge -k_1(1+|x|)$. Fix $T>0$, and note that for $(t,x,p)\in[0,T]\times\mathbb{R}^n\times\mathbb{R}^n$, (A1) says that

$$H(t,x,p) \le H_1(t,x,p) := \mu(|p|) + (T+|x|)(\beta+\alpha|p|).$$

Let L_1 be the corresponding Lagrangian to H_1. For $(t,x,p)\in[0,T]\times\mathbb{R}^n\times\mathbb{R}^n$ it follows that $L_1(t,x,v)\le L(t,x,v)$. Thus if we let V_1 be the value function corresponding to the Bolza problem

$$\min\left\{ -k_1(1+|x(0)|) + \int_0^\tau L_1\bigl(t,x(t),\dot{x}(t)\bigr)dt \ \mid x(\tau)=\xi\right\},$$

we must have $V_1(\tau,\xi)\le V(\tau,\xi)$ for all $(\tau,\xi)\in[0,T]\times\mathbb{R}^n$. We can consider our Bolza problem to have the form of $(\mathcal{P}_t)$ by letting

$$l(a,x(a),b,x(b)) = \Psi_{\{0\}}(a) + \Psi_{\{\tau\}}(b) - k_1(1+|x(a)|) + \Psi_{\{\xi\}}(x(b)).$$

We want to apply Theorem 5.3. Clearly l is proper and lsc. Since $H_1(t,x,\,\cdot\,)$ is lsc, proper, convex, and finite for all (t,x), L_1 satisfies the hypotheses of Theorem 5.3. Furthermore, H_1 satisfies the epi-continuity condition in (t,x) as well as in p, and so by [26, Prop. 2.2], $(w_1,w_2,v)\in\partial H_1(t,x,p)$ implies that $(w_1,w_2)\in\partial_{(t,x)}H_1(t,x,p)$, from which it follows that $|(w_1,w_2)|\le(\beta+\alpha|p|)$, and thus (A2) holds.

Fix $(\bar{\tau},\bar{\xi})\in[0,T]\times\mathbb{R}^n$ and let $\bar{x}(\cdot)$ be the minimizing arc for $V_1(\bar{\tau},\bar{\xi})$, which exists by Proposition 3.1. Theorem 5.3 gives us an absolutely continuous arc $p(\,\cdot\,)$ which satisfies either normal or singular conditions. First we see that part (c) of the singular conditions cannot occur since it forces $p(\bar{a})=p(0)=0$. So the normal conditions must hold. Since $\partial l(\bar{a},\bar{x}(\bar{a}),\bar{b},\bar{x}(\bar{b}))\subset\mathbb{R}\times k_1\mathbb{B}\times\mathbb{R}\times\mathbb{R}^n$, we must have $|p(0)|\le k_1$. Part (a) of the normal conditions says that for a.e. t, $|\dot{p}(t)|\le(\beta+\alpha|p(t)|)$, and so $p(\,\cdot\,)$ must be bounded on $[0,\bar{\tau}]$, implying the existence of a constant k_2, independent of $(\bar{\tau},\bar{\xi})$, such that $\partial\mu(|p(t)|)\subset k_2\mathbb{B}$ for all $t\in[0,\bar{\tau}]$. Also implicit in normal condition (a) is that

$$\begin{aligned}\dot{\bar{x}}(t) &\in \partial_p H_1(t,\bar{x}(t),p(t))\\ &= \partial_p(\mu(|p|)+(T+|x|)(\beta+\alpha|p|))(\bar{x}(t),p(t))\\ &\subset \partial_p(\mu(|p|))(\bar{x}(t),p(t)) + \partial_p((T+|x|)(\beta+\alpha|p|))(\bar{x}(t),p(t))\\ &\subset k_2\mathbb{B}+\alpha(T+|\bar{x}(t)|)\mathbb{B}.\end{aligned}$$

So there exists a constant k_3, also independent of $(\bar{\tau},\bar{\xi})$, such that $|\dot{\bar{x}}(t)|\le k_3(1+|\bar{x}(t)|)$. This implies that $|\bar{x}(t)|\le(|\bar{\xi}|+1)e^{k_3T}$ for $t\in[0,\bar{\tau}]$. Now since $L_1(t,x,v)\ge -H_1(t,x,0)$,

$$\begin{aligned}\int_0^{\bar{\tau}} L_1(t,\bar{x}(t),\dot{\bar{x}}(t))dt &\ge \int_0^{\bar{\tau}} -\mu(0)-\beta(|T|+|\bar{x}(t)|)dt\\ &\ge \int_0^{\bar{\tau}} -\mu(0)-\beta(|T|+(|\bar{\xi}|+1)e^{k_3T})dt\\ &\ge -k_4(|\bar{\xi}|+1)\end{aligned}$$

for some constant $k_4 > 0$. Since $|\bar{x}(0)| \le (|\bar{\xi}| + 1)e^{k_3 T}$, we see then that

$$\begin{aligned} V_1(\tau, \xi) &\ge -k_1((|\xi| + 1)e^{k_3 T} + 1) - k_4(|\xi| + 1) \\ &= -k(1 + |\xi|) \end{aligned}$$

for some $k > 0$ and for all $(\tau, \xi) \in [0, T] \times \mathbb{R}^n$. So V satisfies the ULLG condition.

To see that the value function is a solution to (1.1) as given by Definition 1.2, we first note that Proposition 3.1 shows V is proper and lsc. By Definition 1.1, V satisfies (a) and (b) of Definition 1.2. The only remaining part that requires checking is the subgradient condition (c).

First we check for $\tau = 0$. Take $\xi \in \mathbb{R}^n$ and $(\sigma, \eta) \in \partial V(0, \xi)$. By Proposition 3.2, V satisfies property (P4). That is, $\operatorname{epi} V$ is invariant with respect to the differential inclusion $(\dot{t}(s), \dot{x}(s), \dot{y}(s)) \in \widetilde{E}_L((t(s), x(s), y(s))$. Assumptions (A) and (A2) imply that the mapping $\operatorname{epi} L(t, x, \,\cdot\,)$, by Proposition 2.6, is locally sub-Lipschitz. Then $\widetilde{E}_L(t, x, y) = \{1\} \times E_L(t, x)$ will also be locally sub-Lipschitz and Proposition 6.1 says that

$$\begin{aligned} &\sup_{(v,\beta) \in \operatorname{epi} L(0,\xi,\,\cdot\,)} \big\{ \big\langle\, (\sigma, \eta, -1), (1, v, \beta) \,\big\rangle \big\} \le 0, \\ &\Rightarrow \sup_{v \in \mathbb{R}^n} \big\{ \sigma + \langle \eta, v \rangle \; - L(0, \xi, v) \big\} \le 0, \\ &\Rightarrow \sigma + H(0, \xi, \eta) \le 0. \end{aligned}$$

Now take $(\sigma, \eta) \in \widehat{\partial} V(\tau, \xi)$ for some (τ, ξ) with $\tau > 0$. Let $\bar{x}(\,\cdot\,)$ be a minimizing arc for $(\mathcal{P}_{\tau,\xi})$. As mentioned in the introduction, there exists a differentiable function $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ such that $g(\tau, \xi) = V(\tau, \xi)$, but for any $(\tau', \xi') \ne (\tau, \xi)$, we have $g(\tau', \xi') < V(\tau', \xi')$. Furthermore, $\nabla g(\tau, \xi) = (\sigma, \eta)$. Consider the following general time Bolza problem:

$$(\mathcal{P}_g) \qquad\qquad \text{minimize} \quad l(a, x(a), b, x(b)) + \int_a^b L\big(t, x(t), \dot{x}(t)\big) dt,$$

where $l(a, x(a), b, x(b)) = \Psi_{\{0\}}(a) + \varphi(x(a)) - g(b, x(b))$ and we are minimizing over all nondegenerate time intervals $[a, b]$ and arcs $x \in \mathcal{A}_n^1([a, b])$. If $(\mathcal{P}_g)$ is to have a finite value, a must be 0. Evaluating the above expression for any feasible arc $x(\,\cdot\,)$ and interval $[0, b]$ must give us a nonnegative value, since

$$\begin{aligned} l(0, x(0), b, x(b)) + \int_0^b L\big(t, x(t), \dot{x}(t)\big) dt \; &\ge \varphi(x(0)) - g(b, x(b)) + \int_0^b L\big(t, x(t), \dot{x}(t)\big) dt \\ &\ge V(b, x(b)) - g(b, x(b)) \\ &\ge 0. \end{aligned}$$

It should be clear however, that taking $[a, b] = [0, \tau]$ and $x = \bar{x}(\,\cdot\,)$ gives the value 0 in the above. Thus we have an optimal solution. Now we wish to use the necessary conditions given in Theorem 5.3, which can be applied to our general time Bolza problem $(\mathcal{P}_g)$. First checking the singular conditions described in Theorem 5.3, we have

$$\begin{aligned} \partial^\infty l(0, \bar{x}(0), \tau, \bar{x}(\tau)) &\subset \partial^\infty \Psi_{\{0\}}(0) \times \partial^\infty \varphi(x(0)) \times \partial^\infty g(\tau, \bar{x}(\tau)) \\ &= \mathbb{R} \times \partial^\infty \varphi(x(0)) \times (0, 0). \end{aligned}$$

For the first containment, we have used the calculus rules in ([27, Prop. 10.5]). So there does not exist an arc (h,p) with $|(h(t),p(t))| > 0$ for $t \in [0,\tau]$ satisfying the singular conditions, since condition (c^∞) forces $(h(\tau),p(\tau)) = (0,0)$. So the normal conditions must hold. That is, there exist absolutely continuous arcs h and p satisfying the transversality condition

$$\begin{aligned}(-h(0),p(0),h(\tau),-p(\tau)) &\in \partial l(0,\bar{x}(0),\tau,\bar{x}(\tau))\\ &= \mathbb{R}\times\partial\varphi(x(0))\times(-\sigma,-\eta),\end{aligned}$$

and so $h(\tau) = -\sigma$ and $p(\tau) = \eta$. Also, condition (b) of the normal conditions says that $h(t) = H(t,\bar{x}(t),p(t))$ for all $t \in [0,\tau]$, so in particular $h(\tau) = H(\tau,\bar{x}(\tau),p(\tau))$, giving us that $\sigma + H(\tau,\xi,\eta) = 0$.

Now if we have $(\sigma,\eta) \in \partial V(\tau,\xi)$ for $\tau > 0$, by definition of the general subgradient, there exist sequences $(\tau^\nu,\xi^\nu) \to (\tau,\xi)$ with $V(\tau^\nu,\xi^\nu) \to V(\tau,\xi)$ and $(\sigma^\nu,\eta^\nu) \to (\sigma,\eta)$ with $(\sigma^\nu,\eta^\nu) \in \widehat{\partial}V(\tau^\nu,\xi^\nu)$ and $\tau^\nu > 0$. So by the result just proved, $H(\tau^\nu,\xi^\nu,\eta^\nu) = -\sigma^\nu$. But H being continuous implies $H(\tau^\nu,\xi^\nu,\eta^\nu) \to H(\tau,\xi,\eta)$ which forces $H(\tau,\xi,\eta) = -\sigma$. Thus V is a solution as given by Definition 1.2.

Let u be a solution to (1.1) as in Definition 1.2 that satisfies the ULLG condition. To prove the uniqueness part of the theorem, we need to show that $u = V$. Let $\mathcal{V} : \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n \to \overline{\mathbb{R}}$ be the value function for

$$(\mathcal{P}_{\theta,\tau,\xi}) \qquad \min \quad u(t(0),x(0)) + \int_0^\theta \mathcal{L}\big(s,t(s),x(s),\dot{t}(s),\dot{x}(s)\big)ds.$$

Here we are minimizing over $(t(\,\cdot\,),x(\,\cdot\,)) \in \mathcal{A}^1_{n+1}([0,\theta])$ with $(t(\theta),x(\theta)) = (\tau,\xi)$, and the Lagrangian is

$$\mathcal{L}(s,t,x,v_1,v_2) := L(t,x,v_2) + \Psi_{\{1\}}(v_1).$$

Note that this is a generalized problem of Bolza in which we are considering (t,x) as the state variable and s as the time parameter. The corresponding Hamiltonian is then

$$\begin{aligned}\mathcal{H}(s,t,x,p_1,p_2) &= \sup_{(v_1,v_2)\in\mathbb{R}\times\mathbb{R}^n} \big\{ \langle (v_1,v_2),(p_1,p_2)\rangle - \mathcal{L}(s,t,x,v_1,v_2) \big\}\\ &= \sup_{(v_1,v_2)\in\mathbb{R}\times\mathbb{R}^n} \big\{ \langle v_1,p_1\rangle + \langle v_2,p_2\rangle \ - L(t,x,v_2) - \Psi_{\{1\}}(v_1) \big\}\\ &= \sup_{v_2\in\mathbb{R}^n} \big\{ p_1 + \langle v_2,p_2\rangle - L(t,x,v_2) \big\}\\ &= p_1 + H(t,x,p_2).\end{aligned}$$

We would like to use the existence part of Theorem 2.2, just proved above, to show that $\mathcal{V}$ is a solution to

$$\begin{aligned}\mathcal{V}_s(s,t,x) + \mathcal{H}(s,t,x,\mathcal{V}_t(s,t,x),\mathcal{V}_x(s,t,x)) = 0, &\qquad (s,t,x) \in (0,\infty)\times\mathbb{R}\times\mathbb{R}^n,\\ \mathcal{V}(0,t,x) = u(t,x), &\quad (t,x) \in \mathbb{R}\times\mathbb{R}^n,\end{aligned}$$

in the sense of Definition 1.2. Clearly $\mathcal{H}$ satisfies (A), (A1), and (A2) as long as H does. Assumption (A0) is not necessarily satisfied however, since u could behave in a countercoercive manner with respect to its first variable. We can get around this problem as follows. Let $T > 0$ and consider adding an indicator function $\Psi_{[0,T]\times\mathbb{R}^n}(t(0),x(0))$ to the initial cost $u(t(0),x(0))$ in $(\mathcal{P}_{\theta,\tau,\xi})$. Now the initial cost does satisfy (A0) and the new resulting value function $\bar{\mathcal{V}}$ is therefore a solution to (1.1) as given by Definition 1.2, and it has equation

$$\bar{\mathcal{V}}(\theta,\tau,\xi) = \begin{cases} \mathcal{V}(\theta,\tau,\xi) & \text{if } \ \tau \le T+\theta, \\ +\infty & \text{otherwise.} \end{cases}$$

This implies that $\mathcal{V}$ must satisfy (a)–(c) of Definition 1.2, and for any given $(\theta,\tau,\xi) \in \operatorname{dom}\mathcal{V}$, it suffices to take $T > \theta+\tau$ to compare $\mathcal{V}$ with $\bar{\mathcal{V}}$ and see that (1.8) holds. Thus $\mathcal{V}$ is a solution to (1.1).

First consider the case where $0 < \theta$. If we were to go through the existence proof above, we can again apply the technique of introducing a differentiable function $g \le \mathcal{V}$ whose gradient is (α,σ,η) at the point (θ,τ,ξ), and using the results from Theorem 5.3, assert the existence of absolutely continuous arcs $h(\cdot)$ and $p(\cdot)$ with the properties

(a) $h(s) = \mathcal{H}(s,\bar{t}(s),\bar{x}(s),p(s))$ for all $s \in [0,\theta]$,

(b) $(-h(0),p(0),h(\theta),-p(\theta)) \in \mathbb{R}\times\partial u(\bar{t}(0),\bar{x}(0)) \ \times(-\alpha,-\sigma,-\eta)$,

(c) $\dot{h}(s) = 0$ for a.e. $s \in [0,\theta]$.

Conditions (b) and (c) imply then that $h(s) = -\alpha$ for all $s \in [0,\theta]$. If we denote $p(0)$ by (σ_0,η_0), then (b) implies $(\sigma_0,\eta_0) \in \partial u(\bar{t}(0),\bar{x}(0))$. But then (a) gives us that

$$\begin{aligned} -\alpha &= \mathcal{H}(\theta,\tau,\xi,\sigma,\eta) \\ &= \mathcal{H}(0,\bar{t}(0),\bar{x}(0),\sigma_0,\eta_0) \\ &= \sigma_0 + H(\bar{t}(0),\bar{x}(0),\eta_0). \end{aligned}$$

By assumption, this last quantity is equal to 0 if $\bar{t}(0) > 0$ and is less than or equal to 0 if $\bar{t}(0) = 0$. Since $\dot{\bar{t}}(s) = 1$, and $\bar{t}(\theta) = \tau$, the only way $\bar{t}(0) = 0$ is if $\theta = \tau$. Thus we have

$$\begin{cases} -\alpha = 0 & \text{if } 0 < \theta < \tau, \\ -\alpha \le 0 & \text{if } 0 < \theta = \tau. \end{cases}$$

Now consider $(\alpha,\sigma,\eta) \in \widehat{\partial}\mathcal{V}(0,\tau,\xi)$ with $\tau > 0$. Again applying the existence part of Theorem 2.2 to $\mathcal{V}$, we have $\alpha + \mathcal{H}(0,\tau,\xi,\sigma,\eta) \le 0$. Using the calculus of regular subgradients (see Corollary 10.11 in [27]), we see that $(\sigma,\eta) \in \widehat{\partial}_{(t,x)}\mathcal{V}(0,\tau,\xi) = \widehat{\partial}u(\tau,\xi)$. By assumption then, $\mathcal{H}(0,\tau,\xi,\sigma,\eta) = \sigma + H(\tau,\xi,\eta) = 0$. It follows that $\alpha \le 0$.

Now looking at a general subgradient $(\alpha,\sigma,\eta) \in \partial\mathcal{V}(0,\tau,\xi)$ with $\tau > 0$, there exist sequences $(\alpha^\nu,\sigma^\nu,\eta^\nu) \to (\alpha,\sigma,\eta)$ and $(\theta^\nu,\tau^\nu,\xi^\nu) \to (0,\tau,\xi)$ with $(\alpha^\nu,\sigma^\nu,\eta^\nu) \in \widehat{\partial}\mathcal{V}(\theta^\nu,\tau^\nu,\xi^\nu)$ and $\mathcal{V}(\theta^\nu,\tau^\nu,\xi^\nu) \to \mathcal{V}(0,\tau,\xi)$. Furthermore, since $\tau > 0$, we can take $\theta^\nu < \tau^\nu$. We have just seen though that for this sequence $\alpha^\nu \le 0$, so we must have $\alpha \le 0$.

In summary then, for $(\alpha,\sigma,\eta) \in \partial\mathcal{V}(\theta,\tau,\xi)$, we have

$$\begin{cases} \alpha \le 0 & \text{if } 0 = \theta < \tau, \\ \alpha = 0 & \text{if } 0 < \theta < \tau, \\ \alpha \ge 0 & \text{if } 0 < \theta = \tau. \end{cases}$$

It should be clear from Proposition 3.1 that $\mathcal{V}$ is lsc and proper. So Proposition 4.4 says that in fact $\mathcal{V} = \widetilde{u}$. That is,

$$\mathcal{V}(\theta,\tau,\xi) = \begin{cases} u(\tau,\xi) & \text{if } 0 \le \theta \le \tau, \\ +\infty & \text{otherwise.} \end{cases}$$

But epi $\mathcal{V}(\theta,\,\cdot\,,\,\cdot\,)$ is the reachable set of epi u at time θ under the differential inclusion $\dot{z}(s) \in \widetilde{E}_L(z(s))$. That is,

$$\operatorname{epi} \mathcal{V}(\theta,\,\cdot\,,\,\cdot\,) = \mathcal{R}_{\operatorname{epi} u}(\theta).$$

Now from Proposition 4.3, we have that u satisfies conditions (P1)–(P4), and finally Proposition 3.2 implies that u is the value function V for $(\mathcal{P}_{\tau,\xi})$. □

Acknowledgment. The author wishes to express his gratitude for the assistance provided by Professor R. T. Rockafellar.

REFERENCES

[1] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[2] J.-P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[3] J.-P. Aubin, *Viability Theory*, Birkhäuser, Boston, 1991.

[4] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[5] G. Barles and B. Perthame, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 557–579.

[6] G. Barles, *Discontinuous viscosity solutions of first-order Hamilton-Jacobi equations: A guided visit*, Nonlinear Anal., 20(1993), pp. 1123–1134.

[7] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1983.

[9] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[10] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[11] M. G. Crandall, L. C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[12] M. G. Crandall and P.-L. Lions, *On existence and uniqueness of solutions of Hamilton-Jacobi equations*, Nonlinear Anal., 10 (1986), pp. 353–370.

[13] M. G. Crandall and P.-L. Lions, *Remarks on the existence and uniqueness of unbounded viscosity solutions of Hamilton-Jacobi equations*, Illinois J. Math., 31 (1987), pp. 665–688.

[14] M. G. Crandall, H. Ishii, and P.-L. Lions, *Uniqueness of viscosity solutions of Hamilton-Jacobi equations revisited*, J. Math. Soc. Japan, 39 (1987), pp. 581–595.

[15] K. Diemling, *Multivalued Differential Equations*, Walter de Gruyter, Berlin, 1992.

[16] H. Frankowska, *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[17] H. Frankowska and S. Plaskacz, *Semicontinuous solutions of Hamilton-Jacobi-Bellman equations with state constraints*, Lecture Notes in Nonlinear Anal. 2, 1998, pp. 145–161.

[18] H. Ishii, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.

[19] H. Ishii, *Representation of solutions of Hamilton-Jacobi equations*, Nonlinear Anal., 12 (1988), pp. 121–146.

[20] P. D. Loewen, *Optimal Control via Nonsmooth Analysis*, CRM-AMS Lecture Notes in Math. 2, AMS, Providence, RI, 1993.

[21] P. D. Loewen and R. T. Rockafellar, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.

[22] P. D. Loewen and R. T. Rockafellar, *New necessary conditions for the generalized problem of Bolza*, SIAM J. Control Optim., 34 (1996), pp. 1496–1511.

[23] P. D. Loewen and R. T. Rockafellar, *Bolza problems with general time constraints*, SIAM J. Control Optim., 35 (1997), pp. 2050–2069.

[24] B. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[25] R. T. Rockafellar, *Existence theorems for general control problems of Bolza and Lagrange*, Adv. Math., 15 (1975), pp. 312–333.

[26] R. T. Rockafellar, *Equivalent subgradient versions of Hamiltonian and Euler-Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1314.

[27] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer–Verlag, Berlin, 1997.

[28] R. T. Rockafellar and P. R. Wolenski, *Convexity in Hamilton-Jacobi theory,* 1: *Dynamics and duality*, SIAM J. Control Optim., to appear.

[29] A. I. Subbotin, *A generalization of the basic equation in the theory of differential games*, Soviet Math. Dokl., 22 (1980), pp. 358–362.

[30] A. I. Subbotin, *Existence and uniqueness results for Hamilton-Jacobi equations*, Nonlinear Anal., 16 (1991), pp. 683–699.

[31] A. I. Subbotin, *Generalized Solutions of First-Order PDEs: The Dynamic Optimization Perspective*, Birkhäuser, Boston, 1995.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 1, pp. 306–329

UTILITY MAXIMIZATION WITH DISCRETIONARY STOPPING*

IOANNIS KARATZAS† AND HUI WANG‡

Abstract. Utility maximization problems of mixed *optimal stopping/control* type are considered, which can be solved by reduction to a family of related *pure optimal stopping* problems. Sufficient conditions for the existence of optimal strategies are provided in the context of continuous-time, Itô process models for complete markets. The mathematical tools used are those of optimal stopping theory, continuous-time martingales, convex analysis, and duality theory. Several examples are solved explicitly, including one which demonstrates that optimal strategies need not always exist.

Key words. utility maximization, stochastic control, optimal stopping, variational inequality, duality, convex analysis, martingale representation

AMS subject classifications. Primary, 93E20, 90A09, 60H30; Secondary, 60G44, 49N15, 90A16

PII. S0363012998346323

1. Introduction.

Problems of expected utility maximization go back at least to the seminal articles of Samuelson and Merton (1969) and Merton (1971), and have been studied extensively in recent years, for instance by Pliska (1986), Karatzas, Lehoczky, and Shreve [KLS] (1987), and Cox and Huang (1989). Most of this literature shares the common setting of an agent who receives a deterministic initial capital, which he must then invest in a market (complete or incomplete) so as to maximize the expected utility of his wealth and/or consumption, up to a *prespecified* terminal time.

In this paper we consider a variant of these problems by allowing the agent freely to stop *before or at* a prespecified final time in order to maximize the expected utility of his wealth and/or consumption up to the stopping time. The assets available to the agent can be traded continuously, without restrictions, frictions, or transaction costs; they consist of a locally riskless money-market, and m risky stocks. (One can think, for example, of an investor or mutual fund manager who tries to invest/consume as skillfully as possible before "retiring" from the stock market and putting all his holdings in the money-market.) The stock prices are driven by m independent Brownian motions; these represent the sources of uncertainty in the market model, which is assumed to be complete in the sense of Harrison and Pliska (1981). The market coefficients, i.e., the money-market rate, the stock-appreciation rates, and the matrix of stock volatilities, are bounded random processes adapted to the driving m-dimensional Brownian motion.

The utility maximization problem studied here involves aspects of both *optimal stopping* and *stochastic control.* Such problems also arise in situations like pricing American contingent claims under constraints, selecting trading strategies in the presence of transaction costs with an American option held in the portfolio, target-tracking

*Received by the editors October 19, 1998; accepted for publication (in revised form) February 16, 2000; published electronically August 24, 2000. This research was supported in part by the National Science Foundation under grant NSF-DMS-97-32810.

http://www.siam.org/journals/sicon/39-1/34632.html

†Departments of Mathematics and Statistics, 619 Mathematics Building, Columbia University, New York, NY 10027 (ik@math.columbia.edu).

‡Department of Statistics, 618 Mathematics Building, Columbia University, New York, NY 10027 (wanghui@stat.columbia.edu).

followed by a decision (to engage the target or not), etc.; see Karatzas and Kou (1998), Davis and Zariphopoulou (1995), Davis and Zervos (1994), as well as Karatzas and Sudderth (1999) for such problems in different contexts. The free-boundary problem approach, based on an associated Hamilton–Jacobi–Bellman (HJB) equation of dynamic programming, is inadequate for the analysis of the general version of our model, which is not Markovian. Instead, duality theory plays an important role and leads to a family of pure optimal stopping problems which is even more amenable to analysis. Duality approaches have been used with success in treating portfolio optimization problems for financial markets which are incomplete or impose constraints on portfolio choice, as in Karatzas, Lehoczky, Shreve, and Xu [KLSX] (1991), Shreve and Xu (1992), and Cvitanić and Karatzas (1992).

The model and the utility maximization problem are described in sections 2–5. We present a solution in section 6 using a *duality approach.* However, this solution is not quite satisfactory in the sense that it leads to computationally tractable results only in very special cases and does not shed much light on the general question of existence of optimal strategies. We then introduce and analyze *a family of pure optimal stopping problems* in sections 7–8. In terms of these, we are able to provide conditions which guarantee the existence of optimal strategies. In section 9, several examples are presented, one of which demonstrates that optimal strategies need *not* always exist. For completeness, we treat in Appendix A an example which can be solved explicitly using a free-boundary problem for the associated HJB equation. In Appendix B we formulate an open problem, suggested by the referee, where consumption continues past the time of retirement from the stock market.

It is hoped that the analysis in this paper will serve as a step towards establishing a general theory for stochastic control problems with discretionary stopping in continuous time, possibly along the lines of the Dubins–Savage (1965) theory for discrete-time "leavable gambling problems" developed in Chapter 3 of Maitra and Sudderth (1996).

Remark 1.1. We denote by "standing assumption" those conditions that are always in force throughout the paper; they will not be cited in theorems. And "assumption" stands for those conditions which are in force only when theorems specifically cite them.

2. The market model.

We adopt a model consisting of a money-market, with price $P_0(\cdot)$ given by

$$dP_0(t) = P_0(t) r(t)\, dt, \qquad P_0(0) = 1, \tag{2.1}$$

and of m stocks with prices-per-share $P_i(\cdot)$ satisfying the equations

$$dP_i(t) = P_i(t) \left[b_i(t)\, dt + \sum_{j=1}^{m} \sigma_{ij}(t) dW_j(t) \right], \qquad i = 1, \dots, m. \tag{2.2}$$

Here $W(\cdot) = (W_1(\cdot), \dots, W_m(\cdot))^*$ is an m-dimensional Brownian motion on a complete probability space $(\Omega, \mathbb{F}, \mathbb{P})$. We shall denote by $\mathbb{F} = \{\mathcal{F}_t\}_{0 \le t \le T}$ the $\mathbb{P}$-augmentation of the filtration generated by $W(\cdot)$. The *coëfficients* of the model, that is, the scalar *interest rate* process $r(\cdot)$, the vector process $b(\cdot) = (b_1(\cdot), \dots, b_m(\cdot))^*$ of *appreciation rates*, and the matrix-valued *volatility process* $\sigma(\cdot) = (\sigma_{ij}(\cdot))_{1 \le i,j \le m}$, are assumed to be bounded, and progressively measurable with respect to $\mathbb{F}$. All processes encountered throughout sections 2–9 of the paper will be defined on the fixed, finite horizon $[0, T]$.

Standing Assumption 2.1. We assume that $\|b(t)\| \leq L,\ |r(t)| \leq L,\ \forall\, 0 \leq t \leq T$ hold almost surely (a.s.) for some given real constant $L > 0$.

Standing Assumption 2.2. The process $\sigma(\cdot)$ satisfies the strong nondegeneracy condition

$$\xi^* \sigma(t)\sigma^*(t)\xi \geq \epsilon \|\xi\|^2 \qquad \forall\ (t, \xi) \in [0, T] \times \mathbb{R}^m$$

a.s. for some given real constant $\epsilon > 0$. From Standing Assumption 2.2, the matrices $\sigma(t), \sigma^*(t)$ are invertible, and the norms of $(\sigma(t))^{-1}$ and $(\sigma^*(t))^{-1}$ are bounded from above and below by δ and δ^{-1}, respectively, for some $\delta \in (1, \infty)$; cf. Karatzas and Shreve (1991), p. 372. We also define the "relative risk" process

$$\theta(t) \overset{\triangle}{=} \sigma^{-1}(t)[b(t) - r(t)\mathbf{1}_m], \tag{2.3}$$

where $\mathbf{1}_m = (1, \ldots, 1)^*$, the discount process

$$\gamma(t) \overset{\triangle}{=} \frac{1}{P_0(t)} = \exp\left\{ -\int_0^t r(s)\, ds \right\}, \tag{2.4}$$

the exponential martingale (or *likelihood ratio* process)

$$Z_0(t) \overset{\triangle}{=} \exp\left\{ -\int_0^t \theta^*(s)\, dW(s) - \frac{1}{2}\int_0^t \|\theta(s)\|^2\, ds \right\}, \tag{2.5}$$

and the state-price-density process

$$H(t) \overset{\triangle}{=} \gamma(t) Z_0(t). \tag{2.6}$$

3. Portfolio and wealth processes. A *portfolio process* $\pi(\cdot) = (\pi_1(\cdot), \ldots, \pi_m(\cdot))^*$ is $\mathbb{R}^m$-valued, and a *consumption process* $c(\cdot)$ takes values in $[0, \infty)$; these are both $\mathbb{F}$-progressively measurable and satisfy

$$\int_0^T c(t)\, dt + \int_0^T \|\pi(t)\|^2\, dt < \infty$$

a.s. We regard $\pi_i(t)$ as the proportion of an agent's wealth invested in stock i at time t; the remaining proportion $1 - \pi^*(t)\mathbf{1}_m = 1 - \sum_{i=1}^m \pi_i(t)$ is invested in the money-market. These proportions are *not* constrained to take values in the interval $[0, 1]$; in other words, we allow both short-selling of stocks and borrowing at the interest rate of the bond. For a given, nonrandom, initial capital $x > 0$, let $X(\cdot) \equiv X^{x,\pi,c}(\cdot)$ denote the *wealth-process* corresponding to a portfolio/consumption process pair $\big(\pi(\cdot), c(\cdot)\big)$ as above. This wealth-process is defined by the initial condition $X^{x,\pi,c}(0) = x$ and the equation

$$\begin{aligned} dX(t) &= \sum_{i=1}^m \pi_i(t) X(t) \left\{ b_i(t)\, dt + \sum_{j=1}^m \sigma_{ij}(t) dW_j(t) \right\} \\ &\quad + \left\{ 1 - \sum_{i=1}^m \pi_i(t) \right\} X(t) r(t)\, dt - c(t)\, dt \\ &= r(t)X(t)dt + X(t)\pi^*(t)\sigma(t)dW_0(t) - c(t)\, dt, \qquad X(0) = x > 0, \end{aligned} \tag{3.1}$$

where we have set

$$W_0(t) \overset{\triangle}{=} W(t) + \int_0^t \theta(s)\, ds, \qquad 0 \le t \le T. \tag{3.2}$$

In other words,

$$d\left(\gamma(t)X^{x,\pi,c}(t)\right) = \gamma(t)X^{x,\pi,c}(t)\pi^*(t)\sigma(t)\, dW_0(t) - \gamma(t)c(t)\, dt, \quad 0 \le t \le T. \tag{3.3}$$

The process $W_0(\cdot)$ of (3.2) is Brownian motion under the *equivalent martingale measure*

$$\mathbb{P}_0(A) \overset{\triangle}{=} \mathbb{E}\big[Z_0(T)\mathbf{1}_A\big], \quad A \in \mathcal{F}_T\ , \tag{3.4}$$

by the Girsanov theorem (section 3.5 in Karatzas and Shreve (1991)). We shall say that a portfolio/consumption process pair (π, c) is *available at initial capital* $x > 0$ if the corresponding wealth-process $X^{x,\pi,c}(\cdot)$ of (3.3) is strictly positive on $[0, T]$ a.s.

An application of Itô's rule to the product of the processes $Z_0(\cdot)$ and $\gamma(\cdot)X^{x,\pi,c}(\cdot)$ leads to

$$\begin{aligned} H(t)X^{x,\pi,c}(t) &+ \int_0^t H(s)c(s)\, ds \\ &= x + \int_0^t H(s)X^{x,\pi,c}(s)(\sigma^*(s)\pi(s) - \theta(s))^*\, dW(s). \end{aligned} \tag{3.5}$$

This shows, in particular, that for any pair (π, c) available at initial capital $x > 0$, the process $H(\cdot)X^{x,\pi,c}(\cdot) + \int_0^\cdot H(s)c(s)\, ds$ is a continuous, positive local martingale, hence a supermartingale, under $\mathbb{P}$. Consequently, the optional sampling theorem gives

$$\mathbb{E}\left[H(\tau)X^{x,\pi,c}(\tau) + \int_0^\tau H(s)c(s)\, ds\right] \le x \qquad \forall\ \tau \in \mathcal{S}. \tag{3.6}$$

Here and in what follows, we denote by $\mathcal{S}_{s,t}$ the class of $\mathbb{F}$-stopping times $\tau : \Omega \longrightarrow [s, t]$ for $0 \le s \le t \le T$, and let $\mathcal{S} \equiv \mathcal{S}_{0,T}$.

4. Utility function. A function $U : (0, \infty) \longrightarrow \mathbb{R}$ will be called *utility function* if it is strictly increasing, strictly concave, continuously differentiable, and satisfies

$$U'(0+) \ \overset{\triangle}{=}\ \lim_{x \downarrow 0} U'(x) = \infty, \qquad U'(\infty)\ \overset{\triangle}{=}\ \lim_{x \uparrow \infty} U'(x) = 0. \tag{4.1}$$

We shall denote by $I(\cdot)$ the (continuous, strictly decreasing) inverse of the marginal-utility function $U'(\cdot)$; this function maps $(0, \infty)$ onto itself and satisfies $I(0+) = \infty,\ I(\infty) = 0$. We also introduce the Legendre–Fenchel transform

$$\tilde{U}(y)\ \overset{\triangle}{=}\ \max_{x>0}\, [U(x) - xy]\ =\ U(I(y)) - yI(y), \qquad 0 < y < \infty, \tag{4.2}$$

of $-U(-x)$; this function $\tilde{U}(\cdot)$ is strictly decreasing, strictly convex, and satisfies

$$\tilde{U}'(y) = -I(y), \qquad 0 < y < \infty, \tag{4.3}$$

$$U(x) = \min_{y>0}\, [\tilde{U}(y) + xy]\ =\ \tilde{U}(U'(x)) + xU'(x), \qquad 0 < x < \infty. \tag{4.4}$$

The inequality

$$U(I(y)) \ge U(x) + y[I(y) - x] \qquad \forall\ x > 0,\ y > 0, \tag{4.5}$$

is a direct consequence of (4.2).

5. The optimization problem. The agent in our model has time-dependent utility of the form $\int_0^t e^{-\beta s} U_1\big(c(s)\big)\,ds + e^{-\beta t} U_2(x)$, with $\beta \geq 0$ a real constant. The utility functions $U_1(\cdot)$, $U_2(\cdot)$ measure his utility from consumption and wealth, respectively, whereas β stands for a discount factor. If the agent uses the portfolio/consumption strategy (π, c) available at initial capital $x > 0$, and the stopping rule $\tau \in \mathcal{S}$, his *expected discounted utility* is

$$J(x; \pi, c, \tau) \overset{\triangle}{=} \mathbb{E}\left[\int_0^\tau e^{-\beta t} U_1\big(c(t)\big)\,dt + e^{-\beta\tau} U_2(X^{x,\pi,c}(\tau))\right]. \tag{5.1}$$

The optimization problem considered in this paper is the following: *to maximize the expected discounted utility in* (5.1), over the class $\mathcal{A}(x)$ of triples (π, c, τ) as above, for which the expectation in (5.1) is well defined, i.e.,

$$\mathbb{E}\left[\int_0^\tau e^{-\beta t} U_1^-\big(c(t)\big)\,dt + e^{-\beta\tau} U_2^-(X^{x,\pi,c}(\tau))\right] < \infty. \tag{5.2}$$

(Here and in what follows, x^- denotes the negative part of the real number x, namely, $x^- = \max(-x, 0)$.) The value-function of this problem will be denoted by

$$V(x) \overset{\triangle}{=} \sup_{(\pi,c,\tau)\in\mathcal{A}(x)} J(x; \pi, c, \tau), \qquad x \in (0, \infty). \tag{5.3}$$

We say that the value $V(x)$ is "attainable" if we can find a triple $(\hat\pi, \hat c, \hat\tau) \in \mathcal{A}(x)$ with $V(x) = J(x, \hat\pi, \hat c, \hat\tau)$; such a triple is then called "optimal" for problem (5.3). To ensure that this problem is meaningful, we impose the following assumption throughout.

Standing Assumption 5.1. $V(x) < \infty \quad \forall x \in (0, \infty)$.

It is fairly straightforward that the function $V(\cdot)$ is increasing on $(0, \infty)$. However, it is not clear at this stage whether $V(\cdot)$ is concave or not. We shall discuss this issue in section 8.

Remark 5.2. A sufficient condition for Standing Assumption 5.1 is that

$$\max\{U_1(x),\ U_2(x)\} \leq k_1 + k_2 x^\delta \qquad \forall\ x \in (0, \infty) \tag{5.4}$$

holds for some $k_1 > 0,\ k_2 > 0,\ \delta \in (0, 1)$; cf. Remark 3.6.8 in Karatzas and Shreve (1998).

6. Duality approach. For any *fixed* stopping time $\tau \in \mathcal{S}$, we denote by $\Pi_\tau(x)$ the set of portfolio/consumption process pairs (π, c) for which $(\pi, c, \tau) \in \mathcal{A}(x)$. The solution of the utility maximization problem

$$V_\tau(x) \overset{\triangle}{=} \sup_{(\pi,c)\in\Pi_\tau(x)} J(x; \pi, c, \tau) \tag{6.1}$$

can be derived as in KLS (1987). We review briefly the results in this section. For any triple $(\pi, c, \tau) \in \mathcal{A}(x)$ and any real number $\lambda > 0$, it follows from (4.2), (3.6) that

$$\begin{aligned}
J(x; \pi, c, \tau) &= \mathbb{E}\left[\int_0^\tau e^{-\beta t} U_1\big(c(t)\big)\,dt + e^{-\beta\tau} U_2(X^{x,\pi,c}(\tau))\right] \\
&\leq \mathbb{E}\left[\int_0^\tau e^{-\beta t} \tilde U_1(\lambda e^{\beta t} H(t))\,dt + e^{-\beta\tau} \tilde U_2(\lambda e^{\beta\tau} H(\tau))\right] \\
&\qquad + \lambda \cdot \mathbb{E}\left[H(\tau) X^{x,\pi,c}(\tau) + \int_0^\tau H(t) c(t)\,dt\right] \\
&\leq \mathbb{E}\left[\int_0^\tau e^{-\beta t} \tilde U_1(\lambda e^{\beta t} H(t))\,dt + e^{-\beta\tau} \tilde U_2(\lambda e^{\beta\tau} H(\tau))\right] + \lambda x,
\end{aligned}$$

with equality if and only if

$$X^{x,\pi,c}(\tau) = I_2(\lambda e^{\beta\tau} H(\tau)) \text{ and } c(t) = I_1\big(\lambda e^{\beta t} H(t)\big) \quad \forall\, 0 \le t \le \tau \quad \text{a.s.,} \tag{6.2}$$

$$\mathbb{E}\left[H(\tau)X^{x,\pi,c}(\tau) + \int_0^\tau H(t)c(t)\,dt\right] = x \tag{6.3}$$

hold. It develops that we have $V_\tau(x) \;\le\; \inf_{\lambda>0}\left[\tilde{J}(\lambda;\tau) + \lambda x\right] \quad \forall\, \tau \in \mathcal{S},$ as well as

$$V(x) \;=\; \sup_{\tau\in\mathcal{S}} V_\tau(x) \;\le\; \sup_{\tau\in\mathcal{S}} \inf_{\lambda>0} \left[\tilde{J}(\lambda;\tau) + \lambda x\right] \tag{6.4}$$

with the notation

$$\tilde{J}(\lambda;\tau) \overset{\triangle}{=} \mathbb{E}\left[\int_0^\tau e^{-\beta t}\tilde{U}_1\big(\lambda e^{\beta t}H(t)\big)\,dt + e^{-\beta\tau}\tilde{U}_2(\lambda e^{\beta\tau}H(\tau))\right]. \tag{6.5}$$

In order to proceed, we shall need the following assumption (see Remark 6.7 for discussion).

Assumption 6.1. $\mathbb{E}\left[\sup_{0\le t\le T}\big(H(t)\cdot I_2(\lambda e^{\beta t}H(t)\big) + \int_0^T H(t)I_1(\lambda e^{\beta t}H(t))\,dt\right]$ $< \infty \quad \forall \lambda \in (0,\infty)$.

Under this assumption, for any given $\tau \in \mathcal{S}$, the function $\mathcal{X}_\tau : (0,\infty) \to (0,\infty)$ defined by

$$\mathcal{X}_\tau(\lambda) \overset{\triangle}{=} \mathbb{E}\left[\int_0^\tau H(t)I_1\big(\lambda e^{\beta t}H(t)\big)\,dt + H(\tau)\cdot I_2(\lambda e^{\beta\tau}H(\tau))\right],\ 0<\lambda<\infty, \tag{6.6}$$

is a continuous, strictly decreasing mapping of $(0,\infty)$ onto itself with $\mathcal{X}_\tau(0+) = \infty$, $\mathcal{X}_\tau(\infty) = 0$; thus $\mathcal{X}_\tau(\cdot)$ has a continuous, strictly decreasing inverse $\mathcal{Y}_\tau(\cdot)$ from $(0,\infty)$ onto itself. We define

$$\xi^x(\tau) \overset{\triangle}{=} I_2\big(\mathcal{Y}_\tau(x)e^{\beta\tau}H(\tau)\big) \text{ and } \eta^x(t) \overset{\triangle}{=} I_1\big(\mathcal{Y}_\tau(x)e^{\beta t}H(t)\big),\ \ 0 \le t \le T, \tag{6.7}$$

so that, in particular,

$$\mathbb{E}\left[H(\tau)\xi^x(\tau) + \int_0^\tau H(t)\eta^x(t)\,dt\right] = x. \tag{6.8}$$

LEMMA 6.2. *For any $\tau \in \mathcal{S}$, the random variables of* (6.7) *satisfy*

$$\mathbb{E}\left[e^{-\beta\tau}U_2^-\big(\xi^x(\tau)\big) + \int_0^\tau e^{-\beta t}U_1^-\big(\eta^x(t)\big)\,dt\right] < \infty, \tag{6.9}$$

and for every portfolio/consumption pair $(\pi, c) \in \Pi_\tau(x)$ we have

$$\begin{aligned}&\mathbb{E}\left[\int_0^\tau U_1\big(c(t)\big)\,dt + e^{-\beta\tau}U_2(X^{x,\pi,c}(\tau))\right]\\ &\qquad \le\; \mathbb{E}\left[\int_0^\tau U_1\big(\eta^x(t)\big)\,dt + e^{-\beta\tau}U_2(\xi^x(\tau))\right].\end{aligned} \tag{6.10}$$

Lemma 6.2 can be proved by arguments similar to those used in the proof of Theorem 3.6.3 in Karatzas and Shreve (1998). We conclude from Lemma 6.2 that, if there exists a portfolio $\hat{\pi}_\tau(\cdot)$ such that $(\hat{\pi}_\tau, \hat{c}_\tau)$ is available at initial capital $x > 0$, where $\hat{c}_\tau(\cdot) \overset{\triangle}{=} \eta^x(\cdot)\mathbf{1}_{[0,\tau[}(\cdot)$, and if

$$X^{x,\hat{\pi}_\tau,\hat{c}_\tau}(\tau) = \xi^x(\tau) \tag{6.11}$$

holds a.s., then the pair $(\hat{\pi}_\tau, \hat{c}_\tau)$ belongs to $\Pi_\tau(x)$ and is optimal for the utility maximization problem (6.1). The existence of such a portfolio will need the assumption of market completeness, as we shall see in the next lemma.

LEMMA 6.3. *For any $\tau \in \mathcal{S}$, any $\mathcal{F}_\tau$-measurable random variable B with $\mathbb{P}[B > 0] = 1$, and any progressively measurable process $c(\cdot) \geq 0$ that satisfies $c(\cdot) \equiv 0$ almost everywhere (a.e.) on $[\![\tau, T]\!]$ as well as $\mathbb{E}\left[H(\tau)B + \int_0^T H(t)c(t)\,dt\right] = x$, there exists a portfolio process $\pi(\cdot)$ such that, a.s.*

$$X^{x,\pi,c}(t) > 0, \quad 0 \leq t \leq T, \qquad \text{and} \qquad X^{x,\pi,c}(\tau) = B.$$

Proof. We begin with the strictly positive, continuous process $X(\cdot)$ defined by

$$X(t) \overset{\triangle}{=} \frac{1}{\gamma(t)} \cdot \mathbb{E}_0\left[\gamma(\tau)B + \int_{t\wedge\tau}^{\tau} \gamma(s)c(s)\,ds \middle| \mathcal{F}_t\right], \qquad 0 \leq t \leq T.$$

This process satisfies

$$X(0) = \mathbb{E}_0\left[\gamma(\tau)B + \int_0^\tau \gamma(s)c(s)\,ds\right] = \mathbb{E}\left[H(\tau)B + \int_0^\tau H(s)c(s)\,ds\right] = x,$$

and $X(\tau) = B$ a.s. On the other hand, the $\mathbb{P}_0$-martingale

$$M(\cdot) \overset{\triangle}{=} \gamma(\cdot)X(\cdot) + \int_0^{\cdot} \gamma(s)c(s)\,ds = \mathbb{E}_0\left[\gamma(\tau)B + \int_0^\tau \gamma(s)c(s)\,ds|\mathcal{F}.\right]$$

admits the stochastic integral representation

$$M(t) = x + \int_0^t \psi^*(s)\,dW_0(s), \qquad 0 \leq t \leq T,$$

for some $\mathbb{F}$-adapted process $\psi(\cdot)$ that satisfies $\int_0^T \|\psi(s)\|^2\,ds < \infty$ a.s. (e.g., Karatzas and Shreve (1998), Lemma 1.6.7). Define $\pi(t) \overset{\triangle}{=} (\sigma^*(t))^{-1}\psi(t)/M(t)$, $0 \leq t \leq T$, and check from (3.3) that $X(\cdot) = X^{x,\pi,c}(\cdot)$ a.e. on $[0,T] \times \Omega$. □

Remark 6.4. Note that the martingale $M(\cdot)$ is constant, and thus we have $\psi(\cdot) \equiv 0$, $\pi(\cdot) \equiv 0$ a.e. on the stochastic interval $[\![\tau, T]\!]$; in particular, $X^{x,\pi,c}(t,\omega) = B(\omega)e^{\int_{\tau(\omega)}^t r(u,\omega)\,du}$ a.e. on $[\![\tau, T]\!]$. In other words, at the stopping time τ all investment in the stock market ceases, and all proceeds are invested in the money-market from then on.

We have proved the following result.

PROPOSITION 6.5. *Under Assumption* 6.1, *for any $\tau \in \mathcal{S}$ we have*

$$V_\tau(x) = \inf_{\lambda>0}\left[\tilde{J}(\lambda;\tau) + \lambda x\right] = \tilde{J}(\mathcal{Y}_\tau(x);\tau) + x\mathcal{Y}_\tau(x), \tag{6.12}$$

and the supremum in (6.1) *is attained by the consumption strategy* $\hat{c}_\tau(t) = I_1\big(\mathcal{Y}_\tau(x)e^{\beta t} H(t)\big)\mathbf{1}_{[0,\tau)}(t)$ *and some portfolio* $\hat{\pi}_\tau(\cdot)$ *that satisfies* (6.11). *Moreover,*

(6.13)
$$V(x) = \sup_{\tau\in\mathcal{S}} V_\tau(x) = \sup_{\tau\in\mathcal{S}}\inf_{\lambda>0}\left[\tilde{J}(\lambda;\tau) + \lambda x\right] = \sup_{\tau\in\mathcal{S}}\left[\tilde{J}(\mathcal{Y}_\tau(x);\tau) + x\mathcal{Y}_\tau(x)\right].$$

Example 6.6 (logarithmic utility functions). $U_1(x) = \delta\log x,\ U_2(x) = \log x$ for $x > 0$ and some $\delta \in [0,1]$. In this case, Assumption 6.1 is satisfied, and we have $I_1(y) = \delta/y,\ \tilde{U}_1(y) = \delta\log\delta - \delta[1+\log y]$, and $I_2(y) = 1/y,\ \tilde{U}_2(y) = -1-\log y$. Hence, with

$$Q(t) \triangleq \int_0^t \theta^*(s)\,dW(s) + \int_0^t \left(r(s) + \frac{\|\theta(s)\|^2}{2} - \beta\right) ds$$

and with the convention $\delta\log\delta \equiv 0$ for $\delta = 0$, we have

$$\tilde{J}(\lambda;\tau) = \mathbb{E}\left[e^{-\beta\tau}\left(Q(\tau) - (1+\log\lambda)\right)\right] + \delta\cdot\mathbb{E}\int_0^\tau e^{-\beta t}\big(Q(t) - (1+\log\lambda)\big)\,dt$$
$$+\delta\log\delta\cdot\mathbb{E}\int_0^\tau e^{-\beta t}\,dt$$

for any stopping time τ. It develops that $\mathcal{X}_\tau(\lambda) = K_\tau/\lambda$ and thus $\mathcal{Y}_\tau(x) = K_\tau/x$, where

$$K_\tau \triangleq \mathbb{E}\left[e^{-\beta\tau} + \delta\int_0^\tau e^{-\beta t}\,dt\right].$$

From Proposition 6.5, the value-function of problem (5.3) is given by

$$V(x) = \sup_{\tau\in\mathcal{S}}\ \mathbb{E}\left[e^{-\beta\tau}\big\{\log\big(x/K_\tau\big) + Q(\tau)\big\} + \delta\cdot\int_0^\tau e^{-\beta t}\big\{\log\big(x/K_\tau\big) + Q(t)\big\}\,dt\right],$$

a quantity that is, in general, very difficult to compute. It is not even clear whether the supremum in this expression is attained (see Example 9.3 in this regard). However, *in the special case* $\beta = 0$ *and* $\delta = 0$, the above expression can be reduced significantly to

$$V(x) = \log x + \sup_{\tau\in\mathcal{S}}\mathbb{E}\int_0^\tau \left[r(u) + \frac{1}{2}\|\theta(u)\|^2\right] du$$

and amounts to solving a standard optimal stopping problem. The latter has the trivial solution $\tau^* \equiv T$ for $r(\cdot) \geq 0$.

Remark 6.7. A sufficient condition for Assumption 6.1 is that

$$I_1(y) + I_2(y) \leq k_1 + k_2 y^{-\alpha} \qquad \forall\ y \in (0,\infty) \tag{6.14}$$

holds for some constants $k_1 > 0$, $k_2 > 0$, and $\alpha > 0$. Indeed, under (6.14) we have

$$\mathbb{E}\left[\sup_{0\leq s\leq T}\left(H(s)\cdot I_j(\lambda e^{\beta s}H(s))\right)\right] \leq k_1\mathbb{E}\left[\sup_{0\leq s\leq T}(H(s))\right] + k_2\lambda^{-\alpha}\mathbb{E}\left[\sup_{0\leq s\leq T}(H(s))^{1-\alpha}\right]$$
$$< \infty$$

for $j = 1, 2$, as is easy to check using Hölder's inequality, Doob's maximal inequality, and the boundedness of market coefficients. This is because, for any $\rho \in \mathbf{R}$, there exist positive constants $C_1,\ C_2$ such that

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T}\big(H(t)\big)^{\rho}\right] &= \mathbf{E}\left[\sup_{0\le t\le T}\big(\gamma(t)Z_0(t)\big)^{\rho}\right] \ \le\ C_1\cdot\mathbb{E}\left[\sup_{0\le t\le T}\big(Z_0(t)\big)^{\rho}\right]\\ &\le C_1\cdot\mathbb{E}\left[\sup_{0\le t\le T}\left(e^{-\rho\int_0^t\theta^*(s)\,dW(s)-\frac{\rho^2}{2}\int_0^t\|\theta(s)\|^2\,ds}\right)\right.\\ &\qquad\qquad\left.\cdot\sup_{0\le t\le T}\left(e^{\frac{\rho(\rho-1)}{2}\int_0^t\|\theta(s)\|^2\,ds}\right)\right]\\ &\le C_2\cdot\mathbb{E}\left[\sup_{0\le t\le T}\left(e^{-\rho\int_0^t\theta^*(s)\,dW(s)-\frac{\rho^2}{2}\int_0^t\|\theta(s)\|^2\,ds}\right)\right]<\infty.\end{aligned}$$

7. Pure optimal stopping problems. The representation (6.13) for the solution of the utility maximization problem in section 5 is not entirely satisfactory. It is not clear how the quantities $\mathcal{Y}_\tau(x)$ are related to each other for different stopping times $\tau\in\mathcal{S}$, except in some very special cases. Furthermore, it is not easy to compute the last supremum in (6.13), or even to decide whether it is attained or not. All these points are illustrated in Example 6.6 of a logarithmic utility function. In this section, we shall try to convert the original problem into a family of *pure optimal stopping* problems, for which we can obtain a better understanding. To this end, we define, for every $\lambda\in(0,\infty)$, the *dual optimization problem*

(7.1)
$$\tilde{V}(\lambda)\overset{\triangle}{=}\sup_{\tau\in\mathcal{S}}\tilde{J}(\lambda;\tau)=\sup_{\tau\in\mathcal{S}}\ \mathbb{E}\left[\int_0^\tau e^{-\beta t}\tilde{U}_1\big(\lambda e^{\beta t}H(t)\big)\,dt+e^{-\beta\tau}\tilde{U}_2\big(\lambda e^{\beta\tau}H(\tau)\big)\right]$$

of pure optimal stopping type, in the notation of (6.5), (4.2), (2.6). To ensure that the problem of (7.1) is meaningful, we impose the following assumption throughout.

Standing Assumption 7.1. For any $\lambda\in(0,\infty)$ we have $\tilde{V}(\lambda)<\infty$, and there exists some stopping time $\hat{\tau}_\lambda$ which is optimal in (7.1), i.e., such that $\tilde{V}(\lambda)=\tilde{J}(\lambda;\hat{\tau}_\lambda)$.

Here and in what follows, we denote by $\hat{\mathcal{S}}_\lambda$ the set of stopping times that attain the supremum in (6.5) for every given $\lambda>0$. It follows from (6.4) that we have, in the notation of (7.1),

(7.2)
$$V(x)\le\sup_{\tau\in\mathcal{S}}\inf_{\lambda>0}\left[\tilde{J}(\lambda;\tau)+\lambda x\right]\le\inf_{\lambda>0}\left[\sup_{\tau\in\mathcal{S}}\tilde{J}(\lambda;\tau)+\lambda x\right]=\inf_{\lambda>0}\left[\tilde{V}(\lambda)+\lambda x\right].$$

We wish that the inequalities in (7.2) would always hold as equalities. Unfortunately, it turns out that the second inequality in (7.2) might be strict, depending on the coefficients of the model and on the initial capital x. We shall see this more clearly in the following sections.

Remark 7.2. Standing Assumption 7.1 holds if condition (5.4) is satisfied. This is because the continuous process $Y^\lambda(t)\overset{\triangle}{=}\int_0^t e^{-\beta s}\tilde{U}_1\big(\lambda e^{\beta s}H(s)\big)+e^{-\beta t}\tilde{U}_2(\lambda e^{\beta t}H(t))$, $0\le t\le T$, satisfies in this case $\mathbb{E}[\sup_{0\le t\le T}|Y^\lambda(t)|]<\infty$. Indeed, it is easy to check that (5.4) implies

(7.3)
$$\max\{\tilde{U}_1(y),\tilde{U}_2(y)\}\le k_1+k_3y^{-\alpha}\qquad\forall\ 0<\lambda<\infty$$

with $\alpha = \delta/(1-\delta)$, $k_3 = (1-\delta)(k_2\delta^\delta)^{1/(1-\delta)}$ (cf. KLSX (1991)), and it follows from Remark 6.7 that $\tilde{V}(\lambda) \le \mathbb{E}\left[\sup_{0\le t\le T} |Y^\lambda(t)|\right] \le k_4 + k_5\lambda^{-\alpha} \cdot \mathbb{E}\big[\sup_{0\le t\le T} \big(H(t)\big)^{-\alpha}\big] < \infty$. Standard results in the theory of optimal stopping (e.g., Theorem D.12 in Karatzas and Shreve (1998)) guarantee then the existence of an optimal stopping time.

8. Analysis of the optimal stopping problem. In this section we shall derive our main results for the optimization problem of (5.3), by first establishing several properties of the "dual" value function $\tilde{V}(\cdot)$ defined in (7.1). It is not a trivial matter to decide whether the value function $V(\cdot)$ of our "primal" problem (5.3) inherits the concavity of $U(\cdot)$. Indeed, even the continuity of $V(\cdot)$ is not quite clear a priori. However, properties of convexity and monotonicity are relatively straightforward for the dual value function $\tilde{V}(\cdot)$ of (7.1).

LEMMA 8.1. *The function $\tilde{V}(\cdot)$ of (7.1) is strictly convex and strictly decreasing. In particular, it is continuous and a.e. differentiable.*

Proof. For any $0 < \lambda_1 < \lambda_2 < \infty$, $0 < s < 1$, and $\lambda_0 \overset{\triangle}{=} s\lambda_1 + (1-s)\lambda_2$, we have $\tilde{V}(\lambda_2) = \tilde{J}(\lambda_2; \hat{\tau}_2) < \tilde{J}(\lambda_1; \hat{\tau}_2) \le \tilde{V}(\lambda_1)$ from Standing Assumption 7.1, where $\hat{\tau}_i \in \hat{\mathcal{S}}_{\lambda_i}$, $i = 0, 1, 2$ are optimal stopping times, and $\tilde{V}(\lambda_0) = \tilde{J}(\lambda_0; \hat{\tau}_0) < s\tilde{J}(\lambda_1; \hat{\tau}_0) + (1-s)\tilde{J}(\lambda_2; \hat{\tau}_0) \le s\tilde{V}(\lambda_1) + (1-s)\tilde{V}(\lambda_2)$. □

It follows from Lemma 8.1 that the right- and left-derivatives

$$\triangle^\pm \tilde{V}(\lambda) \overset{\triangle}{=} \lim_{h\to 0\pm} \frac{1}{h}[\tilde{V}(\lambda+h) - \tilde{V}(\lambda)] \tag{8.1}$$

of the convex function $\tilde{V}(\cdot)$ exist, and are finite for every $\lambda \in (0,\infty)$. Furthermore, the strict convexity of $\tilde{V}(\cdot)$ implies

$$\triangle^+ \tilde{V}(\lambda_1) < \triangle^- \tilde{V}(\lambda_2) \le \triangle^+ \tilde{V}(\lambda_2) \le 0 \quad \forall\, 0 < \lambda_1 < \lambda_2 < \infty, \tag{8.2}$$

and $\triangle^+\tilde{V}(\cdot)$ (respectively, $\triangle^-\tilde{V}(\cdot)$) is right- (respectively, left-) continuous.

LEMMA 8.2. *For every $\lambda \in (0,\infty)$ and any optimal stopping time $\hat{\tau}_\lambda \in \hat{\mathcal{S}}_\lambda$, we have*

$$\triangle^- \tilde{V}(\lambda) \le -\mathcal{X}_{\hat{\tau}_\lambda}(\lambda) \le \triangle^+ \tilde{V}(\lambda). \tag{8.3}$$

Proof. The convexity of $\tilde{U}_j(\cdot)$, $j = 1, 2$, gives

$$\tilde{U}_j'(y)(x-y) \le \tilde{U}_j(x) - \tilde{U}_j(y) \le \tilde{U}_j'(x)(x-y) \quad \forall\, 0 < x, y < \infty, \tag{8.4}$$

and for any real number h with $|h| < \lambda$ we obtain

$$\begin{aligned}
\tilde{V}(\lambda+h) - \tilde{V}(\lambda) &= \tilde{V}(\lambda+h) - \tilde{J}(\lambda; \hat{\tau}_\lambda) \ge \tilde{J}(\lambda+h; \hat{\tau}_\lambda) - \tilde{J}(\lambda; \hat{\tau}_\lambda) \\
&\ge h \cdot \mathbb{E}\left[\int_0^{\hat{\tau}_\lambda} H(t)\tilde{U}_1'\big(\lambda e^{\beta t} H(t)\big)\,dt + H(\hat{\tau}_\lambda)\tilde{U}_2'(\lambda e^{\beta\hat{\tau}_\lambda} H(\hat{\tau}_\lambda))\right] \\
&= -h\mathcal{X}_{\hat{\tau}_\lambda}(\lambda).
\end{aligned}$$

The last equality follows from (4.3) and the definition (6.6) of $\mathcal{X}_{\hat{\tau}}(\cdot)$. Letting $h \to 0$, we deduce for arbitrary $\lambda \in (0,\infty)$:

$$\begin{aligned}
\triangle^+\tilde{V}(\lambda) &= \lim_{h\to 0+} \frac{1}{h}[\tilde{V}(\lambda+h) - \tilde{V}(\lambda)] \\
&\ge -\mathcal{X}_{\hat{\tau}_\lambda}(\lambda) \ge \lim_{h\to 0-} \frac{1}{h}[\tilde{V}(\lambda+h) - \tilde{V}(\lambda)] \\
&= \triangle^-\tilde{V}(\lambda). \qquad \square
\end{aligned}$$

COROLLARY 8.3. *If $\tilde{V}(\cdot)$ is differentiable at $\lambda > 0$, then $\tilde{V}'(\lambda) = -\mathcal{X}_{\hat{\tau}_\lambda}(\lambda)$.*

LEMMA 8.4. *We have $\lim_{\lambda \downarrow 0} \triangle^{\pm} \tilde{V}(\lambda) = -\infty$. Moreover, if Assumption 6.1 holds, we also have $\lim_{\lambda \uparrow \infty} \triangle^{\pm} \tilde{V}(\lambda) = 0$.*

Proof. From the decrease of the function $I(\cdot)$, the monotone convergence theorem, and $I(0+) = \infty$, it follows that

$$\lim_{\lambda \downarrow 0} \mathcal{X}_{\hat{\tau}_\lambda}(\lambda) \geq \lim_{\lambda \downarrow 0} \mathbb{E}\left[\inf_{0 \leq s \leq T} \left(H(s) \cdot I_2\left(\lambda e^{\beta T} \sup_{0 \leq s \leq T} H(s)\right)\right)\right] = \infty,$$

and so by Lemma 8.2 and the inequality (8.2) we obtain $\lim_{\lambda \downarrow 0} \triangle^{\pm} \tilde{V}(\lambda) = -\infty$. Now suppose that Assumption 6.1 holds; we have then

$$\begin{aligned} 0 &\leq \lim_{\lambda \uparrow \infty} \mathcal{X}_{\hat{\tau}_\lambda}(\lambda) \\ &\leq \lim_{\lambda \uparrow \infty} \mathbb{E}\left[\sup_{0 \leq s \leq T} \left(H(s) \cdot I_2\left(\lambda e^{\beta s} H(s)\right)\right) + \int_0^T H(s) \cdot I_1\big(\lambda e^{\beta s} H(s)\big)\, ds\right] = 0 \end{aligned}$$

from the decrease of the functions $I_j(\cdot)$, the dominated convergence theorem, and $I_j(\infty) = 0$, $j = 1, 2$. It follows again from Lemma 8.2 and (8.2) that $\lim_{\lambda \uparrow \infty} \triangle^{\pm} \tilde{V}(\lambda) = 0$. □

We shall define, for each given $\lambda > 0$, the subset

$$\mathcal{G}_\lambda \overset{\triangle}{=} \left\{\mathcal{X}_{\hat{\tau}_\lambda}(\lambda) \Big/ \hat{\tau}_\lambda \text{ is optimal in (7.1), i.e., } \hat{\tau}_\lambda \in \hat{\mathcal{S}}_\lambda\right\} \tag{8.5}$$

of $\mathbb{R}^+$. It follows from (8.2) and (8.3) that the sets $\{\mathcal{G}_\lambda\}_{\lambda > 0}$ satisfy the following properties:

(i) $\mathcal{G}_\lambda$ is nonempty for every $\lambda > 0$,

(ii) $\mathcal{G}_\lambda \cap \mathcal{G}_\nu = \emptyset$, if $\lambda \neq \nu$, and

(iii) for any $0 < \nu < \lambda < \infty$ and $x \in \mathcal{G}_\lambda$, $y \in \mathcal{G}_\nu$, we have $x < y$.

Let us also introduce the set

$$\mathcal{G} \overset{\triangle}{=} \bigcup_{\lambda > 0} \mathcal{G}_\lambda. \tag{8.6}$$

We can state now the main result of the paper. This explains, in particular, when we can expect to find an optimal triple in (5.3) and to have equality in (7.2).

THEOREM 8.5. *For any $x \in \mathcal{G}$, the value $V(x)$ of (5.3) is attainable and we have*

$$V(x) = \inf_{\lambda > 0} \left[\tilde{V}(\lambda) + \lambda x\right]. \tag{8.7}$$

Conversely, for any $x \in (0, \infty)$ that satisfies (8.7) and for which the value $V(x)$ of (5.3) is attainable, we have $x \in \mathcal{G}$, provided that Assumption 6.1 holds.

Proof. Suppose $x \in \mathcal{G}_\nu$ for some $\nu > 0$, and $x = \mathcal{X}_{\hat{\tau}_\nu}(\nu)$ for some stopping time $\hat{\tau}_\nu \in \hat{\mathcal{S}}_\nu$ which is optimal in (7.1) with $\lambda = \nu$, i.e., with

$$\tilde{V}(\nu) = \tilde{J}(\nu; \hat{\tau}_\nu) = \mathbb{E}\left[\int_0^{\hat{\tau}_\nu} e^{-\beta t} \tilde{U}_1\big(\nu e^{\beta t} H(t)\big)\, dt + e^{-\beta \hat{\tau}_\nu} \tilde{U}_2\big(\nu e^{\beta \hat{\tau}_\nu} H(\hat{\tau}_\nu)\big)\right]. \tag{8.8}$$

Then we claim

$$V(x) = \tilde{V}(\nu) + \nu x = \inf_{\lambda > 0} [\tilde{V}(\lambda) + \lambda x]. \tag{8.9}$$

Indeed, by Lemma 8.2, we have $-x \in [\triangle^- \tilde{V}(\nu), \triangle^+ \tilde{V}(\nu)]$, so that $\tilde{V}(\lambda) - \tilde{V}(\nu) \geq (-x) \cdot (\lambda - \nu)$, or, equivalently, $\tilde{V}(\lambda) + \lambda x \geq \tilde{V}(\nu) + \nu x \quad \forall \lambda > 0$.

Since $x = \mathcal{X}_{\hat{\tau}_\nu}(\nu) = \mathbb{E}\big[H(\hat{\tau}_\nu) I_2(\nu e^{\beta \hat{\tau}_\nu} H(\hat{\tau}_\nu)) + \int_0^{\hat{\tau}_\nu} H(t) I_1\big(\nu e^{\beta t} H(t)\big)\, dt\big]$, it follows from Lemma 6.3 and Lemma 6.2 that there exists a portfolio process $\hat{\pi}(\cdot)$ with $X^{x,\hat{\pi},\hat{c}}(\hat{\tau}_\nu) = I_2(\nu e^{\beta \hat{\tau}_\nu} H(\hat{\tau}_\nu))$, where $\hat{c}(t) \overset{\triangle}{=} I_1\big(\nu e^{\beta t} H(t)\big) \mathbf{1}_{[0,\tau)}(t)$. The expected utility $J(x; \hat{\pi}, \hat{c}, \hat{\tau}_\nu)$, under the portfolio/consumption strategy $(\hat{\pi}, \hat{c})$ and the stopping time $\hat{\tau}_\nu$, is thus

$$\begin{aligned}
V(x) \geq J(x; \hat{\pi}, \hat{c}, \hat{\tau}_\nu) &= \mathbb{E}\left[\int_0^{\hat{\tau}_\nu} e^{-\beta t} U_1\big(I_1(\nu e^{\beta t} H(t))\big)\, dt + e^{-\beta \hat{\tau}_\nu} U_2(I_2(\nu e^{\beta \hat{\tau}_\nu} H(\hat{\tau}_\nu)))\right] \\
&= \mathbb{E}\left[\int_0^{\hat{\tau}_\nu} e^{-\beta t} \tilde{U}_1\big(\nu e^{\beta t} H(t)\big) + e^{-\beta \hat{\tau}_\nu} \tilde{U}_2(\nu e^{\beta \hat{\tau}_\nu} H(\hat{\tau}_\nu))\right] \\
&\qquad + \nu \cdot \mathbb{E}\left[H(\hat{\tau}_\nu) X^{x,\hat{\pi},\hat{c}}(\hat{\tau}_\nu) + \int_0^{\hat{\tau}_\nu} H(t) \hat{c}(t)\, dt\right] \\
&= \tilde{V}(\nu) + \nu x = \inf_{\lambda > 0} [\tilde{V}(\lambda) + \lambda x],
\end{aligned}$$

and (8.9) follows then from (7.2). In particular, the triple $(\hat{\pi}, \hat{c}, \hat{\tau}_\nu)$ in $\mathcal{A}(x)$ is optimal for the original optimization problem of (5.3).

Conversely, suppose that (8.7) holds for some positive real number x, for which the value $V(x)$ of (5.3) is attained by some optimal triple $(\pi^*, c^*, \tau^*) \in \mathcal{A}(x)$. In other words,

$$V(x) = \inf_{\lambda > 0} [\, \tilde{V}(\lambda) + \lambda x\,] = J(x; \pi^*, c^*, \tau^*) \leq V_{\tau^*}(x) \tag{8.10}$$

in the notation of (6.1). Suppose also that Assumption 6.1 holds. By Lemma 8.1 the function $\lambda \longmapsto \tilde{V}(\lambda) + \lambda x =: G(\lambda)$ is strictly convex, with $G(0+) = \tilde{V}(0+)$ and $G(\infty) = \infty$. Thus, either there exists a unique $\nu > 0$ such that

$$\tilde{V}(\nu) + \nu x = \inf_{\lambda > 0} [\, \tilde{V}(\lambda) + \lambda x\,], \tag{8.11}$$

or else we have $\tilde{V}(0+) \leq \tilde{V}(\lambda) + \lambda x \quad \forall \lambda > 0$. This latter possibility can be ruled out easily; it cannot hold if $\tilde{V}(0+) = \infty$, whereas with $\tilde{V}(0+) < \infty$ it leads to $\lim_{\lambda \downarrow 0}\big(-\triangle^+ \tilde{V}(\lambda)\big) \leq x$, which is impossible by Lemma 8.4. Therefore, (8.11) holds for a unique $\nu > 0$ and leads, with (8.10) and Proposition 6.4, to

$$V(x) = \tilde{V}(\nu) + \nu x \geq \tilde{J}(\nu; \tau^*) + \nu x \geq \inf_{\lambda > 0} [\tilde{J}(\lambda; \tau^*) + \lambda x] = V_{\tau^*}(x) \geq V(x). \tag{8.12}$$

We obtain $\tilde{V}(\nu) = \tilde{J}(\nu; \tau^*)$ as well as $\tilde{J}(\nu; \tau^*) + \nu x = \inf_{\lambda > 0} [\tilde{J}(\lambda; \tau^*) + \lambda x]$ from (8.10), (8.12), or, equivalently, $\tau^* \in \hat{\mathcal{S}}_\nu$ and $\nu = \mathcal{Y}_{\tau^*}(x)$. Thus $x = \mathcal{X}_{\tau^*}(\nu) \in \mathcal{G}_\nu$, which concludes the proof. □

COROLLARY 8.6. *Under Assumption* 6.1*, for any* $x \notin \mathcal{G} \equiv \bigcup_{\lambda > 0} \mathcal{G}_\lambda$*, we have the strict inequality ("duality gap")* $V(x) < \inf_{\lambda > 0} [\tilde{V}(\lambda) + \lambda x]$.

COROLLARY 8.7. *Under Assumption* 6.1*, and if* $\tilde{V}(\cdot)$ *is differentiable everywhere, the value* $V(x)$ *of* (5.3) *is attainable and* (8.7) *holds for every* $x \in (0, \infty)$.

Proof. Since every differentiable convex function is continuously differentiable (cf. Rockafellar (1970), Corollary 25.5.1), $\tilde{V}'(\cdot)$ is continuous. By Lemma 8.4, the range of $\tilde{V}'(\cdot)$ is $(-\infty, 0)$. It follows from Corollary 8.3 that $\mathcal{G} = (0, \infty)$, and Theorem 8.5 applies. □

COROLLARY 8.8. *Under Assumption* 6.1*, suppose that for any* $\lambda \in (0, \infty)$ *there exist two sequences* $\{\lambda_n^{(\pm)}\}$ *with* $\lambda_n^{(+)} \downarrow \lambda$, $\lambda_n^{(-)} \uparrow \lambda$, *as well as stopping times* $\hat{\tau} \in \hat{\mathcal{S}}_\lambda$, $\hat{\tau}_n^{(\pm)} \in \hat{\mathcal{S}}_{\lambda_n^{(\pm)}}$ *such that* $\hat{\tau}_n^{(\pm)} \to \hat{\tau}$ *a.s.; then the value* $V(x)$ *of* (5.3) *is attainable and* (8.7) *holds for every* $x > 0$*.*

Proof. By Corollary 8.7, we need only show that $\tilde{V}(\cdot)$ is differentiable everywhere. From (8.4) and (4.3) we have

$$\begin{aligned}
&\tilde{V}(\lambda_n^{(\pm)}) - \tilde{V}(\lambda) \le \tilde{J}(\lambda_n^{(\pm)}; \hat{\tau}_n^{(\pm)}) - \tilde{J}(\lambda; \hat{\tau}_n^{(\pm)}) \\
&\le -(\lambda_n^{(\pm)} - \lambda) \cdot \mathbb{E}\left[\int_0^{\hat{\tau}_n^{\pm}} H(t) I_1\big(\lambda_n^{(\pm)} e^{\beta t} H(t)\big)\, dt + H(\hat{\tau}_n^{(\pm)}) I_2(\lambda_n^{(\pm)} e^{\beta \hat{\tau}_n^{(\pm)}} H(\hat{\tau}_n^{(\pm)}))\right] \\
&= -(\lambda_n^{(\pm)} - \lambda) \cdot \mathcal{X}_{\hat{\tau}_n^{(\pm)}}(\lambda_n^{(\pm)}),
\end{aligned}$$

which implies

$$\begin{aligned}
\triangle^+ \tilde{V}(\lambda) &= \lim_{\lambda_n^{(+)} \downarrow \lambda} \frac{\tilde{V}(\lambda_n^{(+)}) - \tilde{V}(\lambda)}{\lambda_n^{(+)} - \lambda} \le \limsup_{\lambda_n^{(+)} \downarrow \lambda} \big(-\mathcal{X}_{\hat{\tau}_n^{(+)}}(\lambda_n^{(+)})\big) = -\mathcal{X}_{\hat{\tau}}(\lambda), \\
\triangle^- \tilde{V}(\lambda) &= \lim_{\lambda_n^{(-)} \uparrow \lambda} \frac{\tilde{V}(\lambda_n^{(-)}) - \tilde{V}(\lambda)}{\lambda_n^{(-)} - \lambda} \ge \liminf_{\lambda_n^{(-)} \downarrow \lambda} \big(-\mathcal{X}_{\hat{\tau}_n^{(-)}}(\lambda_n^{(-)})\big) = -\mathcal{X}_{\hat{\tau}}(\lambda)
\end{aligned}$$

by the dominated convergence theorem. From (8.2), $\tilde{V}'(\lambda) = \triangle^+ \tilde{V}(\lambda) = \triangle^- \tilde{V}(\lambda) = -\mathcal{X}_{\hat{\tau}}(\lambda)$. □

Corollaries 8.7 and 8.8 provide simple sufficient (but *not* necessary) conditions, under which there is no "duality gap" in (7.2), i.e., its leftmost and rightmost members are equal. The following proposition will characterize this kind of interchangeability of "inf" and "sup" operations from another point of view, namely, the concavity of the "primal" value function $V(\cdot)$.

PROPOSITION 8.9. *Under Assumption* 6.1*, the following two statements are equivalent:*

(A) $V(\cdot)$ *is concave on* $(0, +\infty)$*,*

(B) $V(x) = \inf_{\lambda > 0} [\tilde{V}(\lambda) + \lambda x]$ *holds for every* $x \in (0, \infty)$*.*

Proof of (B) $\Longrightarrow$ (A). Under condition (B), the number $-V(x)$ is the pointwise supremum of the affine functions $g(\lambda) = -\lambda x - \mu$ such that (x, μ) belongs to the epigraph of $\tilde{V}(\cdot)$. Hence $-V(\cdot)$ is a convex function (Rockafellar (1970), Theorem 12.1), or, equivalently, $V(\cdot)$ is concave.

Proof of (A) $\Longrightarrow$ (B). By Lemma 8.4 and (8.2), it is sufficient to show that for any $(\nu, x) \in (0, \infty) \times (0, \infty)$ such that $-\triangle^+ \tilde{V}(\nu) \le x \le -\triangle^- \tilde{V}(\nu)$, we have $V(x) = \tilde{V}(\nu) + \nu x$.

Let $x_0 \overset{\triangle}{=} -\triangle^+ \tilde{V}(\nu)$, $x_1 \overset{\triangle}{=} -\triangle^- \tilde{V}(\nu)$. Since $\tilde{V}(\cdot)$ is strictly convex and differentiable except on a countable set, we can find a sequence of positive real numbers $\{\lambda_n\}$, such that $\lambda_n \downarrow \nu$ as $n \to \infty$, and $\tilde{V}(\cdot)$ is differentiable at each λ_n. Define $y_n \overset{\triangle}{=} -\tilde{V}'(\lambda_n)$. It follows from the right-continuity of $\triangle^+ \tilde{V}(\cdot)$ that $-y_n = \triangle^+ \tilde{V}(\lambda_n) \downarrow \triangle^+ \tilde{V}(\nu) = -x_0$. However, Theorem 8.5 and Corollary 8.3 assert that

$$V(y_n) = \inf_{\lambda > 0} [\tilde{V}(\lambda) + \lambda y_n] = \tilde{V}(\lambda_n) + \lambda_n y_n. \tag{8.13}$$

Letting $n \to \infty$, we obtain

$$V(x_0) = \tilde{V}(\nu) + \nu x_0, \tag{8.14}$$

thanks to the continuity of $V(\cdot)$ (which is concave by assumption (A)) and of $\tilde{V}(\cdot)$ (which is convex by Lemma 8.1). Furthermore, we claim that $\triangle^- V(x_0) \le \nu$. Indeed, it follows from (8.13) and (8.14) that

$$\begin{aligned} V(y_n) - V(x_0) &= \tilde{V}(\lambda_n) + \lambda_n y_n - \tilde{V}(\nu) - \nu x_0 \ge \triangle^+ \tilde{V}(\nu)(\lambda_n - \nu) + \lambda_n y_n - \nu x_0 \\ &= \lambda_n (y_n - x_0), \end{aligned}$$

and hence

$$\triangle^- V(x_0) = \lim_{n\to\infty} \frac{V(y_n) - V(x_0)}{y_n - x_0} \le \lim_{n\to\infty} \lambda_n = \nu. \tag{8.15}$$

Similarly, we obtain

$$V(x_1) = \tilde{V}(\nu) + \nu x_1 \qquad \text{and} \qquad \triangle^+ V(x_1) \ge \nu. \tag{8.16}$$

However, $\triangle^- V(x_0) \ge \triangle^+ V(x_1)$ holds from the concavity of $V(\cdot)$. It follows from (8.15) and (8.16) that $\triangle^- V(x_0) = \nu = \triangle^+ V(x_1)$, or equivalently, $\triangle^- V(x) = \triangle^+ V(x) = V^{'}(x) = \nu \ \ \forall\ x_0 \le x \le x_1$. It is clear now that $V(x) = \tilde{V}(\nu) + \nu x = \inf_{\lambda>0} [\tilde{V}(\lambda) + \lambda x]$ holds for any $x_0 \le x \le x_1$.

9. Examples. Using the technique developed in the preceding section, we study here several examples, including one which shows that optimal strategies need not always exist (see Example 9.3). The first of these examples can also be treated using the methods of section 6, but for the second and third examples the methodology of section 8 is indispensable. The reader of this section should not fail to notice the rarity of a setting where utility functions of power-type are much easier to handle than logarithmic ones.

Example 9.1 (utility functions of power-type). $U_j(x) = x^\alpha/\alpha$, where $0 < \alpha < 1,\ j = 1, 2$. In this case, condition (5.4) is satisfied and we have $I_j(y) = y^{-1/(1-\alpha)}$ and $\tilde{U}_j(y) = y^{-\gamma}/\gamma$ with $\gamma = \alpha/(1-\alpha),\ j = 1, 2$, so that Assumption 6.1 is also satisfied (see Remark 6.7) and implies $K < \infty$ in (9.2) below. We obtain easily

(9.1)
$$\tilde{V}(\lambda) = \sup_{\tau\in\mathcal{S}} \mathbb{E}\left[\int_0^\tau e^{-\beta t}\tilde{U}_1\big(\lambda e^{\beta t} H(t)\big)\, dt + e^{-\beta\tau}\tilde{U}_2(\lambda e^{\beta\tau} H(\tau))\right] = \frac{K}{\gamma}\lambda^{-\gamma},$$

with

$$K \overset{\triangle}{=} \sup_{\tau\in\mathcal{S}} K_\tau := \sup_{\tau\in\mathcal{S}} \mathbb{E}\left[\int_0^\tau e^{-(1+\gamma)\beta t}\big(H(t)\big)^{-\gamma}\, dt + e^{-(1+\gamma)\beta\tau}\big(H(\tau)\big)^{-\gamma}\right]. \tag{9.2}$$

Clearly $\tilde{V}(\cdot)$ is differentiable everywhere, and it follows from Corollary 8.7 that $V(x) = \inf_{\lambda>0} [\tilde{V}(\lambda) + \lambda x] = K^{1-\alpha}\, x^\alpha/\alpha$. In other words, with utility functions of power-type, the original optimization problem is reduced to the pure optimal stopping problem (9.2). We can arrive at this conclusion also using Proposition 6.5, since we have $\mathcal{X}_\tau(\lambda) = K_\tau \lambda^{-1/(1-\alpha)}$, $\mathcal{Y}_\tau(x) = (K_\tau/x)^{1-\alpha}$, $\tilde{J}(\lambda; \tau) = \frac{K_\tau}{\gamma}\lambda^{-\gamma}$, and thus $V(x) = \frac{x^\alpha}{\alpha} K^{1-\alpha}$ from (6.12), (6.13).

The optimal stopping time $\hat{\tau}$ for the original problem is also optimal for the problem of (9.2); the corresponding optimal consumption $\hat{c}(\cdot)$ and wealth-level $X^{x,\hat{\pi},\hat{c}}(\hat{\tau}) \equiv \xi^x(\hat{\tau})$ are given as

$$\hat{c}(t) = \frac{x}{K} e^{-\frac{\beta t}{1-\alpha}} \big(H(t)\big)^{-\frac{1}{1-\alpha}}, \quad 0 \le t \le \hat{\tau}, \qquad \xi^x(\hat{\tau}) = \frac{x}{K} e^{-\frac{\beta\hat{\tau}}{1-\alpha}} \left(H(\hat{\tau})\right)^{-\frac{1}{1-\alpha}}$$

by (6.11), and the optimal portfolio process $\hat{\pi}(\cdot)$ can then be obtained from Lemma 6.3.

It is straightforward to check that $\hat{\tau} \equiv 0,\ K = 1$ if

$$\beta \ge \gamma \left[\frac{r(t)}{1+\gamma} + \frac{1}{2}\|\theta(t)\|^2\right] \qquad \forall\ 0 \le t \le T$$

holds a.s., and that $\hat{\tau} \equiv T,\ K = K_T$ if

$$\beta \le \gamma \left[\frac{r(t)}{1+\gamma} + \frac{1}{2}\|\theta(t)\|^2\right] \qquad \forall\ 0 \le t \le T$$

holds a.s. This observation provides a complete solution to the optimal stopping problem of (9.2) in the case of constant interest-rate $r(t) \equiv r \in \mathbb{R}$ and relative risk $\theta(t) \equiv \theta \in \mathbb{R}^m$; in particular, if $\beta = \gamma\big(\frac{r}{1+\gamma} + \frac{\|\theta\|^2}{2}\big)$, every stopping time $\tau \in \mathcal{S}_{0,T}$ is optimal in (9.2) and $K = K_\tau = 1$.

Example 9.2 (logarithmic utility function from terminal wealth only, with $\beta > 0$). $U_2(x) = \log x$ for $x > 0$ and $U_1(\cdot) \equiv 0$. This is the setting of Example 6.6 with $\delta = 0$; Assumption 6.1 is now satisfied trivially.

(i) $b(\cdot) \equiv r(\cdot)\mathbf{1}_m$. Since we have $\theta(\cdot) \equiv 0$ in this case, it follows that $\tilde{J}(\lambda;\tau) = -\mathbb{E}[e^{-\beta\tau}(1 + \log\lambda + A(\tau))]$, where

$$A(t,\omega) \overset{\triangle}{=} \beta t - \int_0^t r(s,\omega)\,ds \qquad \forall\ 0 \le t \le T.$$

We claim that

if $\frac{dA(t,\omega)}{dt} - \beta A(t,\omega)$ *is strictly increasing for almost every* $\omega \in \Omega$ *(e.g., if* $r(t) \equiv r > \beta$*), then* (8.7) *holds.*

In order to check this, let $\hat{\tau}_\lambda \overset{\triangle}{=} \inf\{t \ge 0\ /\ \frac{dA(t)}{dt} - \beta A(t) \ge \beta(1 + \log\lambda)\} \wedge T$. It is not difficult to see that $\hat{\tau}_\lambda \in \hat{\mathcal{S}}_\lambda$, since $-e^{-\beta\hat{\tau}_\lambda(\omega)}(1 + \log\lambda + A(\hat{\tau}_\lambda(\omega),\omega))$ is then the minimum of the path $e^{-\beta t}(1 + \log\lambda + A(t,\omega)),\ 0 \le t \le T$. Moreover, the condition of Corollary 8.8 is satisfied, and $\hat{\tau}_{\lambda_n} \to \hat{\tau}_\lambda$ if $\lambda_n \to \lambda$. It follows that

$$V(x) = \inf_{\lambda > 0}\,[\tilde{J}(\lambda;\hat{\tau}_\lambda) + \lambda x].$$

The optimal stopping time for the original optimization problem is $\hat{\tau} \equiv \hat{\tau}_{\hat{\lambda}}$, where $\hat{\lambda} > 0$ attains the infimum in the above expression. The corresponding optimal level of wealth $X^{x,\hat{\pi},0}(\hat{\tau}) \equiv \xi^x(\hat{\tau})$ is given by (6.11) as

$$\xi^x(\hat{\tau}) = \frac{x}{\mathbb{E}\left(e^{-\beta\hat{\tau}}\right)}\ e^{\int_0^{\hat{\tau}} r(s)\,ds - \beta\hat{\tau}},$$

and the optimal portfolio process $\hat{\pi}(\cdot)$ can be derived from Lemma 6.3.

(ii) A general result for the logarithmic utility function (from terminal wealth only) seems difficult to obtain, as we saw already in Example 6.6. Nevertheless, using the theory of section 8, we shall establish the following property:

$$\text{(9.3)} \quad \left\{ \begin{array}{l} \textit{$V(x)$ is attainable and } (8.7) \textit{ holds for every } x>0, \textit{ if there exists a} \\ \text{unique } \textit{optimal stopping time solving problem } (7.1) \textit{ for every } \lambda>0 \end{array} \right\}.$$

The rest of this paragraph is dedicated to the proof of statement (9.3). Consider the continuous process

$$Y^{\lambda}(t) \overset{\triangle}{=} e^{-\beta t}\tilde{U}(\lambda e^{\beta t}H(t)) = -e^{-\beta t}(1+\log\lambda+\beta t+\log H(t))$$

and its Snell envelope, given as a right continuous with limits from the left (RCLL) modification of the supermartingale

$$Z^{\lambda}(t) \overset{\triangle}{=} \ \mathrm{esssup}_{\tau\in\mathcal{S}_{t,T}}\,\mathbb{E}[Y^{\lambda}(\tau)|\mathcal{F}_t], \qquad 0\le t\le T,$$

with $Z^{\lambda}(0) = \sup_{\tau\in\mathcal{S}_{0,T}}\mathbb{E}Y^{\lambda}(\tau) = \tilde{V}(\lambda)$. We claim that $Z^{\lambda}(\cdot)$ is actually continuous. Indeed, since the random variable $\sup_{0\le t\le T}Y^{\lambda}(t)$ is integrable by Remark 7.2, the Snell envelope $Z^{\lambda}(\cdot)$ admits the Doob–Meyer decomposition $Z^{\lambda}(\cdot) = Z^{\lambda}(0)+M^{\lambda}(\cdot)-A^{\lambda}(\cdot)$ (Karatzas and Shreve (1998), Theorem D.13), where $M^{\lambda}(\cdot)$ is an RCLL martingale and $A^{\lambda}(\cdot)$ is continuous and nondecreasing. But any RCLL martingale of the Brownian filtration is continuous (Karatzas and Shreve (1991), Problem 3.4.16); hence $M^{\lambda}(\cdot)$ is continuous, and thus so is $Z^{\lambda}(\cdot)$. The stopping time $\tau^*_{\lambda} \overset{\triangle}{=} \inf\left\{t\in[0,T)\ /\ Z^{\lambda}(t)=Y^{\lambda}(t)\right\}\wedge T$ is the *smallest* optimal stopping time in $\hat{\mathcal{S}}_{\lambda}$, whereas the stopping time $\rho^*_{\lambda} \overset{\triangle}{=} \inf\left\{t\in[0,T)\ /\ A^{\lambda}(t)>0\right\}\wedge T$ is the *largest* optimal stopping time in $\hat{\mathcal{S}}_{\lambda}$ (Karatzas and Shreve (1998), Theorems D.12 and D.9; El Karoui (1981)). In particular, the uniqueness property (9.3) amounts to the statement $\mathbb{P}[\tau^*_{\lambda}=\rho^*_{\lambda}]=1\ \ \forall\, 0<\lambda<\infty$.

Moreover, $\lambda\mapsto\tau^*_{\lambda}$ is *increasing*; that is, for any $\lambda\ge\nu$ we have $\tau^*_{\lambda}\ge\tau^*_{\nu}$ a.s. To see this, observe that $Y^{\lambda}(t)-Y^{\nu}(t) = -e^{-\beta t}\log(\lambda/\nu)$ and obtain

$$\begin{aligned} Z^{\lambda}(t)-Z^{\nu}(t) &= \mathrm{esssup}_{\tau\in\mathcal{S}_{t,T}}\,\mathbb{E}[Y^{\lambda}(\tau)|\mathcal{F}_t] - \mathrm{esssup}_{\tau\in\mathcal{S}_{t,T}}\,\mathbb{E}\left[Y^{\lambda}(\tau)+e^{-\beta\tau}\log\left(\frac{\lambda}{\nu}\right)\middle|\mathcal{F}_t\right] \\ &\ge \mathrm{esssup}_{\tau\in\mathcal{S}_{t,T}}\,\mathbb{E}[Y^{\lambda}(\tau)|\mathcal{F}_t] - \mathrm{esssup}_{\tau\in\mathcal{S}_{t,T}}\,\mathbb{E}[Y^{\lambda}(\tau)|\mathcal{F}_t] - e^{-\beta t}\log\left(\frac{\lambda}{\nu}\right) \\ &= Y^{\lambda}(t)-Y^{\nu}(t) \end{aligned}$$

a.s. for any given $0\le t\le T$. By the continuity of $Z(\cdot)$ and $Y(\cdot)$, it follows that

$$\mathbb{P}\left[Z^{\lambda}(t)-Y^{\lambda}(t)\ge Z^{\nu}(t)-Y^{\nu}(t)\ \ \forall\, 0\le t\le T\right]=1,$$

which implies that $\tau^*_{\lambda}\ge\tau^*_{\nu}$ a.s., since $Z(\cdot)$ always dominates $Y(\cdot)$. It is not difficult to see that $\tau^{\pm}_{\lambda} \overset{\triangle}{=} \lim_{n\to\infty}\tau^*_{\lambda\pm\frac{1}{n}}$ are stopping times, thanks to the continuity of the filtration $\mathbb{F}$. Moreover, they both belong to $\hat{\mathcal{S}}_{\lambda}$, which is an easy exercise on the dominated convergence theorem (we omit the details).

Now we can prove our assertion (9.3). Clearly it must hold that $\tau^*_{\lambda}=\tau^+_{\lambda}=\tau^-_{\lambda}$ by uniqueness of optimal stopping time. It follows from Corollary 8.8 that $V(x)$ is attainable and (8.7) holds for every $x>0$.

Example 9.3 (a case where no optimal strategy exists). We present now an example which shows that optimal strategies *need not* always exist for every initial capital $x \in (0, \infty)$.

Consider the logarithmic utility functions as in Example 6.6 with $\delta = 0$, i.e., $U_1(\cdot) \equiv 0$ and $U_2(x) = \log x$, discount factor $\beta = 1$, and model parameters $m = 1$, $r(\cdot) \equiv 0$, $b(\cdot) \equiv 0$, $\sigma(\cdot) \equiv 1$ in (2.1), (2.2). In this case we may take $c(\cdot) \equiv 0$ since there is no utility from consumption, and for a given initial capital $x > 0$ the wealth-process $X^{x,\pi}(\cdot) \equiv X^{x,\pi,0}(\cdot)$ corresponding to a portfolio $\pi(\cdot)$ satisfies

$$dX^{x,\pi}(t) = X^{x,\pi}(t)\pi(t)\,dW(t), \qquad X^{x,\pi}(0) = x. \tag{9.4}$$

It is not difficult to check that

$$\tilde{V}(\lambda) = \sup_{\tau \in \mathcal{S}} \tilde{J}(\lambda;\tau) = \sup_{\tau \in \mathcal{S}} \mathbb{E}\left[-e^{-\tau}(1 + \log\lambda + \tau)\right] = \max_{0 \le t \le T} F(\lambda;t), \tag{9.5}$$

where $F(\lambda;t) \triangleq -e^{-t}(1 + \log\lambda + t)$, $\lambda > 0$, $t > 0$. Note that the function $t \mapsto F(\lambda;t)$ attains its maximum on the interval $[0,T]$ at one of its endpoints; that is, $\max_{0 \le t \le T} F(\lambda;t) = \max\{F(\lambda;0),\, F(\lambda;T)\}$, since $e^t \frac{dF}{dt}(\lambda;t) = \log\lambda + t$ is increasing. It follows then from (9.5) that

$$\tilde{V}(\lambda) = \left\{ \begin{array}{cc} -(1+\log\lambda), & 0 < \lambda \le \lambda^*(T) \\ -e^{-T}(1+\log\lambda + T), & \lambda^*(T) \le \lambda < \infty \end{array} \right\}, \tag{9.6}$$

where $\lambda^*(s) \triangleq \exp\left\{\left(s/(e^s - 1)\right) - 1\right\} \in (0,1)$ is determined by the equation

$$1 + \log\lambda^*(s) = e^{-s}(1 + \log\lambda^*(s) + s). \tag{9.7}$$

Clearly, $\tilde{V}(\cdot)$ is *not* differentiable at $\lambda = \lambda^*(T)$. Moreover, it is easy to verify that $\mathcal{G}_\lambda = \{1/\lambda\}$ for $0 < \lambda < \lambda^*(T)$ and that $\mathcal{G}_\lambda = \left\{e^{-T}/\lambda\right\}$ for $\lambda > \lambda^*(T)$, and thus

$$\mathcal{G} = \bigcup_{\lambda > 0} \mathcal{G}_\lambda = \left(0, x_0(T)\right] \cup \left[x_1(T), \infty\right) \tag{9.8}$$

with $x_0(s) \triangleq \frac{e^{-s}}{\lambda^*(s)} \in (0,1)$ and $x_1(s) \triangleq \frac{1}{\lambda^*(s)} \in (1,\infty)$; we omit the details of these computations. It should be noted that $x_1(\cdot)$ is increasing with $x_1(0+) = 1$, $x_1(\infty) = e$, whereas $x_0(\cdot)$ is decreasing with $x_0(0+) = 1$, $x_0(\infty) = 0$.

Now with $V_0(x) \triangleq e^{-T}\log x$ and $V_1(x) \triangleq \log x$, let us consider the concave function

$$\begin{aligned} G(x) &\triangleq \inf_{\lambda > 0}[\tilde{V}(\lambda) + \lambda x] \\ &= \left\{ \begin{array}{cc} V_0(x), & 0 < x \le x_0(T) \\ V_0(x_0(T))\frac{x_1(T) - x}{x_1(T) - x_0(T)} + V_1(x_1(T))\frac{x - x_0(T)}{x_1(T) - x_0(T)}, & x_0(T) < x < x_1(T) \\ V_1(x), & x_1(T) \le x < \infty \end{array} \right\} \end{aligned}$$

(see Remark 9.4 for discussion). We have $V(x) = G(x)$ for $x \in \mathcal{G}$ from Theorem 8.5, or

$$V(x) = \left\{ \begin{array}{cc} V_0(x), & 0 < x \le x_0(T) \\ V_1(x), & x_1(T) \le x < \infty \end{array} \right\}. \tag{9.9}$$

In particular, the optimal strategy is to keep all the wealth in the money-market (i.e., $\pi(\cdot) \equiv 0$) and to wait until the terminal time T, if the initial capital x is in $(0, x_0(T)]$, whereas the optimal strategy for $x \geq x_1(T)$ is to stop immediately.

But how about an initial capital $x \in (\, x_0(T), x_1(T)\,)$? From Theorem 8.5 and Proposition 8.9, we know that either $V(x) < G(x)$ for some $x \in (\, x_0(T), x_1(T)\,)$ (which will give us a nonconcave value function $V(\cdot)$), or else $V(x) \equiv G(x) \;\; \forall x \in (\, x_0(T), x_1(T)\,)$ (in which case no optimal strategy exists).

We claim that the latter is the case. In other words, $V(x) \equiv G(x) \;\; \forall x \in \mathbf{R}^+$, but *no* optimal strategy exists for $x \in (\, x_0(T), x_1(T)\,)$. Actually, for every $x \in (\, x_0(T), x_1(T)\,)$, a maximizing sequence of strategy pairs $\{(\pi_n, \tau_n)\}_{n=1}^{\infty}$ can be constructed so that $J(x; \pi_n, \tau_n) \to G(x)$ as $n \to \infty$; this proves, in particular, that $V(\cdot) \equiv G(\cdot)$ on $\big(x_0(T), x_1(T)\big)$. Indeed, consider the wealth-process $dX^{x,n}(t) = nX^{x,n}(t)\, dW(t)$, $X^{x,n}(0) = x$, and let

$$T_0^n \overset{\triangle}{=} \inf\big\{t \geq 0 \;/\; X^{x,n}(t) \leq x_0(T-t)\big\} \wedge T, \tag{9.10}$$

$$T_1^n \overset{\triangle}{=} \inf\big\{t \geq 0 \;/\; X^{x,n}(t) \geq x_1(T-t)\big\} \wedge T. \tag{9.11}$$

Recall $x_0(0+) = x_1(0+) = 1$, so that $T_0^n \wedge T_1^n < T$ holds a.s. We define the portfolio/stopping time pair (π_n, τ_n) by

$$\pi_n(t) \overset{\triangle}{=} n \cdot 1_{\{t < T_1^n \wedge T_0^n\}} \quad \text{and} \quad \tau_n \overset{\triangle}{=} T_1^n \cdot 1_{\{T_1^n < T_0^n\}} + T \cdot 1_{\{T_1^n \geq T_0^n\}}. \tag{9.12}$$

This means if the wealth reaches the curve $x_1(T-\cdot)$ before reaching the curve $x_0(T-\cdot)$, stop immediately when this happens; if the wealth reaches the curve $x_0(T-\cdot)$ before reaching the curve $x_1(T-\cdot)$, then put all the money in the bank account and wait until the terminal time T; and up until the first time that one of these curves is reached, keep an amount of n dollars invested in stock. Clearly,

$$X^{x,\pi_n}(\tau_n) = x_0(T - T_0^n) \cdot 1_{\{T_0^n < T_1^n\}} + x_1(T - T_1^n) \cdot 1_{\{T_1^n < T_0^n\}}. \tag{9.13}$$

Moreover, since $\pi_n(\cdot)$ is bounded, the wealth process $X^{x,\pi_n}(\cdot)$ is a martingale, and the optional sampling theorem gives

$$x = \mathbb{E}\left[X^{x,\pi_n}(\tau_n)\right]. \tag{9.14}$$

Because $T_0^n = \inf\big\{t \geq 0 \;/\; W(t) \leq \frac{1}{2}nt + \frac{1}{n}\log\big(\frac{x_0(T-t)}{x}\big)\big\} \wedge T \longrightarrow 0$ a.s. as $n \to \infty$, it follows from (9.13) and (9.14) that $x_0(T)p_n + x_1(T)(1-p_n) \longrightarrow x$ as $n \to \infty$, where $p_n \overset{\triangle}{=} \mathbb{P}(T_0^n < T_1^n) = 1 - \mathbb{P}(T_1^n < T_0^n)$, or, equivalently,

$$p_n \to \frac{x_1(T) - x}{x_1(T) - x_0(T)} \qquad \text{as} \quad n \to \infty. \tag{9.15}$$

On the other hand, the expected discounted utility corresponding to (π_n, τ_n) of (9.12) is

$$J(x; \pi_n, \tau_n) = \mathbb{E}\big[e^{-T}\log\left(x_0(T - T_0^n)\right) \cdot 1_{\{T_0^n < T_1^n\}} + \log\big(e^{-T_1^n} x_1(T - T_1^n)\big) \cdot 1_{\{T_1^n < T_0^n\}}\big].$$

We conclude the proof by noting from (9.15) and the dominated convergence theorem, that

$$\lim_{n\to\infty} J(x; \pi_n, \tau_n) = e^{-T}\log x_0(T) \cdot \frac{x_1(T) - x}{x_1(T) - x_0(T)} + \log x_1(T) \cdot \frac{x - x_0(T)}{x_1(T) - x_0(T)} = G(x).$$

Remark 9.4. The tangent to the graph of $V_0(\cdot)$ at $x = x_0 \overset{\triangle}{=} x_0(T)$ and the tangent to the graph of $V_1(\cdot)$ at $x = x_1 \overset{\triangle}{=} x_1(T)$ coincide. Indeed, $V_1'(x) = \frac{1}{x}$ so that the tangent $f_1(\cdot)$ to the graph of $V_1(\cdot)$, at the point $x = x_1$, is given by

$$f_1(x) = \frac{x - x_1}{x_1} + f_1(x_1) = \left(\frac{x}{x_1} - 1\right) + \log x_1 = \lambda^*(T)x - (1 + \log \lambda^*(T)).$$

On the other hand, $V_0'(x) = \frac{1}{x}e^{-T}$ so that the tangent $f_0(\cdot)$ to the graph of $V_0(\cdot)$, at the point $x = x_0$, is given by

$$\begin{aligned} f_0(x) &= \frac{x - x_0}{x_0}e^{-T} + f_0(x_0) = e^{-T}\left(x\lambda^*(T)e^T - 1\right) + e^{-T}\log x_0 \\ &= \lambda^*(T)x - e^{-T}(1 + \log \lambda^*(T) + T). \end{aligned}$$

Thanks to (9.7), these two expressions are the same.

Appendix A. In this section we provide an example which illustrates briefly, in a Markovian setting and with logarithmic utility from wealth (we set $c(\cdot) \equiv 0$ and write $X^{x,\pi}(\cdot) \equiv X^{x,\pi,0}(\cdot)$ throughout), how the optimization problem of (5.3) can be cast in the form of a free-boundary problem for a suitable HJB equation, which can then be solved explicitly.

In order to obtain such an explicit solution, we place ourselves on an infinite time-horizon so that all stopping times $\tau \in \mathcal{S}_{0,\infty}$ are admissible, and we denote the corresponding value function by

$$V_\infty(x) \;=\; \sup_{(\pi,\tau)\in\mathcal{A}(x)} \mathbb{E}\left[e^{-\beta\tau}\log X^{x,\pi}(\tau)\cdot \mathbf{1}_{\{\tau<\infty\}}\right] \tag{A.1}$$

with $\beta > 0$, for a given initial capital $x > 0$ in the notation of (9.4). Furthermore, we assume that the coefficients of the model $r(\cdot) \equiv r > 0$, $b(\cdot) \equiv b$, $\sigma(\cdot) \equiv \sigma > 0$ are all constant, and we impose the assumption $b \neq r\mathbf{1}_m$, or, equivalently, $\theta(\cdot) \equiv \theta \neq 0$. For the measure-theoretic subtleties associated with working on an infinite time-horizon, we refer the reader to section 1.7 in Karatzas and Shreve (1998).

Consider the differential operator

$$\begin{aligned} \mathcal{L}u\,(x) &\overset{\triangle}{=} -\beta u(x) + rxu'(x) + \max_{\pi\in\mathbb{R}^m}\left(xu'(x)\pi^*\sigma\theta + \frac{1}{2}x^2u''(x) \parallel \pi^*\sigma \parallel^2\right) \\ &= -\beta u(x) + rxu'(x) - \frac{(u'(x))^2\Theta^2}{2u''(x)}, \end{aligned} \tag{A.2}$$

acting on functions $u : (0,\infty) \to \mathbb{R}$ which are twice continuously differentiable with $u''(\cdot) < 0$; here $\Theta \overset{\triangle}{=} \parallel (\sigma^*)^{-1}\theta \parallel = \parallel (\sigma\sigma^*)^{-1}(b - r\mathbf{1}_m) \parallel > 0$. By analogy with section 2.7 in Karatzas and Shreve (1998), we cast the original optimization problem of (A.1) as a *variational inequality*, relying on the familar "principle of smooth–fit."

Variational Inequality A.1. *Find a number* $b \in (1,\infty)$ *and an increasing function* $g(\cdot)$ *in the space* $\mathcal{C}([0,\infty)) \cap \mathcal{C}^1((0,\infty)) \cap \mathcal{C}^2((0,\infty)\setminus\{b\})$, *such that*

$$\mathcal{L}g\,(x) = 0\;, \quad 0 < x < b, \tag{A.3}$$

$$\mathcal{L}g\,(x) < 0\;, \quad x > b, \tag{A.4}$$

$$g(x) > \log x\;, \quad 0 < x < b, \tag{A.5}$$

$$g(x) = \log x\;, \quad x \geq b, \tag{A.6}$$

$$g(x) > 0\;, \quad x > 0, \tag{A.7}$$

$$g''(x) < 0\;, \quad x \in (0,\infty)\setminus\{b\}. \tag{A.8}$$

THEOREM A.2. *Suppose that the pair $(b, g(\cdot))$ solves the Variational Inequality* A.1*, that the ratio $|g'(x)/(xg''(x))|$ is bounded away from both zero and infinity on* $(0,\infty)$*, and that the stochastic differential equation*

$$\text{(A.9)} \qquad d\hat{X}(t) = \hat{X}(t)\left[r\,dt - \frac{g'(\hat{X}(t))}{\hat{X}(t)g''(\hat{X}(t))}\,\theta^*\,dW_0(t)\right], \qquad \hat{X}(0) = x > 0,$$

has a pathwise unique, strictly positive strong solution $\hat{X}(\cdot)$. In terms of this process, define

$$\text{(A.10)} \qquad \hat{\pi}(\cdot) \overset{\triangle}{=} -\big(\sigma^*\big)^{-1}\theta\,\frac{g'(\xi)}{\xi g''(\xi)}\bigg|_{\xi=\hat{X}(\cdot)}, \qquad \hat{\tau} \overset{\triangle}{=} \inf\left\{t \geq 0/\hat{X}(t) \geq b\right\}.$$

Then the function $g(\cdot)$ coincides with the optimal expected utility $V_\infty(\cdot)$ of (A.1)*, the pair $\big(\hat{\pi}(\cdot),\,\hat{\tau}\big)$ attains the supremum in* (A.1)*, and we have $\hat{X}^{x,\hat{\pi}}(\cdot) \equiv \hat{X}(\cdot)$.*

Proof. Fix $x \in (0,\infty)$. For any available portfolio process $\pi(\cdot)$, an application of Itô's rule to $\mathcal{G}^{x,\pi}(t) \overset{\triangle}{=} e^{-\beta t}g(X^{x,\pi}(t))$, $0 \leq t < \infty$, yields, in conjunction with (3.1), (A.3), and (A.4),

$$\begin{aligned} \text{(A.11)} \quad & e^{-\beta t}g(X^{x,\pi}(t)) - g(x) - \int_0^t e^{-\beta s}\pi^*\sigma\cdot\xi g'(\xi)\big|_{\xi=X^{x,\pi}(s)}\,dW(s) \\ &= \int_0^t e^{-\beta s}\left((\pi^*\sigma\theta + r)\cdot\xi g'(\xi) + \frac{1}{2}g''(\xi)\xi^2\parallel\pi^*\sigma\parallel^2 - \beta g(\xi)\right)\bigg|_{\xi=X^{x,\pi}(s)} ds \\ &\leq \int_0^t e^{-\beta s}\mathcal{L}g(X^{x,\pi}(s))\,ds \;\leq\; 0. \end{aligned}$$

It follows that the process $\mathcal{G}^{x,\pi}(t) = e^{-\beta t}g(X^{x,\pi}(t))$, $0 \leq t < \infty$ is a local supermartingale under $\mathbf{P}$, hence also a true supermartingale because it is positive. In particular, $\mathcal{G}^{x,\pi}(\infty) \overset{\triangle}{=} \limsup_{t\to\infty}\mathcal{G}^{x,\pi}(t) \geq 0$ exists a.s., and $\{\mathcal{G}^{x,\pi}(t),\ 0 \leq t \leq \infty\}$ is a $\mathbf{P}$-supermartingale. Thus

$$\begin{aligned} \text{(A.12)} \qquad \mathbb{E}[e^{-\beta\tau}\log X^{x,\pi}(\tau)\cdot\mathbf{1}_{\{\tau<\infty\}}] &\leq \mathbb{E}[e^{-\beta\tau}g\big(X^{x,\pi}(\tau)\big)\cdot\mathbf{1}_{\{\tau<\infty\}}] \\ &\leq \mathbb{E}[\mathcal{G}^{x,\pi}(\tau))] \leq g(x) \end{aligned}$$

holds for any stopping time $\tau \in \mathcal{S}_{0,\infty}$, by the optional sampling theorem and (A.5)–(A.6); in other words, $V_\infty(x) \leq g(x)$. We complete the proof upon noticing that, thanks to (A.3) and (A.6), all the inequalities in (A.11) and (A.12) hold as equalities for the choice

$$\text{(A.13)} \qquad \hat{\pi}(t) \overset{\triangle}{=} -\frac{g'(\hat{X}(t))}{\hat{X}(t)g''(\hat{X}(t))}(\sigma^*)^{-1}\theta, \qquad \hat{\tau}_b \overset{\triangle}{=} \inf\left\{t \geq 0/\hat{X}(t) \geq b\right\},$$

since we have $0 < g(\hat{X}(\hat{\tau}_b)) \leq \log b$ and $e^{-\beta\hat{\tau}_b}g(\hat{X}(\hat{\tau}_b)) = 0$ on the event $\{\hat{\tau}_b = \infty\}$. □

We have now to construct the solution of Variational Inequality A.1 and to verify the properties for (A.9) assumed in Theorem A.2.

PROPOSITION A.3. *Let α be the unique solution of the quadratic equation*

$$\text{(A.14)} \qquad \alpha^2 - \left(1 + \frac{\Theta^2}{2r} + \frac{\beta}{r}\right)\alpha + \frac{\beta}{r} = 0$$

in the interval $(0,1)$*, set* $b \stackrel{\triangle}{=} e^{1/\alpha}$*, and consider the function*

$$g(x) \stackrel{\triangle}{=} \left\{ \begin{array}{ll} x^{\alpha}/e\alpha, & 0 \le x < b \\ \log x, & b \le x < \infty \end{array} \right\}. \tag{A.15}$$

Then the pair $(b, g(\cdot))$ *solves Variational Inequality* A.1*, and the stochastic differential equation* (A.9) *has a pathwise unique, strictly positive strong solution* $\hat{X}(\cdot)$*.*

Proof. Note that the function

$$F(u) \stackrel{\triangle}{=} u^2 - \left(1 + \frac{\frac{\Theta^2}{2} + \beta}{r}\right) u + \frac{\beta}{r}\ , \qquad 0 \le u < \infty, \tag{A.16}$$

is convex with $F(0) = \beta/r > 0$, $F(1) = -\Theta^2/2r < 0$. Thus $F(\cdot)$ has exactly one root in the interval $(0,1)$. It is clear now that (A.6)–(A.8) are satisfied since $b > 1$. Furthermore, notice from (A.15) that

$$g'(x) = \left\{ \begin{array}{ll} x^{\alpha-1}/e, & 0 < x < b \\ 1/x, & b < x < \infty \end{array} \right\} \tag{A.17}$$

is continuous across $x = b$ (principle of smooth-fit), which implies that the function $g(\cdot)$ belongs to the space of functions $\mathcal{C}([0,\infty)) \cap \mathcal{C}^1((0,\infty)) \cap \mathcal{C}^2((0,\infty) \setminus \{b\})$. It is fairly straightforward to check that (A.3) holds for $0 < x < b$, and that $|g'(x)/(xg''(x))|$ is bounded away from both zero and infinity on $(0,\infty)$ (cf. (A.19) below). As for (A.4), we need to prove that $-\beta \log x + r + \Theta^2/2 < 0 \ \ \forall\, x > b$. Since $\log b = 1/\alpha$ and $\beta > 0$, it is sufficient to verify $\alpha < \alpha^* \stackrel{\triangle}{=} \beta/(r + \frac{\Theta^2}{2})$. Indeed

$$\begin{aligned} F(\alpha^*) &= \alpha^* \left(\alpha^* - \frac{\frac{\Theta^2}{2} + \beta}{r} - 1 \right) + \frac{\beta}{r} \\ &< \alpha^* \left(\frac{\beta}{r} - \frac{\frac{\Theta^2}{2} + \beta}{r} - 1 \right) + \frac{\beta}{r} = \alpha^* \left(- \frac{\frac{\Theta^2}{2} + r}{r} \right) + \frac{\beta}{r} \ = \ 0, \end{aligned}$$

which yields $\alpha < \alpha^*$. Finally, (A.5) follows readily from

$$g'(x) - (\log x)' = \frac{1}{x}\left(\frac{1}{e}x^{\alpha} - 1\right) < \frac{1}{x}\left(\frac{1}{e}b^{\alpha} - 1\right) = 0, \quad 0 < x < b.$$

It is now clear that the pair $(b, g(\cdot))$ solves Variational Inequality A.1.

For the function $g(\cdot)$ of (A.15), the optimal wealth-process $\hat{X}(\cdot)$ of Theorem A.2 satisfies the stochastic differential equation (A.9), namely,

$$d\hat{X}(t) = \hat{X}(t)\big[r\,dt + \nu\big(\hat{X}(t)\big)\theta^*\,dW_0(t)\big], \qquad \hat{X}(0) = x > 0, \tag{A.18}$$

where

$$\nu(x) \ \stackrel{\triangle}{=} \ -\frac{g'(x)}{xg''(x)} \ = \ \left\{ \begin{array}{ll} 1/(1-\alpha), & 0 < x < b \\ 1, & b \le x < \infty \end{array} \right\}. \tag{A.19}$$

Equivalently, the process $\hat{Y}(\cdot) \stackrel{\triangle}{=} \log \hat{X}(\cdot)$ solves the stochastic differential equation

$$d\hat{Y}(t) = \left[r - \frac{||\theta||^2}{2} \cdot \nu^2\big(e^{\hat{Y}(t)}\big)\right] dt + \nu\big(e^{\hat{Y}(t)}\big)\theta^*\,dW_0(t), \ \ \hat{Y}(0) = \log x, \tag{A.20}$$

which has a pathwise unique, strong solution (cf. Nakao (1972)). This, in turn, means that (A.15) for $\hat{X}(\cdot) \equiv e^{\hat{Y}(\cdot)}$ also has a strictly positive, pathwise unique strong solution, as postulated in Theorem A.2. □

Remark A.4. For $x \geq b$, we have $\hat{\tau} \equiv 0$; on the other hand, for $0 < x < b$, we can write the stopping time $\hat{\tau} \overset{\triangle}{=} \inf \left\{ t \geq 0 \Big/ \hat{X}(t) \geq x \right\} = \inf \left\{ t \geq 0 \Big/ \hat{Y}(t) \geq \log b \right\}$ in the form of the time

$$\hat{\tau} = \inf \left\{ t \geq 0 \Bigg/ \left(r + \frac{||\theta||^2}{2} \frac{1-2\alpha}{(1-\alpha)^2} \right) t + \frac{\theta^*}{1-\alpha} W(t) \ \geq \ \log\left(\frac{b}{x}\right) \right\}$$

of first-passage to a positive level by a Brownian motion with drift. Clearly, we have $\mathbf{P}[\hat{\tau} < \infty] = 1$ if and only if $(1-\alpha)^2 + ||\theta||^2(1-2\alpha)/2r \geq 0$, and in light of (A.14) this last condition is equivalent to

$$\text{(A.21)} \qquad \left(\beta - r - ||\theta||^2 + \frac{\Theta^2}{2} \right) \cdot \alpha \ \geq \ \left(\beta - r - \frac{||\theta||^2}{2} \right).$$

In particular, if $\sigma = I_m$, the condition (A.21) amounts to

$$\text{(A.22)} \qquad \beta \leq r + ||b - r\mathbf{1}_m||^2.$$

Remark A.5. From (A.13), the optimal portfolio process is actually given as

$$\text{(A.23)} \qquad \hat{\pi}(t) \ \equiv \ \frac{(\sigma^*)^{-1}}{1-\alpha}\theta \ = \ \frac{(\sigma\sigma^*)^{-1}}{1-\alpha}[b - r\mathbf{1}_m], \qquad 0 \leq t < \hat{\tau} \ ;$$

this means that the optimal strategy is to invest a *fixed* proportion of total wealth in every stock, given by (A.3), up to the optimal stopping time $\hat{\tau}$.

Remark A.6. The assumption $\theta \neq 0$ is crucial for solving Variational Inequality A.1. When $\theta = 0$, we can have situations, as in Example 9.3, for which *no* optimal strategy exists. Actually, for $\theta = 0$ and $\beta > r$, it is easy to show that Variational Inequality A.1 has no solution (see Example 9.2 for discussion of the case $\theta = 0$, $\beta < r$).

Appendix B. As the referee points out, it would be very interesting to study optimization over a consumption stream that extends beyond the stopping time τ. Consider, for instance, the situation of an investor who remains in the stock-market up until a "retirement" time τ of his choice. At that point he consumes a lump-sum amount $\xi \geq 0$ of his choice (say, to buy a new house, or to finance some other "retirement-related" activity); and from then on he keeps his holdings in the money-market, making withdrawals for consumption at some rate, up until $t = T$.

We can capture such a situation by changing the wealth-equation of (3.1) to read

$$\text{(B.1)} \qquad dX(t) = r(t)X(t)dt + X(t)\pi^*(t)\sigma(t)dW_0(t) - dC(t), \quad X(0) = x > 0.$$

Here

$$\text{(B.2)} \qquad C(t) = \int_0^t c(u)\,du + \xi \cdot 1_{[\tau,T]}(t), \qquad 0 \leq t \leq T,$$

is the "cumulative consumption up to time t." This process consists of a stopping time $\tau \in \mathcal{S}$, a consumption-rate process $c(\cdot)$ as before, and an $\mathcal{F}_\tau$-measurable random variable $\xi : \Omega \to [0, \infty)$ representing lump-sum consumption at time τ. We say that a portfolio/cumulative-consumption process pair (π, C) is “available” to an investor with initial capital x, if the portfolio process $\pi(\cdot)$ and the wealth-process $X(\cdot) \equiv X^{x,\pi,C}(\cdot)$ of (B.1) satisfy

$$\pi(t) = 0, \qquad \tau \le t \le T, \tag{B.3}$$

$$X^{x,\pi,C}(t) > 0 \quad \forall\, 0 \le t < T, \qquad \text{and} \qquad X^{x,\pi,C}(T) \ge 0\ , \tag{B.4}$$

a.s. For any such pair (π, C), the investor’s expected discounted utility is given as

$$J^*(x; \pi, C) \ \stackrel{\triangle}{=}\ \mathbb{E}\left[\alpha \int_0^\tau e^{-\beta t} U_1\big(c(t)\big)\, dt + e^{-\beta\tau} U_2(\xi) + \gamma \int_\tau^T e^{-\beta t} U_1\big(c(t)\big)\, dt\right] \tag{B.5}$$

for some given constants $\alpha \ge 0,\ \gamma \ge 0$ and utility functions $U_1(\cdot),\ U_2(\cdot)$. With $\alpha = 1,\ \gamma = 0,$ we recover the problem of section 5. With $\alpha = 0,\ \gamma = 1,$ the expression of (B.5) tries to capture the situation of an investor who consumes nothing up until retirement, consumes a lump-sum amount ξ at that time, and afterwards keeps all holdings in the money-market while consuming at some rate $c(\cdot)$. The objective now is to maximize the expression of (B.5) over the class $\mathcal{A}^*(x)$ of pairs (π, C) that satisfy the analogue

$$\mathbb{E}\left[\alpha \int_0^\tau e^{-\beta t} U_1^-\big(c(t)\big)\, dt + e^{-\beta\tau} U_2^-(\xi) + \gamma \int_\tau^T e^{-\beta t} U_1^-\big(c(t)\big)\, dt\right] < \infty \tag{B.6}$$

of (5.2), and to see whether the value-function

$$V^*(x) \ \stackrel{\triangle}{=}\ \sup_{(\pi,C)\in\mathcal{A}^*(x)} J^*(x; \pi, C), \qquad x \in (0, \infty), \tag{B.7}$$

is attained by some optimal $(\hat{\pi}, \hat{C}) \in \mathcal{A}^*(x)$. We have not yet been able to obtain a satisfactory answer to these questions and would like to suggest their resolution as an interesting open problem.

Acknowledgments. We are indebted to the associate editor and the referees for their very careful reading of the first version of this paper and for their many helpful suggestions. In particular, the open problem presented in Appendix B was inspired by suggestions from one of the referees.

REFERENCES

[1] J. Cox and C.F. Huang (1989), *Optimal consumption and portfolio policies when asset prices follow a diffusion process*, J. Econom. Theory, 49, pp. 33–83.

[2] J. Cvitanić and I. Karatzas (1992), *Convex duality in constrained portfolio optimization*, Ann. Appl. Probab., 2, pp. 767–818.

[3] M.H.A. Davis and Th. Zariphopoulou (1995), *American options and transaction fees*, in Mathematical Finance, IMA Vol. Math. Appl. 65, Springer-Verlag, New York, pp. 47–62.

[4] M.H.A. Davis and M. Zervos (1994), *A problem of singular stochastic control with discretionary stopping*, Ann. Appl. Probab., 4, pp. 226–240.

[5] L.E. Dubins and L.J. Savage (1965), *How to Gamble if You Must: Inequalities for Stochastic Processes*, McGraw-Hill, New York.

[6] N. El Karoui (1981), *Les Aspects Probabilistes du Contrôle Stochastique*, Lecture Notes in Math., Springer-Verlag, Berlin, pp. 73–238.

[7] J.M. Harrison and S.R. Pliska (1981), *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11, pp. 215–260.

[8] I. Karatzas and S.G. Kou (1998), *Hedging American contingent claims with constrained portfolios*, Finance Stoch., 2, pp. 215–258.

[9] I. Karatzas, J.P. Lehoczky, and S.E. Shreve (1987), *Optimal portfolio and consumption decisions for a "small investor" on a finite horizon*, SIAM J. Control Optim., 25, pp. 1557–1586.

[10] I. Karatzas, J.P. Lehoczky, S.E. Shreve, and G.L. Xu (1991), *Martingale and duality methods for utility maximization in an incomplete market*, SIAM J. Control Optim., 29, pp. 702–730.

[11] I. Karatzas and S.E. Shreve (1991), *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York.

[12] I. Karatzas and S.E. Shreve (1998), *Methods of Mathematical Finance*, Springer-Verlag, New York.

[13] I. Karatzas and W.D. Sudderth (1999), *Control and stopping of a diffusion on an interval*, Ann. Appl. Probab., 9, pp. 188–196.

[14] A.P. Maitra and W.D. Sudderth (1996), *Discrete Gambling and Stochastic Games*, Springer-Verlag, New York.

[15] R.C. Merton (1971), *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3, pp. 373–413. *Erratum*, 6 (1973), pp. 213–214.

[16] S. Nakao (1972), *On the pathwise uniqueness of solutions of stochastic differential equations*, Osaka J. Math., 9, pp. 513–518.

[17] S.R. Pliska (1986), *A stochastic calculus model of continuous trading: Optimal portfolios*, Math. Oper. Res., 11, pp. 371–382.

[18] R.T. Rockafellar (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

[19] P.A. Samuelson and R.C. Merton (1969), *A complete model of warrant-pricing that maximizes utility*, Industr. Mangmt. Review, 10, pp. 17–46.

[20] S.E. Shreve and G.L. Xu (1992), *A duality method for optimal consumption and investment under short-selling prohibition,* I: *General market coefficients, and* II: *Constant market coefficients*, Ann. Appl. Probab., 2, pp. 87–112 and pp. 314–328.

SIAM J. Control Optim.
Vol. 39, No. 2, pp. 331–351

EXACT BOUNDARY CONTROLLABILITY FOR THE LINEAR KORTEWEG–DE VRIES EQUATION ON THE HALF-LINE[*]

LIONEL ROSIER[†]

Abstract. This paper is concerned with the controllability of the linear Korteweg–de Vries equation on the domain $\Omega = (0, +\infty)$, the control being applied at the left endpoint $x = 0$. It is shown that the *exact* boundary controllability holds true in $L^2(0, +\infty)$ provided that the solutions are not required to be in $L^\infty(0, T, L^2(0, +\infty))$. The proof rests on a Carleman's estimate and an approximation theorem. A similar result is obtained for the heat equation and for the Schrödinger equation.

Key words. exact boundary controllability, unbounded domain, Carleman's estimate

AMS subject classifications. 35Q20, 93B05, 93C20

PII. S0363012999353229

1. Introduction and main results.

The Korteweg–de Vries (KdV) equation

$$u_t + u_x + uu_x + u_{xxx} = 0, \quad t \geq 0,\ x \in \Omega \subset \mathbb{R}, \tag{1.1}$$

may serve as a model for (among other things) propagation of small amplitude long water waves in a uniform channel. In this context, t is time, x is the space variable, and u stands for the deviation of the liquid's surface from the equilibrium position. The boundary (resp., internal) controllability of (1.1) has been extensively studied (see [21], [22], [19], [20], and also [16] for the Benjamin–Bona–Mahony equation) when Ω is bounded, say $\Omega = (0, L)$. The (local) exact boundary controllability of (1.1) follows in [19] from the *exact* boundary controllability of the associated linear KdV equation, namely

$$u_t + u_x + u_{xxx} = 0. \tag{1.2}$$

To date, there is no result as far as the boundary controllability of (1.1) or (1.2) on some *unbounded* domain (say $\Omega = (0, +\infty)$) is concerned. The aim of this paper is to fill this gap in providing a study of the exact boundary controllability of (1.2) on $(0, +\infty)$, which may be seen as a first step in the knowledge of the control theory for (1.1) on unbounded domains. It should be observed that the *approximate* boundary controllability of (1.2) in $L^2(0, +\infty)$ is quite easy to prove, whereas the *exact* boundary controllability requires a more sophisticated analysis, due to a lack of compactness. An enlightening example of the difference between exact and approximate (internal) controllabilities for linear PDEs in unbounded domains is provided by the following result, whose (simple) proof is sketched in the appendix.

Proposition 1.1. *Consider a (real) constant coefficients differential operator* $Au = \sum_{i=0}^{n} a_i \frac{d^i u}{d x^i}$, *with domain* $\mathcal{D}(A) = \{u \in L^2(\mathbb{R});\ Au \in L^2(\mathbb{R})\}$. *Assume that*

[*] Received by the editors March 12, 1999; accepted for publication (in revised form) April 5, 2000; published electronically August 31, 2000.
http://www.siam.org/journals/sicon/39-2/35322.html

[†] Laboratoire d'Analyse Numérique et EDP, Université Paris-Sud, bâtiment 425, 91405 Orsay Cedex, France (Lionel.Rosier@math.u-psud.fr).

$n \geq 2$ (with $a_n \neq 0$) and that A generates a continuous semigroup $\big(S(t)\big)_{t\geq 0}$ on $L^2(\mathbb{R})$. Let $T > 0$ and $L_1 < L_2$ be some numbers. Set

$$\mathcal{R} = \left\{ \int_0^T S(T-t)f(t,\cdot)dt; \quad f \in L^2(\mathbb{R}^2),\ \operatorname{supp} f \subset [0,T] \times [L_1, L_2] \right\},$$

where supp f *denotes the support of f. Then $\mathcal{R}$ is a* strict dense *subspace of $L^2(\mathbb{R})$.*

In other words, when considering *mild* solutions (in $C([0,T], L^2(\mathbb{R}))$) of the forced initial-value problem

$$\begin{cases} \frac{du}{dt} - Au &= f, \\ u(0) &= 0, \end{cases}$$

where f denotes any square integrable function supported in $[0,T] \times [L_1, L_2]$, the space $\mathcal{R}$ of all reachable states is dense in (but different from) $L^2(\mathbb{R})$. Notice that for $Au = -u_{xxx} - u_x$, letting $L_1 = -1 < L_2 = 0$ and taking the restrictions to $(0, +\infty)$ of the mild solutions, we readily infer the approximate *boundary* controllability of (1.2) in $L^2(0, +\infty)$. It turns out that the *exact* boundary controllability of (1.2) in $L^2(0,+\infty)$ also fails to be true if we restrict ourselves to solutions with bounded energy, that is, which belong to $L^\infty(0,T,L^2(0,+\infty))$. An *implicit* formulation (that is, without specification of the boundary conditions) of this fact is given in the following theorem, to be proved later in this paper.

THEOREM 1.2. *Let $T > 0$. Then there exists $u_0 \in L^2(0, +\infty)$ such that if u is any function in $L^\infty(0,T,L^2(0,+\infty))$ satisfying*

$$\begin{cases} u_t + u_x + u_{xxx} &= 0 \quad \text{in } \mathcal{D}'\big((0,T) \times (0,+\infty)\big), \\ u_{|_{t=0}} &= u_0, \end{cases} \tag{1.3}$$

then $u_{|_{t=T}} \neq 0$.

(Notice that $u_{|_{t=0}}$ and $u_{|_{t=T}}$ are meaningful in $H^{-3}(0,+\infty)$ for any $u \in L^\infty(0,T, L^2(0,+\infty))$ satisfying (1.3): Indeed, such a function belongs to the space $W^{1,\infty}(0,T, H^{-3}(0,+\infty))$.) Theorem 1.2 tells us that even the (boundary) *null-controllability* fails to be true for solutions with bounded energy. Nevertheless, when the bounded energy condition ($u \in L^\infty(0,T,L^2(0,+\infty))$) is dropped, the exact boundary controllability of KdV holds true, as is shown in the following theorem, which is the main result of this paper.

THEOREM 1.3. *Let T, ϵ, b be positive numbers, with $\epsilon < \frac{T}{2}$. Let $L^2((0,+\infty), e^{-2bx}dx)$ denote the space of (class of) measurable functions $u : (0,+\infty) \to \mathbb{R}$ such that $\int_0^{+\infty} u^2(x)e^{-2bx}\,dx < \infty$. Let $u_0 \in L^2(0,+\infty)$ and $u_T \in L^2((0,+\infty), e^{-2bx}dx)$. Then there exists a function*

$$u \in L^2_{loc}\big([0,T] \times [0,+\infty)\big) \cap C\big([0,\epsilon], L^2(0,+\infty)\big) \cap C\big([T-\epsilon, T], L^2((0,+\infty), e^{-2bx}dx)\big)$$

which solves

$$\begin{cases} u_t + u_x + u_{xxx} &= 0 \quad \text{in } \mathcal{D}'\big((0,T) \times (0,+\infty)\big), \\ u_{|_{t=0}} &= u_0, \\ u_{|_{t=T}} &= u_T. \end{cases} \tag{1.4}$$

Let us make some comments.

1. The proof of Theorem 1.3 combines Fursikov–Imanuvilov's approach (see [4]) for the boundary controllability of the Burgers equation on bounded domains (which

is based on a global Carleman's estimate) and, for the extension to some unbounded domain, Rosay's clever proof of Malgrange–Ehrenpreis's theorem (see [18]), which uses an approximation theorem. Roughly speaking, the approximation theorem allows us to modify a sequence of solutions of $u_t + u_x + u_{xxx} = f$, defined on an increasing sequence of domains, in such a way that it converges (strongly) in $L^2_{loc}(\mathbb{R}^2)$. It should be emphasized that our approach allows us to consider initial and final states in *different* spaces of functions, thus exploiting an *asymmetric* property of the KdV equation, namely the (forward) wellposedness of (1.2) in the asymmetric space $L^2(\mathbb{R}, e^{2bx}dx)$ for any $b > 0$ (see [9]). Notice that we may require that $u \in C([T-\epsilon, T], L^2(0, +\infty))$ if u_T is also assumed to be in $L^2(0, +\infty)$.

2. As in [2] and [13], the formulation of the previous boundary controllability result is *implicit*. Nevertheless, setting $h_0 = u_{|_{x=0}}$, $h_1 = u_{x|_{x=0}}$, and $h_2 = u_{xx|_{x=0}}$, it may be seen that $h_0, h_1, h_2 \in H^{-1}(0,T)$ and, thanks to Holmgren's uniqueness theorem, that u is the *only* solution (in the same space as above) of the initial-value boundary problem

$$\begin{cases} u_t + u_x + u_{xxx} &= 0 \quad \text{in } \mathcal{D}'\big((0,T)\times(0,+\infty)\big), \\ u_{|_{x=0}} &= h_0, \ u_{x|_{x=0}} = h_1, \ u_{xx|_{x=0}} = h_2, \\ u_{|_{t=0}} &= u_0. \end{cases}$$

Moreover u satisfies $u_{|_{t=T}} = u_T$.

3. The method described in item 1 applies also to many other linear PDEs for which the characteristic hyperplanes take the form $\{t = \text{Const.}\}$: For instance, the heat equation $u_t - \Delta u = 0$ and the Schrödinger equation $iu_t + \Delta u = 0$ are concerned. (See section 5.)

The paper is outlined as follows. The proof of Theorem 1.2 is given in section 2. It rests on a duality argument and on the behavior of the traces $u_{x|_{x=0}}$, $u_{xx|_{x=0}}$ of exponential solutions for (1.2) with the boundary condition $u_{|_{x=0}} = 0$. A global Carleman's estimate for the KdV equation (which is subsequently used) is stated and proved in section 3. The proof of Theorem 1.3 is given in section 4, together with the proof of the approximation theorem (Lemma 4.4). In the last section we sketch the proof of similar results for the heat equation and the Schrödinger equation.

From now on, for the sake of brevity, we shall write P for the operator $(\partial/\partial t) + (\partial/\partial x) + (\partial^3/\partial x^3)$.

2. Proof of Theorem 1.2. The proof of Theorem 1.2 rests on the following key result.

LEMMA 2.1. *There exists a family* $(v^\lambda)_{\lambda>0}$ *of functions in* $\cap_{n\geq 0} C^\infty([0,T], H^n(0,+\infty))$ *such that for every* $\lambda > 0$

$$P v^\lambda = 0 \quad \textit{in } (0,T)\times(0,+\infty), \tag{2.1}$$

$$v^\lambda{}_{|_{x=0}} = 0 \quad \textit{on } (0,T), \tag{2.2}$$

$$\|v^\lambda{}_{|_{t=0}}\|_{L^2(0,+\infty)} = 1, \tag{2.3}$$

and

$$\|v^\lambda_{x|_{x=0}}\|_{H^n(0,T)} + \|v^\lambda_{xx|_{x=0}}\|_{H^n(0,T)} \to 0 \ \textit{as } \lambda \to 0 \ (\textit{for every } n \geq 1). \tag{2.4}$$

Proof. Let us consider the operator $Av := -v_{xxx} - v_x$ with domain

$$D(A) = H^3(0,+\infty) \cap H_0^1(0,+\infty) \subset L^2(0,+\infty).$$

Then A generates a strongly continuous semigroup $\big(S(t)\big)_{t\geq 0}$ on $L^2(0,+\infty)$, and we shall search for v^λ in the form of an exponential solution:

$$v^\lambda(t,\cdot) = S(t)v_0^\lambda = e^{-\lambda t}v_0^\lambda,$$

where $v_0^\lambda \in D(A)$ solves $Av_0^\lambda = -\lambda v_0^\lambda$ and $\lambda \in (0,+\infty)$. The roots of the equation $-z^3 - z = -\lambda$ may be written in the form $r, -\frac{r}{2} \pm i\mu$, where $0 < r \sim \lambda$ as $\lambda \to 0^+$ and $\mu = \big(1 + \frac{3}{4}r^2\big)^{\frac{1}{2}}$. Let $w_0^\lambda(x) := \Im m\,\big(e^{(-\frac{r}{2}+i\mu)x}\big) = e^{-\frac{r}{2}x}\sin(\mu x)$ (hence $w_0^\lambda \in D(A)$ and $Aw_0^\lambda = -\lambda w_0^\lambda$). Easy calculations give

$$\|w_0^\lambda\|_{L^2(0,+\infty)} = \left(\frac{2\mu^2}{r(r^2+4\mu^2)}\right)^{\frac{1}{2}}.$$

Set $c_\lambda := (\frac{r(r^2+4\mu^2)}{2\mu^2})^{\frac{1}{2}}$, $v_0^\lambda := c_\lambda w_0^\lambda$, and $v^\lambda(t,x) := e^{-\lambda t}v_0^\lambda(x)$. Since we know that (2.1)–(2.3) are true, it remains to prove (2.4). Obviously $v_x^\lambda(t,0) = c_\lambda \mu e^{-\lambda t}$ and $v_{xx}^\lambda(t,0) = -c_\lambda \mu\, r\, e^{-\lambda t}$, and since $c_\lambda \to 0$ as $\lambda \to 0^+$, (2.4) follows. □

It will result from the next lemma that the traces $u_{|_{x=0}}$, $u_{x|_{x=0}}$, and $u_{xx|_{x=0}}$ of a bounded energy solution $u = u(t,x)$ of (1.2) belong to the dual space to $H^1(0,T)$ (which is not to be confused with $H^{-1}(0,T) = H_0^1(0,T)')$.

LEMMA 2.2. *Let T and L be positive numbers and let $u \in L^\infty(0,T,L^2(0,L))$ be such that $Pu = 0$ in $\mathcal{D}'\big((0,T)\times(0,L)\big)$. Then $u \in H^3(0,L,H^1(0,T)')$ and we have for some constant $C = C(L,T) > 0$*

$$\begin{aligned}\|u(\cdot,0)\|_{H^1(0,T)'} + \|u_x(\cdot,0)\|_{H^1(0,T)'} + \|u_{xx}(\cdot,0)\|_{H^1(0,T)'}\\ \leq C\|u\|_{L^\infty(0,T,L^2(0,L))}.\end{aligned} \tag{2.5}$$

Proof. Since $u_t = -(u_{xxx} + u_x) \in L^2(0,T,H^{-3}(0,L))$, we see that $u \in H^1(0,T,H^{-3}(0,L))$; hence for every $f \in H^1(0,T,H_0^3(0,L))$

$$\int_0^T \langle u_t, f\rangle\, dt = -\int_0^T\int_0^L u f_t\, dxdt + [\langle u,f\rangle]_{t=0}^T, \tag{2.6}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing $\langle\cdot,\cdot\rangle_{H^{-3}(0,L),H_0^3(0,L)}$. Since $u \in C([0,T], H^{-3}(0,L)) \cap L^\infty(0,T,L^2(0,L))$, u, as a function of t, is weakly continuous in $L^2(0,L)$. Hence

$$\begin{aligned}\left|\int_0^T \langle u_t, f\rangle\, dt\right| &\leq \|u\|_{L^2((0,T)\times(0,L))}\cdot\|f_t\|_{L^2((0,T)\times(0,L))}\\ &\quad+ \|u\|_{L^\infty(0,T,L^2(0,L))}\cdot\big(\|f(0,\cdot)\|_{L^2(0,L)} + \|f(T,\cdot)\|_{L^2(0,L)}\big)\\ &\leq C_1\|u\|_{L^\infty(0,T,L^2(0,L))}\cdot\|f\|_{L^2(0,L,H^1(0,T))}\end{aligned} \tag{2.7}$$

for some constant $C_1 = C_1(T,L) > 0$. Since $H^1(0,T,H_0^3(0,L))$ is dense in $L^2(0,L,H^1(0,T))$, we infer from (2.7) that $u_t \in L^2(0,L,H^1(0,T)')$. Integrating three times with respect to (w.r.t.) x in the equation

$$(u_{xx} + u)_x = -u_t \qquad \text{in } \mathcal{D}'(0,L,H^1(0,T)')$$

($u_t \in L^2(0,L,H^1(0,T)')$ being given by (2.6)), we deduce that $u \in H^3(0,L,H^1(0,T)')$ and (2.5) follows. □

We now proceed to the proof of Theorem 1.2. Arguing by contradiction, we assume that for every $u_0 \in L^2(0,+\infty)$ there exists a function $u \in L^\infty(0,T,L^2(0,+\infty))$ such that $Pu = 0$ in $\mathcal{D}'((0,T)\times(0,+\infty))$, $u_{|t=0} = u_0$, and $u_{|t=T} = 0$. Let E denote the space

$$\{u \in L^\infty(0,T,L^2(0,+\infty)),\ Pu = 0 \text{ in } \mathcal{D}'((0,T)\times(0,+\infty)) \text{ and } u_{|t=T} = 0\},$$

endowed with the norm $\|u\|_E = \|u\|_{L^\infty(0,T,L^2(0,+\infty))}$. It is a Banach space, since

$$\|u_{|t=T}\|_{H^{-3}(0,+\infty)} \le C\|u\|_{H^1(0,T,H^{-3}(0,+\infty))} \le C'\|u\|_E$$

for all $u \in E$ and some constants $C, C' > 0$. Also, the linear map $\Lambda : u \in E \mapsto u_{|t=0} \in H^{-3}(0,+\infty)$ is continuous. Actually, thanks to [14, Lem. 8.1], Λ takes values in $L^2(0,+\infty)$ and we readily infer from the closed graph theorem that Λ is continuous as a map from E into $L^2(0,+\infty)$. Let $N = \ker\Lambda$, let $\tilde{E}$ stand for the quotient space of E by N, and let π denote the natural projection of E onto $\tilde{E}$. Then $\tilde{E}$ is a Banach space for the norm $\|\pi(u)\|_{\tilde{E}} := \inf_{w\in\pi(u)}\|w\|_E$, and the induced map $\tilde{\Lambda} : \tilde{E} \to L^2(0,+\infty)$ (defined by $\tilde{\Lambda}(\pi(u)) = \Lambda(u)$ for any $u \in E$) has a continuous inverse by the open mapping theorem. For every $\lambda > 0$ we pick $u^\lambda \in E$ such that $\pi(u^\lambda) = \tilde{\Lambda}^{-1}(v^\lambda(0,\cdot))$ (with v^λ as in Lemma 2.1) and

$$\|u^\lambda\|_E \le 2\|\pi(u^\lambda)\|_{\tilde{E}} \le 2\|\tilde{\Lambda}^{-1}\|. \tag{2.8}$$

Let L be a positive number. Integrations by part in

$$\int_0^T\int_0^L P(u^\lambda)v^\lambda\,dxdt = 0$$

result in

$$-\int_0^L v^\lambda(0,x)^2\,dx + \left[\langle u^\lambda_{xx} + u^\lambda, v^\lambda\rangle - \langle u^\lambda_x, v^\lambda_x\rangle + \langle u^\lambda, v^\lambda_{xx}\rangle\right]_{x=0}^L = 0, \tag{2.9}$$

where $\langle\cdot,\cdot\rangle$ denotes here the duality pairing $\langle\cdot,\cdot\rangle_{H^1(0,T)',H^1(0,T)}$. Since

$$\begin{aligned}&\|u^\lambda(\cdot,L)\|_{H^1(0,T)'} + \|u^\lambda_x(\cdot,L)\|_{H^1(0,T)'} + \|u^\lambda_{xx}(\cdot,L)\|_{H^1(0,T)'}\\ &\quad\le C\|u\|_{L^\infty(0,T,L^2(L,L+1))}\\ &\quad\le C\|u\|_{L^\infty(0,T,L^2(0,+\infty))}\end{aligned}$$

(where $C = C(1,T)$ is as in Lemma 2.2) and since $v^\lambda(\cdot,L)$, $v^\lambda_x(\cdot,L)$ and $v^\lambda_{xx}(\cdot,L) \to 0$ in $H^1(0,T)$ as $L \to +\infty$, letting $L \to +\infty$ in (2.9) and using (2.2)–(2.3) we get

$$1 = \int_0^{+\infty} v^\lambda(0,x)^2\,dx = \langle u^\lambda_x, v^\lambda_x\rangle_{|x=0} - \langle u^\lambda, v^\lambda_{xx}\rangle_{|x=0}.$$

Hence, by (2.5) and (2.8) (also with $C = C(1,T)$)

$$1 \le 2C\|\tilde{\Lambda}^{-1}\|\left(\|v^\lambda_{x|x=0}\|_{H^1(0,T)} + \|v^\lambda_{xx|x=0}\|_{H^1(0,T)}\right). \tag{2.10}$$

Letting $\lambda \to 0$ in (2.10) and using (2.4) we get a contradiction. The proof of Theorem 1.2 is complete. □

3. A Carleman's estimate. Let T and L be positive numbers. Set

$$\mathcal{Z} = \{q \in C^3([0,T]\times[-L,L]);\ q(t,\pm L) = q_x(t,\pm L) = q_{xx}(t,\pm L) = 0 \text{ for } 0 \le t \le T\}.$$

This section is devoted to the proof of the following global Carleman's estimate for the KdV equation.

PROPOSITION 3.1. *There exists a smooth positive function ψ on $[-L,L]$ (which depends on L) and there exist constants $s_0 = s_0(L,T)$ and $C = C(L,T)$ such that for all $s \ge s_0$ and all $q \in \mathcal{Z}$*

$$\begin{aligned}\int_0^T\int_{-L}^L &\left\{\frac{s^5}{t^5(T-t)^5}|q|^2 + \frac{s^3}{t^3(T-t)^3}|q_x|^2 + \frac{s}{t(T-t)}|q_{xx}|^2\right\} e^{-\frac{2s\psi(x)}{t(T-t)}}\,dxdt \\ &\le C\int_0^T\int_{-L}^L |q_t + q_x + q_{xxx}|^2 e^{-\frac{2s\psi(x)}{t(T-t)}}\,dxdt.\end{aligned} \tag{3.1}$$

Proof. Let $\psi = \psi(x)$ be a positive function (to be specified later) of class C^3 in $[-L,L]$ and let $\varphi(t,x) := \frac{\psi(x)}{t(T-t)}$. Let q be given in $\mathcal{Z}$ and let $s > 0$. Set $u := e^{-s\varphi}q$ and $w := e^{-s\varphi}P(e^{s\varphi}u)$. We readily get

$$w = Au + Bu_x + Cu_{xx} + u_{xxx} + u_t, \tag{3.2}$$

with

$$\begin{aligned} A &:= s(\varphi_t + \varphi_x + \varphi_{xxx}) + 3s^2\varphi_x\varphi_{xx} + (s\varphi_x)^3,\\ B &:= 1 + 3s\varphi_{xx} + 3(s\varphi_x)^2,\\ C &:= 3s\varphi_x.\end{aligned}$$

Set $M_1(u) := u_t + u_{xxx} + Bu_x$ and $M_2(u) := Au + Cu_{xx}$. We deduce the following inequality:

$$2\iint M_1(u)\,M_2(u) \le \iint \big(M_1(u) + M_2(u)\big)^2 = \iint w^2. \tag{3.3}$$

(Here and in what follows, the integrals are extended to $(0,T)\times(-L,L)$.) To compute the integral in the left-hand side of (3.3) we perform integrations by part w.r.t. x or t. We readily get

$$\iint M_1(u)Au = -\frac{1}{2}\iint (A_t + A_{xxx} + (AB)_x)u^2 + \frac{3}{2}\iint A_x u_x^2 \tag{3.4}$$

and

$$\iint (u_{xxx} + Bu_x)Cu_{xx} = -\frac{1}{2}\iint C_x u_{xx}^2 - \frac{1}{2}\iint (BC)_x u_x^2. \tag{3.5}$$

Finally, using (3.2),

$$\begin{aligned}\iint u_tCu_{xx} &= -\iint C_xu_tu_x - \iint Cu_{tx}u_x\\ &= \iint C_x\big(Au + Bu_x + Cu_{xx} + u_{xxx} - w\big)u_x + \frac{1}{2}\iint C_tu_x^2\\ &= -\frac{1}{2}\iint (C_xA)_xu^2 + \frac{1}{2}\iint (2BC_x - (CC_x)_x + C_{xxx} + C_t)u_x^2\\ &\quad -\iint C_xu_{xx}^2 - \iint C_xwu_x.\end{aligned} \tag{3.6}$$

Combining (3.4), (3.5), and (3.6) we get

$$\begin{aligned} 2\iint M_1(u)M_2(u) &= -\iint \big(A_t + A_{xxx} + (AB)_x + (C_xA)_x\big)u^2 \\ &\quad+\iint \big(3A_x - (BC)_x + 2BC_x - (CC_x)_x + C_{xxx} + C_t\big)u_x^2 \qquad (3.7)\\ &\quad-3\iint C_x u_{xx}^2 - 2\iint C_x w u_x\cdot \end{aligned}$$

If ϵ is any number in $(0,1)$, then by the Cauchy–Schwarz inequality

$$2\iint C_x w u_x \le \epsilon \iint C_x^2 u_x^2 + \epsilon^{-1}\iint w^2\cdot$$

Hence, setting

$$\begin{aligned} D &:= -\big(A_t + A_{xxx} + (AB)_x + (C_xA)_x\big),\\ E &:= 3A_x + BC_x - B_xC - (CC_x)_x + C_{xxx} + C_t - \epsilon C_x^2,\\ F &:= -3C_x \end{aligned}$$

and using (3.3), (3.7) we get

$$\iint Du^2 + \iint Eu_x^2 + \iint Fu_{xx}^2 \le (1+\epsilon^{-1})\iint w^2. \qquad (3.8)$$

The function ψ will be chosen in such a way that D, E, and F are positive. Clearly

$$\begin{aligned} D &= -(AB)_x + \frac{1}{t^4(T-t)^4}O(s^4) \qquad (\text{as } s \to +\infty)\\ &= -\big(3(s\varphi_x)^5\big)_x + \frac{O(s^4)}{t^4(T-t)^4}\\ &= -15s^5\frac{\psi'(x)^4\psi''(x)}{t^5(T-t)^5} + \frac{O(s^4)}{t^4(T-t)^4}\cdot \end{aligned}$$

It follows that for s large enough, if

$$|\psi'(x)| > 0 \text{ and } \psi''(x) < 0 \text{ for } x \in [-L,L], \qquad (3.9)$$

we have

$$D \ge C_1\frac{s^5}{t^5(T-t)^5} \qquad (3.10)$$

for some constant $C_1 > 0$. On the other hand, expanding E in a series of powers of s, it is easy to see that there is no term in s^3 (because of cancellations) and that

$$\begin{aligned} E &= 9s^2\big((1-\epsilon)\varphi_{xx}^2 - \varphi_x\varphi_{xxx}\big) + \frac{O(s)}{t^2(T-t)^2}\\ &= 9s^2\frac{(1-\epsilon)\psi''(x)^2 - \psi'(x)\psi'''(x)}{t^2(T-t)^2} + \frac{O(s)}{t^2(T-t)^2}\cdot \end{aligned}$$

Hence for s large enough, if

$$(1-\epsilon)\psi''(x)^2 - \psi'(x)\psi'''(x) > 0 \quad \text{for all } x \in [-L,L], \qquad (3.11)$$

we get for some constant $C_2 > 0$

$$E \geq C_2 \frac{s^2}{t^2(T-t)^2}. \tag{3.12}$$

Finally, for some constant $C_3 > 0$

$$F = -\frac{9\psi''(x)s}{t(T-t)} \geq C_3 \frac{s}{t(T-t)} \tag{3.13}$$

provided that (3.9) holds true. Now pick some smooth positive function ψ on $[-L, L]$ such that (3.9) and (3.11) are fulfilled for some $\epsilon > 0$. (For instance, picking any ϵ in $(0,1)$, $\psi(x) = -x^2 + (2L+1)(x+2L)$ is convenient.) We infer from (3.8), (3.10), (3.12), and (3.13) that, for s large enough,

$$\iint \left\{ \frac{s^5}{t^5(T-t)^5} u^2 + \frac{s^2}{t^2(T-t)^2} u_x^2 + \frac{s}{t(T-t)} u_{xx}^2 \right\} \leq C_4 \iint w^2 \tag{3.14}$$

for some constant $C_4 > 0$. Actually (3.14) may be slightly improved by observing that

$$\begin{aligned} \iint \frac{s^3}{t^3(T-t)^3} u_x^2 &= -\iint \frac{s^3}{t^3(T-t)^3} u u_{xx} \\ &\leq \frac{1}{2} \left(\iint \frac{s^5}{t^5(T-t)^5} u^2 + \iint \frac{s}{t(T-t)} u_{xx}^2 \right) \\ &\leq \frac{C_4}{2} \iint w^2 \end{aligned}$$

(thanks to (3.14)); hence, for s large enough,

$$\iint \left\{ \frac{s^5}{t^5(T-t)^5} u^2 + \frac{s^3}{t^3(T-t)^3} u_x^2 + \frac{s}{t(T-t)} u_{xx}^2 \right\} \leq \frac{3}{2} C_4 \iint w^2. \tag{3.15}$$

Replacing u with $e^{-s\varphi}q$ in (3.15) we readily get (3.1) for some constant $C > 0$ and s large enough. The proof of Proposition 3.1 is complete. □

Corollary 3.2. *Let $L > 0$ and let $f = f(t,x)$ be any function in $L^2\big(\mathbb{R}_t \times (-L,L)_x\big)$ such that* $\operatorname{supp} f \subset [t_1, t_2] \times (-L, L)$, *where $-\infty < t_1 < t_2 < \infty$. Then for every $\epsilon > 0$ there exist a positive number $C = C(L, t_1, t_2, \epsilon)$ (C does not depend on f) and a function $v \in L^2\big(\mathbb{R} \times (-L,L)\big)$ such that*

$$v_t + v_x + v_{xxx} = f \text{ in } \mathcal{D}'(\mathbb{R} \times (-L,L)), \tag{3.16}$$

$$\operatorname{supp} v \subset [t_1 - \epsilon, t_2 + \epsilon] \times (-L, L), \tag{3.17}$$

$$\|v\|_{L^2\left(\mathbb{R}\times(-L,L)\right)} \leq C \|f\|_{L^2\left(\mathbb{R}\times(-L,L)\right)}. \tag{3.18}$$

Proof. Applying a translation w.r.t. time if needed, we may assume without loss of generality that $0 = t_1 - \epsilon < t_1 < t_2 < t_2 + \epsilon =: T$. We readily infer from (3.1) that for some constants $k, C_1 > 0$ and for every $q \in \mathcal{Z}$

$$\int_0^T \int_{-L}^{L} |q|^2 e^{-\frac{k}{t(T-t)}} \, dxdt \leq C_1 \int_0^T \int_{-L}^{L} |Pq|^2 \, dxdt. \tag{3.19}$$

Thus the bilinear form

$$(p,q) := \int_0^T \int_{-L}^{L} Pp\, Pq\, dxdt$$

is a scalar product on $\mathcal{Z}$. Let H denote the completion of $\mathcal{Z}$ for $(\cdot,\cdot)$. Obviously $|q|^2 e^{-\frac{k}{t(T-t)}}$ is integrable on $(0,T)\times(-L,L)$ if $q\in H$, and (3.19) holds true as well. On the other hand the linear form

$$l(q) := -\int_0^T \int_{-L}^{L} fq\, dxdt$$

is well defined and continuous on H. Indeed, using (3.19) and the assumption on the support of f, we get

$$\int_0^T \int_{-L}^{L} |fq|\, dxdt = \int_{t_1}^{t_2} \int_{-L}^{L} |fq|\, dxdt \le C_2 \|f\|_{L^2\big((t_1,t_2)\times(-L,L)\big)} \cdot (q,q)^{\frac{1}{2}} \tag{3.20}$$

for some constant $C_2>0$. It follows from the Riesz representation theorem that there exists a unique $p\in H$ such that

$$\text{for all } q\in H \quad (p,q) = l(q). \tag{3.21}$$

We set $v := P(p) \in L^2\big((0,T)\times(-L,L)\big)$. Taking $q\in\mathcal{D}\big((0,T)\times(-L,L)\big)$ as a test function in (3.21) we get

$$\langle P^*(v), q\rangle_{\mathcal{D}'(Q),\mathcal{D}(Q)} = \langle -f, q\rangle_{\mathcal{D}'(Q),\mathcal{D}(Q)},$$

where $Q=(0,T)\times(-L,L)$ and $P^*=-P$ is the (formal) adjoint to the operator P. Hence $Pv=f$ in $\mathcal{D}'(Q)$. Notice that $v\in H^1(0,T,H^{-3}(-L,L))$, since v and $v_t = f - v_{xxx} - v_x$ belong to $L^2(0,T,H^{-3}(-L,L))$; hence $v_{|_{t=0}}$ and $v_{|_{t=T}}$ are meaningful in $H^{-3}(-L,L)$. Now let $q\in\mathcal{Z}\subset H^1(0,T,H_0^3(-L,L))$. It follows from (3.21) that

$$\begin{aligned} -\int_0^T \int_{-L}^{L} fq\, dxdt &= \int_0^T \int_{-L}^{L} v(q_t + q_x + q_{xxx})\, dxdt \\ &= -\int_0^T \langle v_t + v_x + v_{xxx}, q\rangle\, dt + [\langle v,q\rangle]_{t=0}^T \\ &= -\int_0^T \int_{-L}^{L} fq\, dxdt + [\langle v,q\rangle]_{t=0}^T, \end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing $\langle\cdot,\cdot\rangle_{H^{-3}(-L,L),H_0^3(-L,L)}$. Since $q_{|_{t=0}}$ and $q_{|_{t=T}}$ may be arbitrarily chosen in $\mathcal{D}(-L,L)$, we infer that $v_{|_{t=0}} = v_{|_{t=T}} = 0$ in $H^{-3}(-L,L)$. Extending v by setting $v(t,x)=0$ for $(t,x)\notin(0,T)\times(-L,L)$, we see that (3.16), (3.17), and (3.18) hold true (with $C=C_2$). □

Remark 1. Using [19, Thm. 1.2] instead of Proposition 3.1, one may prove that the result in Corollary 3.2 also holds true for $\epsilon=0$ and that the weight $e^{-\frac{k}{t(T-t)}}$ may be dropped in the integral term of the left-hand side of (3.19). Nevertheless, the proof given here is direct and shorter, and it leads to a self-contained paper. Moreover, this proof also works for the heat equation (see below the proof of Theorem 5.2).

4. The proof of Theorem 1.3. For the comfort of the reader, we first give an outline of the proof of Theorem 1.3. In the first step, we show that we are finished if, for any $f \in L^2_{loc}(\mathbb{R}^2)$ with support in some strip $[0,T] \times \mathbb{R}$ and any $\epsilon > 0$, there exists a function $u \in L^2_{loc}(\mathbb{R}^2)$ supported in $[-\epsilon, T+\epsilon] \times \mathbb{R}$, which solves $P\,u = f$. This problem has already been solved when the whole domain $\mathbb{R}^2$ is replaced by $\mathbb{R}_t \times (-n,n)_x$, $n \geq 1$. (See Corollary 3.2.) At this stage, we are given a sequence of solutions of $P\,u = f$, which are defined on an increasing sequence of domains and are supported in $[-\epsilon, T+\epsilon] \times \mathbb{R}$. To ensure the convergence of this sequence in $L^2_{loc}(\mathbb{R}^2)$, we need an approximation theorem (Lemma 4.4), which differs from the one in [18] by a careful control on the growth of the support in time. Two technical lemmas (namely, Lemmas 4.2 and 4.3) are needed to prove the approximation theorem. The final step is a standard Mittag–Leffler's procedure.

Let $u_0 \in L^2(0,+\infty)$ and $u_T \in L^2\big((0,+\infty), e^{-2bx}dx\big)$. It is well known (see [9]) that the operator $Av = -v_{xxx}$ with domain $H^3(\mathbb{R})$ (resp., $\{v \in L^2(\mathbb{R}, e^{2bx}dx),\ Av \in L^2(\mathbb{R}, e^{2bx}dx)\}$) generates a continuous semigroup on $L^2(\mathbb{R})$ (resp., $L^2(\mathbb{R}, e^{2bx}dx)$). Thanks to the standard change of functions

$$v(t,x) = u(t, t+x) \tag{4.1}$$

we easily get two functions $u_1(t,x)$, $u_2(t,x)$ such that $u_1 \in C([0,T], L^2(\mathbb{R}))$, $u_2 \in C([0,T], L^2(\mathbb{R}, e^{2bx}dx))$, and

$$Pu_1 = Pu_2 = 0 \quad \text{on } (0,T) \times \mathbb{R},$$

$$u_1(0,x) = \begin{cases} u_0(x) & \text{for a.e. } x > 0, \\ 0 & \text{for a.e. } x < 0, \end{cases} \qquad u_2(0,x) = \begin{cases} u_T(-x) & \text{for a.e. } x < 0, \\ 0 & \text{for a.e. } x > 0. \end{cases}$$

Now set $\tilde{u}_2(t,x) = u_2(T-t, -x)$. Obviously $P\,\tilde{u}_2 = 0$ and $\tilde{u}_{2_{|t=T}} = \tilde{u}_T$ on $(0,+\infty)$. Let ϵ' be any number in $(\epsilon, \frac{T}{2})$ and let $\varphi \in C^\infty([0,T])$ be such that $\varphi(t) = 1$ for $t \leq \epsilon'$ and $\varphi(t) = 0$ for $t \geq T - \epsilon'$. The change of functions

$$u(t,x) = \varphi(t) u_1(t,x) + \big(1 - \varphi(t)\big)\tilde{u}_2(t,x) + w(t,x)$$

transforms (1.4) into

$$\begin{cases} P\,w &= \frac{d\varphi}{dt}\,(\tilde{u}_2 - u_1) \quad \text{in } \mathcal{D}'\big((0,T) \times (0,+\infty)\big), \\ w_{|t=0} &= w_{|t=T} = 0 \quad \text{on } (0,+\infty). \end{cases}$$

Setting $f(t,x) = \varphi'(t)\big(\tilde{u}_2(t,x) - u_1(t,x)\big)$, it is clear that we are finished if the following result is proved.

Proposition 4.1. *Let $f = f(t,x)$ be any function in $L^2_{loc}(\mathbb{R}^2)$ such that*

$$\operatorname{supp} f \subset [t_1, t_2] \times \mathbb{R}$$

where $0 < t_1 < t_2 < T$. Let $\epsilon \in (0, \min(t_1, T - t_2))$. Then there exists $u \in L^2_{loc}(\mathbb{R}^2)$ such that

$$P\,u = f \ \textit{in } \mathcal{D}'(\mathbb{R}^2) \ \textit{ and } \ \operatorname{supp}\, u \subset [t_1 - \epsilon, t_2 + \epsilon] \times \mathbb{R}. \tag{4.2}$$

Remark 2. The question whether Proposition 4.1 remains valid with $\epsilon = 0$ is open. Notice that the answer is negative for the heat equation. (See Remark 3 below.)

As in [18] the proof of Proposition 4.1 rests on an approximation theorem (Lemma 4.4), which in turn is obtained as a consequence of two preliminary lemmas. In what follows S_L will denote the unitary group in $L^2(-L,L)$ generated by the operator $Au = -u_{xxx} - u_x$ with domain

$$\mathcal{D}(A) = \{u \in H^3(-L,L);\ u(-L) = u(L),\ u_x(-L) = u_x(L),\ u_{xx}(-L) = u_{xx}(L)\}\cdot$$

Set $e_n(x) = \frac{1}{\sqrt{2L}} e^{in\frac{\pi}{L}x}$ for $n \in \mathbb{Z}$. e_n is an eigenvector for A associated with the eigenvalue $\omega_n = i\lambda_n$, with

$$\lambda_n = \left(n\frac{\pi}{L}\right)^3 - n\frac{\pi}{L}\cdot \tag{4.3}$$

If u_0 is any complex-valued function in $L^2(-L,L)$, decomposed as $u_0 = \sum_{n\in\mathbb{Z}} c_n e_n$, we have for every $t \in \mathbb{R}$

$$S_L(t)\,u_0 = \sum_{n\in\mathbb{Z}} e^{i\lambda_n t} c_n e_n. \tag{4.4}$$

We are now ready to state the first lemma, which may be seen as a preliminary version to the approximation theorem.

LEMMA 4.2. *Let l_1, l_2, L, t_1, t_2, T be numbers such that $0 < l_1 < l_2 < L$ and $0 < t_1 < t_2 < T$. Let $u \in L^2\left((0,T)\times(-l_2,l_2)\right)$ be such that*

$$P\,u = 0 \quad in\ (0,T)\times(-l_2,l_2) \qquad and \quad \text{supp}\ u \subset [t_1,t_2]\times(-l_2,l_2). \tag{4.5}$$

Let $\delta > 0$ with $2\delta < \min(t_1, T-t_2)$ and $\eta > 0$ be given. Then there exist $v_1, v_2 \in L^2(-L,L)$ and $v \in L^2\left((0,T)\times(-L,L)\right)$ such that

$$P\,v = 0 \quad in\ (0,T)\times(-L,L), \tag{4.6}$$

$$v(t,\cdot) = S_L(t-t_1+2\delta)v_1 \quad for\ \ t_1-2\delta < t < t_1-\delta, \tag{4.7}$$

$$v(t,\cdot) = S_L(t-t_2-\delta)v_2 \quad for\ \ t_2+\delta < t < t_2+2\delta, \tag{4.8}$$

$$\|v-u\|_{L^2((t_1-2\delta,t_2+2\delta)\times(-l_1,l_1))} < \eta. \tag{4.9}$$

Roughly speaking, (4.7)–(4.8) mean that for $t \in (t_1-2\delta, t_1-\delta)\cup(t_2+\delta, t_2+2\delta)$ v satisfies (in addition to $P\,v = 0$) the boundary conditions $v(-L) = v(L)$, $v_x(-L) = v_x(L)$, and $v_{xx}(-L) = v_{xx}(L)$.

Proof of Lemma 4.2. Set $Q = (0,T)\times(-L,L)$, $Q_\delta = (t_1-2\delta, t_2+2\delta)\times(-l_1,l_1)$. Smoothing u by convolution and multiplying the regularized function by a cut-off function (of x), we easily get a function $u' \in \mathcal{D}(\mathbb{R}^2)$ such that

$$\begin{cases} \text{supp}\ u' \subset [t_1-\delta, t_2+\delta]\times[-l_2,l_2], \\ P\,u' = 0 \ \text{ in } (0,T)\times(-l_1,l_1) \text{ and} \\ \|u'-u\|_{L^2\left((0,T)\times(-l_1,l_1)\right)} < \frac{\eta}{2}. \end{cases} \tag{4.10}$$

Let

$$\mathcal{E} = \{v \in L^2(Q);\ \exists v_1, v_2 \in L^2(-L,L) \text{ s.t. } (4.6),\ (4.7),\ \text{and } (4.8) \text{ hold true}\}.$$

The lemma is proved if we may find $v \in \mathcal{E}$ such that $\|v - u'\|_{L^2(Q_\delta)} < \frac{\eta}{2}$. We are finished if we prove $u' \in \overline{\mathcal{E}} = \mathcal{E}^{\perp\,\perp}$, where the closure and the orthogonal complement are taken in the space $L^2(Q_\delta)$. Fix a function $g \in \mathcal{E}^\perp \subset L^2(Q_\delta)$. Before proving $(u', g)_{L^2(Q_\delta)} = 0$, we begin with the following claim.

Claim 1. *Let* $\mathcal{T} = \{\varphi \in C^\infty(\mathbb{R}^2);\ \operatorname{supp} \varphi \subset [t_1 - \delta, t_2 + \delta] \times \mathbb{R}\}$. *Then there exists* $C > 0$ *such that*

$$\text{for all } \varphi \in \mathcal{T} \quad |(\varphi, g)_{L^2(Q_\delta)}| \le C \|P\varphi\|_{L^2(Q)}. \tag{4.11}$$

Proof of Claim 1. Let $\varphi \in \mathcal{T}$, and set $\psi(t) := \int_0^t S_L(t-\tau) P\varphi(\tau)\, d\tau$ for $0 \le t \le T$; that is, ψ is the (strong) solution of the following boundary initial-value problem:

$$\begin{aligned} P\psi &= P\varphi \qquad \text{in } Q,\\ \psi(t,-L) &= \psi(t,L),\\ \psi_x(t,-L) &= \psi_x(t,L),\\ \psi_{xx}(t,-L) &= \psi_{xx}(t,L),\\ \psi(0,\cdot) &= 0. \end{aligned}$$

Clearly $v := \psi - \varphi \in \mathcal{E}$ ((4.7)–(4.8) hold true with $v_1 = 0$, $v_2 = \psi(t_2 + \delta)$); hence $(\psi - \varphi, g)_{L^2(Q_\delta)} = 0$. On the other hand, it is clear that

$$\text{for all } t \in [0,T] \quad \|\psi(t)\|_{L^2(-L,L)} \le \|P\varphi\|_{L^1(0,t,L^2(-L,L))} \le \sqrt{T} \|P\varphi\|_{L^2(Q)};$$

hence

$$|(\varphi, g)|_{L^2(Q_\delta)} = |(\psi, g)|_{L^2(Q_\delta)} \le T \|g\|_{L^2(Q_\delta)} \cdot \|P\varphi\|_{L^2(Q)}.$$

This completes the proof of Claim 1. We now proceed to the next claim.

Claim 2. *There exists a function* $w \in L^2(Q)$ *such that*

$$\text{for all } \varphi \in \mathcal{T} \quad (\varphi, g)_{L^2(Q_\delta)} = (P\varphi, w)_{L^2(Q)}. \tag{4.12}$$

Proof of Claim 2. Let $\mathcal{Z} := \{(P\varphi)_{|_Q};\ \varphi \in \mathcal{T}\}$. Notice first that for any $\zeta \in \mathcal{Z}$, if $\zeta = (P\varphi_1)_{|_Q} = (P\varphi_2)_{|_Q}$ for two functions $\varphi_1, \varphi_2 \in \mathcal{T}$, then $\varphi_1 - \varphi_2 \in \mathcal{E}$; hence $(\varphi_1 - \varphi_2, g)_{L^2(Q_\delta)} = 0$. It follows that the (linear) map $\Lambda : \zeta \in \mathcal{Z} \mapsto (\varphi, g)_{L^2(Q_\delta)} \in \mathbb{R}$ (if $\zeta = (P\varphi)_{|_Q}$, $\varphi \in \mathcal{T}$) is well defined. Let H denote the closure of $\mathcal{Z}$ in $L^2(Q)$. We infer from (4.11) that Λ may be extended to H in such a way that Λ is a continuous linear form on H. It follows from Riesz representation theorem that there exists $w \in H$ such that $\Lambda(\zeta) = (\zeta, w)_{L^2(Q)}$ for all $\zeta \in H$. Then (4.12) holds true.

We are now ready to prove $(u', g)_{L^2(Q_\delta)} = 0$. Extend g and w on $\mathbb{R}^2$ to $\tilde{g}$, $\tilde{w}$ by setting

$$\begin{aligned} \tilde{g}(t,x) &= 0 \quad \text{for } (t,x) \in \mathbb{R}^2 \setminus Q_\delta,\\ \tilde{w}(t,x) &= 0 \quad \text{for } (t,x) \in \mathbb{R}^2 \setminus Q. \end{aligned}$$

Set $\Omega = (t_1 - \delta, t_2 + \delta) \times \mathbb{R}$ and let $\varphi \in \mathcal{D}(\Omega) \subset \mathcal{T}$. Obviously

$$(\varphi, g)_{L^2(Q_\delta)} = (\varphi, \tilde{g})_{L^2(\Omega)} \quad \text{and} \quad (P\varphi, w)_{L^2(Q)} = (P\varphi, \tilde{w})_{L^2(\Omega)};$$

hence it follows from (4.12) that

$$\langle P^* \tilde{w}, \varphi\rangle_{\mathcal{D}'(\Omega),\mathcal{D}(\Omega)} = \langle \tilde{g}, \varphi\rangle_{\mathcal{D}'(\Omega),\mathcal{D}(\Omega)}.$$

Thus

$$P^* \tilde{w} = \tilde{g} \text{ in } \mathcal{D}'(\Omega)$$

and

$$P^* \tilde{w} = 0 \text{ for } t_1 - \delta < t < t_2 + \delta \text{ and } |x| > l_1.$$

Since

$$\tilde{w}(t,x) = 0 \text{ for } t_1 - \delta < t < t_2 + \delta \text{ and } |x| > L$$

we infer from Holmgren's uniqueness theorem (see [5, Thm. 8.6.8]) that

$$\tilde{w}(t,x) = 0 \text{ for } t_1 - \delta < t < t_2 + \delta \text{ and } |x| > l_1. \tag{4.13}$$

Applying (4.12) to $u' \in \mathcal{T}$ and using (4.10), (4.13) we get

$$\begin{aligned}(u',g)_{L^2(Q_\delta)} &= (P\,u', w)_{L^2(Q)}\\ &= (P\,u', w)_{L^2\big((t_1-\delta,t_2+\delta)\times(-l_1,l_1)\big)}\\ &= 0.\end{aligned}$$

The proof of Lemma 4.2 is complete. □

Next result is an observability result.

LEMMA 4.3. *Let l, L, T be positive numbers such that $l < L$. Then there exists a constant $C > 0$ such that for every $u_0 \in L^2(-L,L)$, if u denotes $S_L(\cdot)u_0$, we have*

$$\|u_0\|_{L^2(-L,L)} \le C\|u\|_{L^2\big((0,T)\times(-l,l)\big)}. \tag{4.14}$$

(Hence

$$\|u\|_{L^2((0,T)\times(-L,L))} \le \sqrt{T}C\|u\|_{L^2((0,T)\times(-l,l))}.) \tag{4.15}$$

Proof. Pick $T' \in (0, \frac{T}{2})$ and $\gamma > \frac{\pi}{T'}$. Let $N \in \mathbb{N}$ be such that

$$\lambda_N - \lambda_{-N} = 2\lambda_N \ge \gamma \text{ and } (n \in \mathbb{Z},\ |n| \ge N) \Rightarrow \lambda_{n+1} - \lambda_n \ge \gamma.$$

By Ingham's inequality (see [7]) there exists a constant $C^{T'} > 0$ such that for every sequence $(a_n)_{|n|\ge N}$ of complex numbers, with $a_n = 0$ for $|n|$ large enough, the following inequality holds true:

$$\sum_{|n|\ge N} |a_n|^2 \le C^{T'} \int_{-T'}^{T'} \left| \sum_{|n|\ge N} a_n e^{-i\lambda_n t} \right|^2 dt. \tag{4.16}$$

Let $\mathcal{Z}_n := \operatorname{Span}(e_n)$ for $n \in \mathbb{Z}$ and $\mathcal{Z} = \bigoplus_{n\in\mathbb{Z}} \mathcal{Z}_n \subset L^2(-L,L)$. We define a seminorm p in $\mathcal{Z}$ by

$$p(u) := \left(\int_{-l}^{l} |u(x)|^2\,dx\right)^{\frac{1}{2}} \quad \text{for } u \in \mathcal{Z}.$$

p is clearly a norm in each $\mathcal{Z}_n$. On the other hand, if $u_0 \in \mathcal{Z} \cap (\bigoplus_{|n|<N} \mathcal{Z}_n)^\perp$ (i.e., u_0 may be written in the form $u_0 = \sum_{|n|\geq N} c_n e_n$ with $c_n = 0$ for $|n|$ large enough), then applying (4.16) (with $a_n = \frac{c_n}{\sqrt{2L}} e^{i(\lambda_n T' + n\frac{\pi}{L}x)}$) and integrating w.r.t. x on $(-l, l)$, we get

$$2l \sum_{|n|\geq N} \frac{|c_n|^2}{2L} \leq C^{T'} \int_{-l}^{l} \int_0^{2T'} \left| \sum_{|n|\geq N} e^{i\lambda_n \tau} c_n e_n(x) \right|^2 d\tau dx;$$

hence, by Fubini's theorem,

$$\|u_0\|^2_{L^2(-L,L)} \leq \frac{L}{l} C^{T'} \int_0^{2T'} p\big(S_L(\tau)u_0\big)^2 \, d\tau.$$

Finally, for any $u_0 \in L^2(-L, L)$, we have

$$\int_0^{2T'} p\big(S_L(\tau)u_0\big)^2 \, d\tau \leq \|S_L(\cdot)u_0\|^2_{L^2\left((0,2T')\times(-L,L)\right)} = 2T'\|u_0\|^2_{L^2(-L,L)}.$$

Since $T > 2T'$, it follows from [10, Thm. 5.2] that there exists a constant $C > 0$ such that (4.14) holds true for all $z_0 \in \mathcal{Z}$. We get (4.14) for all $u_0 \in L^2(-L, L)$ by a density argument. □

We now proceed to the proof of the following approximation theorem, which differs from the one in [18] by an additional property on the support.

LEMMA 4.4. *Let $n \in \mathbb{N} \setminus \{0, 1\}$ and let t_1, t_2, T be numbers such that $0 < t_1 < t_2 < T$. Let $u \in L^2\big((0,T) \times (-n, n)\big)$ be such that $Pu = 0$ in $(0,T) \times (-n, n)$ and* supp $u \subset [t_1, t_2] \times (-n, n)$. *Let $0 < \epsilon < \min(t_1, T - t_2)$. Then there exists $v \in L^2\big((0,T) \times (-n-1, n+1)\big)$ such that*

$$Pv = 0 \quad \textit{in } (0,T) \times (-n-1, n+1), \tag{4.17}$$

$$\text{supp } v \subset [t_1 - \epsilon, t_2 + \epsilon] \times (-n-1, n+1), \tag{4.18}$$

$$\|v - u\|_{L^2\left((0,T)\times(-n+1,n-1)\right)} < \epsilon. \tag{4.19}$$

Proof. Let $\eta > 0$ (to be chosen later). By Lemma 4.2, applied with $L = n + 1$, there exists $\tilde{v} \in L^2\big((0,T) \times (-n-1, n+1)\big)$ such that

$$P\tilde{v} = 0 \quad \text{in} \quad (0,T) \times (-n-1, n+1),$$

$$\tilde{v}(t, \cdot) = S_{n+1}\left(t - t_1 + \frac{\epsilon}{2}\right) v_1 \quad \text{for} \quad t_1 - \frac{\epsilon}{2} < t < t_1 - \frac{\epsilon}{4}, \tag{4.20}$$

$$\tilde{v}(t, \cdot) = S_{n+1}\left(t - t_2 - \frac{\epsilon}{4}\right) v_2 \quad \text{for} \quad t_2 + \frac{\epsilon}{4} < t < t_2 + \frac{\epsilon}{2} \tag{4.21}$$

for some $v_1, v_2 \in L^2(-n-1, n+1)$ and

$$\|\tilde{v} - u\|_{L^2\left((t_1-\frac{\epsilon}{2}, t_2+\frac{\epsilon}{2})\times(-n+1,n-1)\right)} < \eta.$$

In order that (4.18) be fulfilled, we multiply $\tilde{v}$ by a cut-off function. Let $\varphi \in \mathcal{D}(0,T)$ be such that $0 \leq \varphi \leq 1$, $\varphi(t) = 1$ for all $t \in [t_1 - \frac{\epsilon}{4}, t_2 + \frac{\epsilon}{4}]$ and $\operatorname{supp}(\varphi) \subset [t_1 - \frac{\epsilon}{2}, t_2 + \frac{\epsilon}{2}]$. Set $\bar{v}(t,x) = \varphi(t)\tilde{v}(t,x)$. It follows that

$$\operatorname{supp}\, \bar{v} \subset \left[t_1 - \frac{\epsilon}{2}, t_2 + \frac{\epsilon}{2}\right] \times (-n-1, n+1)\cdot$$

Hence

$$\begin{aligned} &\|\bar{v} - u\|_{L^2\left((0,T)\times(-n+1,n-1)\right)} \\ &\quad = \|\bar{v} - u\|_{L^2\left((t_1-\frac{\epsilon}{2},t_2+\frac{\epsilon}{2})\times(-n+1,n-1)\right)} \\ &\quad \leq \|(\varphi - 1)\tilde{v}\|_{L^2\left((t_1-\frac{\epsilon}{2},t_2+\frac{\epsilon}{2})\times(-n+1,n-1)\right)} \\ &\qquad + \|\tilde{v} - u\|_{L^2\left((t_1-\frac{\epsilon}{2},t_2+\frac{\epsilon}{2})\times(-n+1,n-1)\right)}\cdot \end{aligned}$$

Since $\operatorname{supp}\, u \subset [t_1, t_2] \times (-n, n)$ and $\varphi(t) = 1$ for $t_1 - \frac{\epsilon}{4} \leq t \leq t_2 + \frac{\epsilon}{4}$, we get

$$\begin{aligned} &\|(\varphi - 1)\tilde{v}\ \|^2_{L^2\left((t_1-\frac{\epsilon}{2},t_2+\frac{\epsilon}{2})\times(-n+1,n-1)\right)} \\ &\quad \leq \|\tilde{v}\|^2_{L^2\left(\{(t_1-\frac{\epsilon}{2},t_1-\frac{\epsilon}{4})\cup(t_2+\frac{\epsilon}{4},t_2+\frac{\epsilon}{2})\}\times(-n+1,n-1)\right)} \\ &\quad \leq \|\tilde{v} - u\|^2_{L^2\left((t_1-\frac{\epsilon}{2},t_2+\frac{\epsilon}{2})\times(-n+1,n-1)\right)} \\ &\quad < \eta^2\cdot \end{aligned} \tag{4.22}$$

Hence

$$\|\bar{v} - u\|_{L^2\left((0,T)\times(-n+1,n-1)\right)} \leq 2\eta\cdot \tag{4.23}$$

Finally $P\,\bar{v} = \frac{d\varphi}{dt}\,\tilde{v}$ in $(0,T) \times (-n-1, n+1)$; hence

$$\begin{aligned} &\|P\,\bar{v}\ \|^2_{L^2\left((0,T)\times(-n-1,n+1)\right)} \\ &\quad \leq \|\tfrac{d\varphi}{dt}\|^2_{L^\infty(0,T)} \cdot \|\tilde{v}\|^2_{L^2\left(\{(t_1-\frac{\epsilon}{2},t_1-\frac{\epsilon}{4})\cup(t_2+\frac{\epsilon}{4},t_2+\frac{\epsilon}{2})\}\times(-n-1,n+1)\right)}\cdot \end{aligned}$$

Since (4.20), (4.21) hold true, we infer from Lemma 4.3 that there exists a constant $C = C(n, \epsilon) > 0$ such that

$$\|\tilde{v}\|_{L^2\left((t_1-\frac{\epsilon}{2},t_1-\frac{\epsilon}{4})\times(-n-1,n+1)\right)} \leq C\|\tilde{v}\|_{L^2\left((t_1-\frac{\epsilon}{2},t_1-\frac{\epsilon}{4})\times(-n+1,n-1)\right)}$$

and also

$$\|\tilde{v}\|_{L^2\left((t_2+\frac{\epsilon}{4},t_2+\frac{\epsilon}{2})\times(-n-1,n+1)\right)} \leq C\|\tilde{v}\|_{L^2\left((t_2+\frac{\epsilon}{4},t_2+\frac{\epsilon}{2})\times(-n+1,n-1)\right)}\cdot$$

Hence, by (4.22),

$$\|P\,\bar{v}\|_{L^2\left((0,T)\times(-n-1,n+1)\right)} \leq C\|\frac{d\varphi}{dt}\|_{L^\infty(0,T)}\,\eta\cdot \tag{4.24}$$

We finally modify $\bar{v}$ in order that (4.17) be satisfied. By Corollary 3.2 there exist a constant $C' > 0$ (which depends on n, t_1, t_2, and ϵ) and a function $w \in L^2\big((0,T) \times (-n-1, n+1)\big)$ such that

$$\begin{gathered} P\,w = P\,\bar{v} \ \text{ in } (0,T) \times (-n-1, n+1), \\ \operatorname{supp}\, w \subset [t_1 - \epsilon, t_2 + \epsilon] \times (-n-1, n+1), \end{gathered}$$

and also

$$\|w\|_{L^2\big((0,T)\times(-n-1,n+1)\big)} \leq C'\|P\,\bar{v}\|_{L^2\big((0,T)\times(-n-1,n+1)\big)}\cdot \tag{4.25}$$

Set $v := \bar{v} - w$. Then (4.17) and (4.18) are obvious, and we infer from (4.23), (4.24), and (4.25) that

$$\|v-u\|_{L^2\big((0,T)\times(-n+1,n-1)\big)} \leq \left(2 + CC'\left\|\frac{d\varphi}{dt}\right\|_{L^\infty(0,T)}\right)\eta\cdot$$

Hence (4.19) holds true provided that η is small enough. □

We now turn to the proof of Proposition 4.1, which is carried out as in [18].

Proof of Proposition 4.1. Let $(t_1^n)_{n\geq 2}$ and $(t_2^n)_{n\geq 2}$ be two sequences of numbers such that

$$\text{for all } n\geq 2 \qquad t_1-\epsilon < t_1^{n+1} < t_1^n < t_1 < t_2 < t_2^n < t_2^{n+1} < t_2+\epsilon\cdot \tag{4.26}$$

We construct (by induction on n) a sequence $(u_n)_{n\geq 2}$ of functions such that, for every $n \geq 2$,

$$u_n \in L^2\big((0,T)\times(-n,n)\big), \tag{4.27}$$

$$\text{supp } u_n \subset [t_1^n, t_2^n]\times(-n,n), \tag{4.28}$$

$$P\,u_n = f \ \text{ in } \ (0,T)\times(-n,n), \tag{4.29}$$

and, if $n > 2$,

$$\|u_n - u_{n-1}\|_{L^2\big((0,T)\times(-n+2,n-2)\big)} < 2^{-n}\cdot \tag{4.30}$$

u_2 is given by Corollary 3.2. Now let $n \geq 2$ and assume that $u_2,\ldots,u_n$ have been constructed in such a way that (4.27)–(4.30) hold true. By Corollary 3.2 there exists $w \in L^2\big((0,T)\times(-n-1,n+1)\big)$ such that

$$\text{supp } w \subset [t_1^2, t_2^2]\times(-n-1,n+1) \ \text{ and } \ P\,w = f \ \text{ in } \ (0,T)\times(-n-1,n+1)\cdot$$

Since $P(u_n - w) = 0$ in $(0,T)\times(-n,n)$ and

$$\text{supp}\,(u_n - w_{|(0,T)\times(-n,n)}) \subset [t_1^n, t_2^n]\times(-n,n),$$

with $t_1^{n+1} < t_1^n < t_2^n < t_2^{n+1}$, it follows from Lemma 4.4 that there exists a function $v \in L^2\big((0,T)\times(-n-1,n+1)\big)$ such that

$$\text{supp } v \subset [t_1^{n+1}, t_2^{n+1}]\times(-n-1,n+1), \quad P\,v = 0 \text{ in } (0,T)\times(-n-1,n+1)$$

and also

$$\|v-(u_n-w)\|_{L^2\big((0,T)\times(-n+1,n-1)\big)} < 2^{-n-1}\cdot$$

We set $u_{n+1} := v + w$. Then (4.27)–(4.30) are fulfilled. Extending the u_n's by setting $u_n(t,x) = 0$ for $(t,x) \notin (0,T)\times(-n,n)$, we infer from (4.30) that the sequence $(u_n)_{n\geq 2}$ converges in $L^2_{loc}(\mathbb{R}^2)$ towards a function u such that

$$\text{supp } u \subset [t_1-\epsilon, t_2+\epsilon]\times\mathbb{R}$$

by (4.26) and (4.28). Finally $P\,u = f$ in $\mathbb{R}^2$, because of (4.29). This completes the proof of Proposition 4.1 and also the proof of Theorem 1.3. □

5. The heat equation and the Schrödinger equation. In this section, we are concerned with the control of the heat equation and of the Schrödinger equation in unbounded domains. Let us first briefly discuss the controllability of the heat equation

$$u_t - \Delta u = 0, \tag{5.1}$$

where $\Delta u = \sum_{i=1}^N \frac{\partial^2 u}{\partial x_i^2}$, $N \geq 1$. By Proposition 1.1, the (boundary or internal) *approximate* controllability of (5.1) in unbounded domains is obvious. Notice that this result is still valid (but not so obvious) for a *semilinear* heat equation; see [23]. As far as the boundary null-controllability of (5.1) is concerned, it has been proved in [17] that no (nontrivial) function in $\mathcal{D}(\Omega)$ (the space of test functions) can be driven to 0 when $\Omega = \mathbb{R}^N_+ := \{(x', x_N) : \ x' \in \mathbb{R}^{N-1}, \ x_N > 0\}$ and solutions of (5.1) are taken in some *transposition* sense. However, if *all* the solutions of (5.1) are taken into consideration, then the null-controllability is recovered, thanks to the following result by Jones (see [8], [13]).

Theorem 5.1. *Let $g \in C^0(\mathbb{R}^N)$ and $T > 0$. Then there exists a function $u \in C^0([0,T] \times \mathbb{R}^N)$ which solves* (5.1) *in the distributional sense for $t > 0$ and satisfies $u_{|t=0} = g$, $u_{|t=T} = 0$.*

Notice that, as it has been pointed out in [13], the boundary control problem is solved once and for all without reference to any *specific* domain or set of boundary conditions. Theorem 5.1 is derived in [8] from the existence, for any $\epsilon > 0$, of a fundamental solution of the heat equation which is supported in the strip $[0,\epsilon] \times \mathbb{R}^N$. A result close to Theorem 5.1 may be proved along the same lines as in section 4.

Theorem 5.2. *Let $\big(S(t)\big)_{t \geq 0}$ denote the continuous semigroup on $L^2(\mathbb{R}^N)$ generated by the operator $Au = \Delta u$ with domain $H^2(\mathbb{R}^N)$. Let T, ϵ be positive numbers with $\epsilon < \frac{T}{2}$ and let $u_0, u_1 \in L^2(\mathbb{R}^N)$. Then there exists a function*

$$u \in L^2_{loc}\big([0,T] \times \mathbb{R}^N)\big) \cap C\big([0,\epsilon] \cup [T-\epsilon, T], L^2(\mathbb{R}^N)\big)$$

which solves

$$\begin{cases} u_t - \Delta u &= 0 \qquad \text{in } \mathcal{D}'\big((0,T) \times \mathbb{R}^N\big), \\ u_{|t=0} &= u_0, \\ u_{|t=T} &= S(T)u_1. \end{cases} \tag{5.2}$$

Proof. Set, for any $L > 0$, $\Omega_L := (-L, L)^N$. Let P_1 denote the operator $\frac{\partial}{\partial t} - \Delta$. Let $\epsilon' \in (\epsilon, \frac{T}{2})$ and let $\varphi \in C^\infty([0,T])$ be such that $\varphi(t) = 1$ for $t \leq \epsilon'$ and $\varphi(t) = 0$ for $t \geq T - \epsilon'$. The change of functions

$$u(t,\cdot) = \varphi(t)S(t)u_0 + \big(1 - \varphi(t)\big)S(t)u_1 + w(t,\cdot)$$

transforms (5.2) into

$$\begin{cases} P_1 w &= \frac{d\varphi}{dt} S(t)(u_1 - u_0) \quad \text{ in } \mathcal{D}'\big((0,T) \times \mathbb{R}^N\big), \\ w_{|t=0} &= w_{|t=T} = 0. \end{cases}$$

Once again, it is clear that we are finished if Proposition 4.1 holds true with P_1 instead of P. The estimate (see [3, Lem. 5.2], [4, Thm. 4.1])

$$\exists k > 0, \ \exists C > 0 \text{ s.t.} \quad \int_0^T \int_{\Omega_L} |q|^2 e^{-\frac{k}{t(T-t)}} dxdt \leq C \int_0^T \int_{\Omega_L} |q_t + \Delta q|^2 \, dxdt$$

for any $q \in C^2([0,T] \times [-L,L]^N)$ such that $q(t,x) = \partial_n q(t,x) = 0$ for $(t,x) \in [0,T] \times \partial\Omega_L$ shows that Corollary 3.2 holds true with P_1 instead of P. The other key ingredient in the proof of Proposition 4.1, namely the internal observability result (Lemma 4.3), may be found in the literature (see [12, Cor. 2]). If $\big(S_L(t)\big)_{t\geq 0}$ now denotes the continuous *semigroup* on $L^2(\Omega_L)$ generated by the operator $Au = \Delta u$ with domain $H^2_{per}(\Omega_L)$, then the proofs of Lemmas 4.2 and 4.4 and of Proposition 4.1 are word for word the same as those given above for the KdV equation. □

Remark 3. For the heat equation, the results in Corollary 3.2 and in Proposition 4.1 are no longer true if we do $\epsilon = 0$. Indeed, if we assume that Corollary 3.2 is true for the one-dimensional heat equation with $\epsilon = 0$ (for any f), then an argument similar to the one used in the proof of Theorem 1.2 shows that for some constant $C > 0$ we have

$$\int_0^T \int_{-L}^L |q|^2 \, dxdt \leq C \int_0^T \int_{-L}^L |q_t - q_{xx}|^2 \, dxdt \tag{5.3}$$

for any $q \in C^2([0,T] \times [-L,L])$ such that $q(t,\pm L) = q_x(t,\pm L) = 0$. Let E be the classical fundamental solution of the one-dimensional heat equation, namely

$$E(t,x) = \begin{cases} (4\pi t)^{-\frac{1}{2}} \exp(-\frac{x^2}{4t}) & \text{if } t > 0,\ x \in \mathbb{R}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\rho \in \mathcal{D}(-L,L)$ be such that $\rho(x) = 1$ for $|x| < \frac{L}{2}$. Set $q(t,x) = \rho(x)\frac{\partial E}{\partial x}(t,x)$ for $x \in (-L,L)$ and $t > 0$. Direct computations show that $\|q\|_{L^2((\epsilon,T+\epsilon)\times(-L,L))} \to +\infty$ as $\epsilon \to 0^+$, whereas $\|q_t - q_{xx}\|_{L^2((\epsilon,T+\epsilon)\times(-L,L))} = O(1)$, contradicting (5.3).

We now turn to the Schrödinger equation

$$iu_t + \Delta u = 0. \tag{5.4}$$

For the sake of simplicity, we restrict ourselves to the one-dimensional case (i.e., $N = 1$). The following result is proved in the same way as Theorem 1.3.

THEOREM 5.3. *Let T, ϵ be positive numbers with $\epsilon < \frac{T}{2}$ and let $u_0, u_T \in L^2(\mathbb{R})$. Then there exists a function*

$$u \in L^2_{loc}\big([0,T] \times \mathbb{R}\big) \cap C\big([0,\epsilon] \cup [T-\epsilon,T], L^2(\mathbb{R})\big)$$

which solves

$$\begin{cases} iu_t + u_{xx} &= 0 \quad \text{in } \mathcal{D}'\big((0,T) \times \mathbb{R}\big), \\ u_{|t=0} &= u_0, \\ u_{|t=T} &= u_T. \end{cases} \tag{5.5}$$

Proof. We shall write P_2 for the operator $\frac{\partial}{\partial t} - i\frac{\partial^2}{\partial x^2}$. Let $\big(S(t)\big)_{t\in\mathbb{R}}$ denote the unitary group on $L^2(\mathbb{R})$ generated by the operator $Au = iu_{xx}$ with domain $H^2(\mathbb{R})$. Let ϵ' and φ be as in the proof of Theorem 5.2. The change of functions

$$u(t,\cdot) = \varphi(t)S(t)u_0 + \big(1 - \varphi(t)\big)S(t-T)u_T + w(t,\cdot)$$

transforms (5.5) into

$$\begin{cases} P_2\, w &= \frac{d\varphi}{dt}\, S(t)\big(S(-T)u_T - u_0\big) \quad \text{in } \mathcal{D}'\big((0,T) \times \mathbb{R}\big), \\ w_{|t=0} &= w_{|t=T} = 0. \end{cases}$$

We are again led to prove Proposition 4.1, but with P_2 instead of P. Let S_L denote here the unitary group on $L^2(-L,L)$ generated by the operator $Au = iu_{xx}$ with domain $\{u \in H^2(-L,L);\ u(-L) = u(L),\ u_x(-L) = u_x(L)\}$. Let $e_n(x) = \frac{1}{\sqrt{2L}} e^{in\frac{\pi}{L}x}$ and $\lambda_n = (n\frac{\pi}{L})^2$ for $n \in \mathbb{Z}$. If $u_0 \in L^2(-L,L)$ is decomposed as $u_0 = \sum_{n\in\mathbb{Z}} c_n e_n$, then $S(t)u_0 = \sum_{n\in\mathbb{Z}} e^{-i\lambda_n t} c_n e_n$ for all $t \in \mathbb{R}$.

A proof of Corollary 3.2 using a controllability result in the literature instead of a Carleman's estimate is as follows. Let $w(t) := \int_{t_1}^t S_L(t-\tau) f(\tau)\, d\tau \in L^2(-L,L)$ for $t_1 \le t \le t_2$. By [15, Thm. 1.2] (the result being in fact true for any final state $y_T \in L^2(\Omega)$ instead of 0) there exists some (internal) control function $h \in L^2((t_1,t_2)\times(L,L+1))$ such that the solution $y \in C([t_1,t_2], L^2(-L,L+1))$ of

$$\begin{cases} y_t - iy_{xx} &= h\chi_{(L,L+1)} \quad \text{in } (t_1,t_2)\times(-L,L+1),\\ y &= 0 \quad\quad\quad\ \ \text{on } (t_1,t_2)\times\{-L,L+1\},\\ y_{|t=t_1} &= 0 \end{cases}$$

satisfies $y_{|t=t_2} = -w(t_2)$ on $(-L,L)$. Clearly, the function

$$v(t) := \begin{cases} w(t) + y(t) & \text{if } t_1 \le t \le t_2,\\ 0 & \text{otherwise} \end{cases}$$

fulfills $v_t - iv_{xx} = f$ in $\mathcal{D}'(\mathbb{R}\times(-L,L))$, (3.17) (with $\epsilon = 0$) and (3.18).

As for the KdV equation the proof of Lemma 4.3 rests on Ingham's inequality and on [10, Thm. 5.2]. Here $\mathcal{Z} = \bigoplus_{n\in\mathbb{N}} \mathcal{Z}_n$, with $\mathcal{Z}_0 = \mathrm{Span}(e_0)$ and $\mathcal{Z}_n = \mathrm{Span}(e_n, e_{-n})$ for $n \ge 1$. To properly handle the left-hand side in Ingham's estimate

$$\sum_{n\ge N} |a_n|^2 \le C^{T'} \int_{-T'}^{T'} \left| \sum_{n\ge N} a_n e^{-i\lambda_n t} \right|^2 dt \tag{5.6}$$

(which is applied with $a_n = (c_n e_n + c_{-n} e_{-n}) e^{-i\lambda_n T'}$) it is sufficient to observe that the estimate

$$\|c_n e_n + c_{-n} e_{-n}\|^2_{L^2(-l,l)} \ge \frac{1}{2}\left(\|c_n e_n\|^2_{L^2(-l,l)} + \|c_{-n} e_{-n}\|^2_{L^2(-l,l)}\right)$$

holds true provided that $|n|$ is large enough, which implies (for N large enough)

$$\int_{-l}^{l} \sum_{n\ge N} |c_n e_n + c_{-n} e_{-n}|^2\, dx \ge \frac{l}{2L} \sum_{|n|\ge N} |c_n|^2 = \frac{l}{2L} \|u_0\|^2_{L^2(-L,L)}.$$

The rest of the proof of Lemma 4.3 and of Proposition 4.1 is as above for the KdV equation. $\square$

Appendix. Proof of Proposition 1.1.

We argue by contradiction and assume that $\mathcal{R} = L^2(\mathbb{R})$. Consider the map

$$\Lambda : f \in L^2\big((0,T)\times(L_1,L_2)\big) \to \int_0^T S(T-t)\tilde{f}(t,\cdot)dt \in L^2(\mathbb{R})$$

(where $\tilde{f}$ is the prolongation of f by 0 on $\mathbb{R}^2$). Let $N = \ker(\Lambda)$. Then the restriction of Λ to the orthogonal complement of N in $L^2\big((0,T)\times(L_1,L_2)\big)$ is a one-to-one continuous linear map which is onto $L^2(\mathbb{R})$; hence its inverse $(\Lambda_{|_{N^\perp}})^{-1}$ is continuous.

Let $f \in \mathcal{D}\big((0,T) \times (L_1, L_2)\big)$ and let $w_T \in D(A^*)$, where A^* denotes the adjoint of the operator A. (Clearly $A^* w = \sum_{i=0}^{n} (-1)^i a_i \frac{d^i w}{dx^i}$.) Recall that A^* generates the continuous semigroup $\big(S^*(t)\big)_{t \geq 0}$ on $L^2(\mathbb{R})$. Set $w(t) := S^*(T-t) w_T$ for $t \in [0,T]$. Then w solves

$$\begin{cases} \frac{dw}{dt} & = -A^* w, \\ w_{|_{t=T}} & = w_T. \end{cases}$$

Let $u(t) = \int_0^t S(t-s) \tilde{f}(s, \cdot) ds$. Integrating by part in

$$\int_0^T \int_{\mathbb{R}} \left(\frac{dw}{dt} + A^* w \right) u \, dxdt = 0,$$

we get

$$\int_0^T \int_{L_1}^{L_2} f(t,x) w(t,x) \, dxdt = \int_{\mathbb{R}} w_T(x) u(T,x) \, dx. \tag{A.1}$$

The same equation holds true (by density) for $f \in L^2\big((0,T) \times (L_1, L_2)\big)$ and $w_T \in L^2(\mathbb{R})$. Letting $f = (\Lambda_{|_{N^\perp}})^{-1}(w_T)$, where w_T is any function in $L^2(\mathbb{R}) \setminus \{0\}$, we get

$$\|w_T\|^2_{L^2(\mathbb{R})} \leq \|f\|_{L^2\big((0,T)\times(L_1,L_2)\big)} \cdot \|w\|_{L^2\big((0,T)\times(L_1,L_2)\big)};$$

hence

$$\|w_T\|_{L^2(\mathbb{R})} \leq \|(\Lambda_{|_{N^\perp}})^{-1}\| \cdot \|w\|_{L^2\big((0,T)\times(L_1,L_2)\big)}.$$

Replacing w_T by $w_T(\cdot + n)$ (and also $w(t,x)$ by $w(t, x+n)$), we get

$$\|w_T\|_{L^2(\mathbb{R})} = \|w_T(\cdot + n)\|_{L^2(\mathbb{R})} \leq \|(\Lambda_{|_{N^\perp}})^{-1}\| \cdot \|w\|_{L^2\big((0,T)\times(L_1+n,L_2+n)\big)}.$$

Letting $n \to \infty$, we get (by Lebesgue's theorem) $w_T = 0$, a contradiction. Thus $\mathcal{R} \neq L^2(\mathbb{R})$. Now let $w_T \in \mathcal{R}^\perp$. We infer from (A.1) that

$$\int_0^T \int_{L_1}^{L_2} f(t,x) w(t,x) \, dxdt = 0$$

for all $f \in L^2\big((0,T) \times (L_1, L_2)\big)$; hence $w_{|_{(0,T)\times(L_1,L_2)}} = 0$. Since $n \geq 2$ (with $a_n \neq 0$) it follows from Holmgren's uniqueness theorem that $w = 0$ in $(0,T) \times \mathbb{R}$; hence $w_T = 0$, and we infer that $\mathcal{R}$ is dense in $L^2(\mathbb{R})$.

Acknowledgment. The author wishes to thank E. Zuazua, who introduced to him Jones and his work.

REFERENCES

[1] J. Bona and R. Winther, *The Korteweg–de Vries equation, posed in a quarter-plane*, SIAM J. Math. Anal., 14 (1983), pp. 1056–1106.

[2] J. M. Coron, *On the controllability of* 2*-D incompressible perfect fluids*, J. Math. Pures Appl. (9), 75 (1996), pp. 155–188.

[3] E. Fernández-Cara, *Null controllability of the semilinear heat equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–103; also available online from http://www.emath.fr/cocv/.

[4] A. V. Fursikov and O. Y. Imanuvilov, *On controllability of certain systems simulating a fluid flow*, in Flow Control, IMA Vol. Math. Appl. 68, M. D. Gunzburger, ed., Springer-Verlag, New York, 1995, pp. 149–184.

[5] L. Hörmander, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, New York, 1983.

[6] T. Horsin, *On the controllability of the Burgers equation*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 83–95; also available online from http://www.emath.fr/cocv/.

[7] A. E. Ingham, *Some trigonometrical inequalities with application to the theory of series*, Math. A., 41 (1936), pp. 367–379.

[8] B. F. Jones, Jr., *A fundamental solution of the heat equation which is supported in a strip*, J. Math. Anal. Appl., 60 (1977), pp. 314–324.

[9] T. Kato, *On the Cauchy problem for the (generalized) Korteweg-de Vries equations*, in Studies in Applied Mathematics, Adv. Math. Suppl. Stud. 8, Academic Press, New York, 1983, pp. 93–128.

[10] V. Komornik, *Exact Controllability and Stabilization, the Multiplier Method*, RAM Res. Appl. Math. 36, John Wiley–Masson, Chichester, UK, Paris, 1994.

[11] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl. (9), 71 (1992), pp. 267–291.

[12] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.

[13] W. Littman, *Boundary control theory for hyperbolic and parabolic partial differential equations with constant coefficients*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 3 (1978), pp. 567–580.

[14] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, Vol. 1, Dunod, Paris, 1968.

[15] E. Machtyngier, *Exact controllability for the Schrödinger equation*, SIAM J. Control Optim., 32 (1994), pp. 24–34.

[16] S. Micu, *On the controllability of the linearized Benjamin-Bona-Mahony equation*, SIAM J. Control Optim., submitted.

[17] S. Micu and E. Zuazua, *On the lack of null-controllability of the heat equation on the half space*, Portugal. Math., to appear.

[18] J. P. Rosay, *A very elementary proof of the Malgrange-Ehrenpreis theorem*, Amer. Math. Monthly, 98 (1991), pp. 518–523.

[19] L. Rosier, *Exact boundary controllability for the Korteweg-de Vries equation on a bounded domain*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 33–55; also available online from http://www.emath.fr/cocv/.

[20] L. Rosier, *Exact boundary controllability for the linear Korteweg-de Vries equation—a numerical study*, in Control and Partial Differential Equations, ESAIM Proc. 4, Société de Mathématiques Appliquées et Industrielles, Paris, 1998, pp. 255–267; also available online from http://www.emath.fr/Maths/Proc/.

[21] D. L. Russell and B. Y. Zhang, *Controllability and stabilizability of the third-order linear dispersion equation on a periodic domain*, SIAM J. Control Optim., 31 (1993), pp. 659–676.

[22] D. L. Russell and B. Y. Zhang, *Exact controllability and stabilizability of the Korteweg-de Vries equation*, Trans. Amer. Math. Soc., 348 (1996), pp. 3643–3672.

[23] L. de Teresa and E. Zuazua, *Approximate controllability of a semilinear heat equation in unbounded domains*, Nonlinear Anal., 37 (1999), pp. 1059–1090.

SIAM J. Control Optim.
Vol. 39, No. 2, pp. 352–358

THE SOLUTION TO THE CONJECTURE ON PROPERNESS OF WEAKLY RELAXED DELAYED CONTROLS*

JAVIER F. ROSENBLUETH†

Abstract. In 1986 Warga proposed a "weak" relaxation procedure applicable to fully nonlinear problems with delays in the control variables and showed that the resulting relaxed problem has a solution. However, in the event of commensurate delays, several examples were found for which weakly relaxed controls cannot be approximated with original controls, so that this extension fails to be "proper." Although the case of commensurate delays was solved by the introduction of a "strong" model, the question of how to properly relax noncommensurately delayed controls has remained open, and a natural candidate has been precisely that of weakly relaxed controls. In this paper, a general counterexample is constructed which rules out the weak relaxation as a proper relaxation when there are two or more delays. It is hoped that this result will provide some insight into the problem of finding a general representation of properly relaxed controls.

Key words. optimal control problems, systems with time delays, proper relaxation procedures

AMS subject classifications. 49A10, 49A50, 49D20

PII. S0363012998333001

1. Introduction.

This paper concerns the problem of finding a proper relaxation procedure for optimal control problems with nonadditively coupled delays (or, more generally, shifts) in the control variables. Usually in relaxation theory the aim is to find a relaxation procedure for a given original problem which leads us to its proper extension. This means that the set of controls for the original problem is dense in the space of controls for the relaxed problem. Such procedures are well studied for delay free problems. For problems with delays the situation is more difficult and less understood. For example, there are natural relaxation procedures which give a proper extension while there are other natural relaxation procedures which do not (see [2, 8]).

Research in this area starts in [11]. There are two basic problems:

1. the existence of a proper relaxation, and
2. the representation of the set of relaxed controls when it exists.

Problem 1 has been affirmatively answered in [11] when there is only one constant delay, in [8] when the constant delays are commensurate, in [13] for arbitrary constant delays, and in [12] for certain variable delays.

This paper deals with problem 2. Specifically, a general counterexample is constructed which rules out the weak relaxation proposed in [11] as a proper relaxation when there are two or more delays. This is a significant step forward in this direction which finally clarifies questions raised in [1, 2, 3, 4, 5, 6, 7, 8, 11]. In order to understand this contribution, let us briefly state the problem we shall be concerned with together with some basic notation and previous results.

For $T \subset \mathbf{R}$ compact and R a compact metric space, denote by $\mathcal{M}(T, R)$ the space of measurable functions mapping T to $rpm(R)$ with the weak star topology of $L^1(T, C(R))^*$, where $rpm(R)$ is the space of Radon probability measures on R with

*Received by the editors January 26, 1998; accepted for publication (in revised form) March 14, 2000; published electronically August 31, 2000.

http://www.siam.org/journals/sicon/39-2/33300.html

†IIMAS–UNAM, Apartado Postal 20-726, México DF 01000, México (jfrl@leibniz.iimas.unam.mx).

the weak star topology of $C(R)^*$. Let $\mathcal{U}(T,R)$ be the space of measurable functions mapping T to R, embedded in $\mathcal{M}(T,R)$ by identifying each $u \in \mathcal{U}(T,R)$ with the function $t \mapsto \delta_{u(t)}$, where δ_a (also written as δa) is the Dirac measure at a.

It is well known that $\mathcal{M}(T,R)$ coincides with $\mathrm{cl}\,\mathcal{U}(T,R)$, the weak star closure of $\mathcal{U}(T,R)$. For optimal control problems where $\mathcal{U}(T,R)$ is the space of ordinary controls, under the usual assumptions on the data of the problem, existence of minimizers in the space $\mathcal{M}(T,R)$ of relaxed controls can thus be assured, and they can be approximated with ordinary controls. This fact is summarized by saying that $\mathcal{M}(T,R)$ provides a "proper" relaxation procedure for $\mathcal{U}(T,R)$. For a full account of these ideas, together with a thorough study of relaxation and its importance in optimal control theory, we refer to Warga's book [10].

For optimal control problems involving delays in the controls, several attempts have been made to find proper relaxation procedures. The space of *ordinary delayed controls* we shall consider (also studied in [1, 2, 3, 4, 5, 6, 7, 8]), which illustrates the main difficulties encountered when addressing relaxation questions, is given by

$$\mathcal{U}(\theta_1,\dots,\theta_k) = \{(u_0,u_1,\dots,u_k) \in \mathcal{U}(T,\Omega^{k+1}) \mid u_i(t) = u_{i-1}(t-\Delta_i) \\ \text{almost everywhere (a.e.) in } T_i \ (i=1,\dots,k)\},$$

where $T=[0,1]$, $0<\theta_1<\cdots<\theta_k<1$ are given real numbers, Ω is a given compact metric space, and $\theta_0=0$, $\Delta_i=\theta_i-\theta_{i-1}$, $T_i=[\Delta_i,1]$ $(i=1,\dots,k)$.

Warga [11] proposed (in a more general setting) a natural extension of $\mathcal{U}(\theta_1,\dots,\theta_k)$, which we call the *weak relaxation procedure*, given by

$$\mathcal{M}_{\mathrm{w}}(\theta_1,\dots,\theta_k) = \{\mu \in \mathcal{M}(T,\Omega^{k+1}) \mid \mathcal{P}_i\mu(t) = \mathcal{P}_{i-1}\mu(t-\Delta_i) \text{ a.e. in } T_i \ (i=1,\dots,k)\},$$

where if, say, $\mu \in \mathcal{M}(T,\Omega^n)$ for some $n \in \mathbf{N}$ and $S \subset \{0,1,\dots,n-1\}$, then $\mathcal{P}_S\mu(t)$ denotes the projection onto the S coordinates of $\mu(t)$. Equivalently, $\mu \in \mathcal{M}(T,\Omega^{k+1})$ is a weakly relaxed control if and only if

$$\int_{T_i} dt \int \varphi(t,r_i)\mu(t)(dr) = \int_{T_i} dt \int \varphi(t,r_{i-1})\mu(t-\Delta_i)(dr)$$

for all φ in $L^1(T,C(\Omega))$ and $i=1,\dots,k$, where $r=(r_0,\dots,r_k)$.

One readily verifies that the set of weakly relaxed controls contains the set of ordinary controls and, regarding it as a subspace of $\mathcal{M}(T,\Omega^{k+1})$ with the weak star topology, it is compact. In [11] the question of properness of this model (that is, if the equality $\mathcal{M}_{\mathrm{w}}(\theta_1,\dots,\theta_k) = \mathrm{cl}\,\mathcal{U}(\theta_1,\dots,\theta_k)$ holds) was posed but could not be proved. For the one delay case, it is shown in [3, 8] that this model is indeed a proper relaxation procedure.

In [8] Rosenblueth and Vinter introduced an abstract relaxation procedure, which we call the $\mathcal{D}$-model, and properness of this procedure was established by Warga and Zhu [13]. However, as we point out in [9], determining the set of $\mathcal{D}$-relaxed controls for specific problems is a very difficult and perhaps even a hopeless task, so there is a need to find more concrete characterizations of the closure of the space of ordinary delayed controls.

Now, the conjecture mentioned in our title, considered in [1, 2, 3, 4, 5, 6, 7, 8, 11], is that

$$\textit{for any } \Omega \subset \mathbf{R}^m \textit{ compact and } 0<\theta_1<\theta_2<1, \ \mathcal{M}_{\mathrm{w}}(\theta_1,\theta_2) = \mathrm{cl}\,\mathcal{U}(\theta_1,\theta_2).$$

This statement is false. In [8], Rosenblueth and Vinter exhibit an element of $\mathcal{M}_{\mathrm{w}}(\theta_1,\theta_2)$ lying outside $\mathrm{cl}\,\mathcal{U}(\theta_1,\theta_2)$ for the case when $\theta_2=2\theta_1$ and $\Omega=[0,1]$.

Rosenblueth [4] extends the inequality $\mathcal{M}_{\rm w}(\theta_1,\theta_2) \neq \operatorname{cl} \mathcal{U}(\theta_1,\theta_2)$ to pairs different than $(\theta_1, 2\theta_1)$ and whose quotient is a rational number (this settles a question raised by Andrews [1]). In the event of commensurate delays, a "strong" relaxation procedure was introduced in [8] and shown to be proper (see also [4, 9]). For the noncommensurate case, the problem of how to characterize $\mathcal{D}$-relaxed controls has remained unsolved, but a natural candidate has been precisely the space of weakly relaxed controls (see [1, 5, 6]). In particular, it is shown in [6] that, if $\Omega = \{0,1\}$ and θ_1/θ_2 is irrational, then any constant element of $\mathcal{M}_{\rm w}(\theta_1,\theta_2)$ can be approximated with elements of $\mathcal{U}(\theta_1,\theta_2)$. The present paper finally solves this question. We show that inequality also holds for noncommensurate θ_1, θ_2:

For any $0 < \theta_1 < \theta_2 < 1$, *there exists* $\mu \in \mathcal{M}_{\rm w}(\theta_1,\theta_2)$ *such that* $\mu \notin \operatorname{cl} \mathcal{U}(\theta_1,\theta_2)$.

It is hoped that the counterexample used to solve this question will provide some insight into the problem of finding a general representation of properly relaxed controls.

2. The solution to the conjecture.

In the following theorem we assume that $\Omega = [0,1]$. Essentially the same arguments apply if $\Omega \subset \mathbf{R}$ is any compact set containing at least two points.

THEOREM 2.1. *For any* $0 < \theta_1 < \theta_2 < 1$ *there exists* $\mu \in \mathcal{M}_{\rm w}(\theta_1,\theta_2)$ *such that* $\mu \notin \operatorname{cl} \mathcal{U}(\theta_1,\theta_2)$.

Proof. Let $0 < \theta_1 < \theta_2 < 1$ and set $\alpha := \theta_2 - \theta_1$. For all $(u,v,w) \in \Omega^3$ let

$$h(u,v,w) := \min\{|(u, v-1, w-1)|, |(u-1, v, w)|\},$$
$$g(u,v,w) := \min\{|(u, v, w-1)|, |(u-1, v-1, w)|\},$$

and, for any $t \in T$, $x_0, x_1 \in \mathbf{R}$, and $(u,v,w) \in \Omega^3$, let

$$f(t,x_0,x_1,u,v,w) := \begin{cases} (x_0 - t/2)^2 + h(u,v,w) & \text{if } t \in [0,\theta_2), \\ (x_0 - t/2)^2 + g(u,v,w) & \text{if } t \in [\theta_2, 1]. \end{cases}$$

Consider the problem (P) of minimizing $x_1(1)$ subject to

$$\begin{cases} (\dot x_0(t), \dot x_1(t)) = (u(t), f(t, x_0(t), x_1(t), u(t), v(t), w(t))) \text{ a.e. in } T, \\ (x_0(0), x_1(0)) = (0,0), \\ (u,v,w) \in \mathcal{U}(\theta_1,\theta_2). \end{cases}$$

Let $\mu \in \mathcal{M}(T, \Omega^3)$ be given by

$$\mu(t) = \begin{cases} \frac{1}{2}\delta(0,1,1) + \frac{1}{2}\delta(1,0,0) & \text{if } t \in [0,\theta_2), \\ \frac{1}{2}\delta(0,0,1) + \frac{1}{2}\delta(1,1,0) & \text{if } t \in [\theta_2, 1]. \end{cases}$$

Since

$$\mathcal{P}_i\mu(t) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1 \qquad \text{for all } t \in T \text{ and } i = 0,1,2,$$

we have $\mu \in \mathcal{M}_{\rm w}(\theta_1,\theta_2)$. Note that its corresponding cost is zero and, since the cost cannot be negative, the minimum of the problem posed on $\mathcal{M}_{\rm w}(\theta_1,\theta_2)$ is zero.

Let $0 < a < \min\{\alpha, \theta_1, 1-\theta_2\}$ so that the intervals $I := [0,a)$, $I+\alpha$, and $I+\theta_2$ are disjoint and belong to $[0,1)$. Let (x_0,x_1,u,v,w) be any admissible original process for (P), so that $(u,v,w) \in \mathcal{U}(T,\Omega^3)$ and

$$v(t) = u(t-\theta_1) \text{ a.e. in } [\theta_1, 1], \tag{2.1}$$
$$w(t) = v(t-\alpha) \text{ a.e. in } [\alpha, 1], \tag{2.2}$$

and observe that the following relations hold a.e. in I:

$$w(t+\alpha) = v(t), \quad v(t+\theta_2) = u(t+\alpha), \quad w(t+\theta_2) = u(t).$$

Indeed, by (2), $w(t+\alpha) = v(t)$ a.e. in $[0, 1-\alpha]$ and, since $a < \theta_1 < 1-\theta_2+\theta_1 = 1-\alpha$, we have $I \subset [0, 1-\alpha]$. By (1), $v(t+\theta_2) = u(t+\alpha)$ a.e. in $[-\alpha, 1-\theta_2] \supset I$. Finally, by (2), $w(t+\theta_2) = v(t+\theta_1)$ a.e. in $[-\theta_1, 1-\theta_2]$ and, by (1), $v(t+\theta_1) = u(t)$ a.e. in $[0, 1-\theta_1]$. Hence $w(t+\theta_2) = u(t)$ a.e. in $[0, 1-\theta_2] \supset I$ and the three relations hold a.e. in I.

Therefore,

$$\begin{aligned} x_1(1) &= \int_0^{\theta_2} \{(x_0(s) - s/2)^2 + h(u(t), v(t), w(t))\} dt \\ &\quad + \int_{\theta_2}^1 \{(x_0(s) - s/2)^2 + g(u(t), v(t), w(t))\} dt \\ &\geq \int_0^a h(u(t), v(t), w(t)) dt + \int_\alpha^{\alpha+a} h(u(t), v(t), w(t)) dt \\ &\quad + \int_{\theta_2}^{\theta_2+a} g(u(t), v(t), w(t)) dt \\ &= \int_0^a \{h(u(t), v(t), w(t)) + h(u(t+\alpha), v(t+\alpha), v(t)) \\ &\quad + g(u(t+\theta_2), u(t+\alpha), u(t))\} dt. \end{aligned}$$

Fix $t \in I$ and let $r_0 = u(t)$, $r_1 = v(t)$, $r_2 = u(t+\alpha)$, $s_0 = w(t)$, $s_1 = v(t+\alpha)$, and $s_2 = u(t+\theta_2)$. Define

$$\varphi(t) := h(r_0, r_1, s_0) + h(r_2, s_1, r_1) + g(s_2, r_2, r_0),$$

and let

$$m_0 := \begin{cases} 1 & \text{if } h(r_0, r_1, s_0) = |(r_0, r_1 - 1, s_0 - 1)|, \\ 0 & \text{if } h(r_0, r_1, s_0) = |(r_0 - 1, r_1, s_0)|, \end{cases}$$

$$m_1 := \begin{cases} 1 & \text{if } h(r_2, s_1, r_1) = |(r_2, s_1 - 1, r_1 - 1)|, \\ 0 & \text{if } h(r_2, s_1, r_1) = |(r_2 - 1, s_1, r_1)|, \end{cases}$$

$$m_2 := \begin{cases} 1 & \text{if } g(s_2, r_2, r_0) = |(s_2, r_2, r_0 - 1)|, \\ 0 & \text{if } g(s_2, r_2, r_0) = |(s_2 - 1, r_2 - 1, r_0)|. \end{cases}$$

Observe now that

$$\begin{aligned} m_0 \neq m_1 &\Rightarrow \varphi(t) \geq |1 - r_1| + |r_1|, \\ m_1 \neq m_2 &\Rightarrow \varphi(t) \geq |1 - r_2| + |r_2|, \\ m_0 = m_2 &\Rightarrow \varphi(t) \geq |1 - r_0| + |r_0|. \end{aligned}$$

On the other hand, if $m_0 = m_1$ and $m_1 = m_2$, then $m_0 = m_2$ and so, in all cases, $\varphi(t) \geq 1$. It follows that

$$x_1(1) \geq \int_0^a \varphi(t) dt \geq a > 0$$

and so the infimum of (P) posed over the original admissible processes is positive. This implies that μ cannot be approximated with elements of $\mathcal{U}(\theta_1, \theta_2)$. □

3. Extensions to other delay-relaxation problems. The term "weak relaxation procedure" was first introduced in [8] referring to the model proposed by Warga in [11]. To be exact, the latter is slightly different from the one studied in [8] and mentioned in section 1. The subtle difference, which we shall explain below, leads to the study of another delay-relaxation problem for which a proof similar to the one of Theorem 2.1 can be applied.

Consider the following optimal control problem involving constant delays in the control variables. Let $T := [0,1]$ and suppose we are given real numbers $0 < \theta_1 < \cdots < \theta_k < 1$, a point $\xi \in \mathbf{R}^n$, a compact set $\Omega \subset \mathbf{R}^m$, and functions g mapping $\mathbf{R}^n$ to $\mathbf{R}$ and f mapping $T \times \mathbf{R}^n \times \mathbf{R}^{m(k+1)}$ to $\mathbf{R}^n$. Let $\hat{T} := [-\theta_k, 1]$ and consider the problem (P) of minimizing $g(x(1))$ subject to

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t), u(t-\theta_1), \ldots, u(t-\theta_k)) \text{ a.e. in } T, \\ x(0) = \xi, \\ u(t) \in \Omega \text{ a.e. in } \hat{T}, \end{cases}$$

where u is any measurable function mapping $\hat{T}$ to $\mathbf{R}^m$.

In [11] Warga reformulated this "original control problem" (P) by treating the control functions as independent variables which satisfy certain compatibility conditions in terms of the delays. The model of relaxation proposed by Warga was obtained by generalizing these conditions in the corresponding space of relaxed controls. To be specific, let

$$\begin{aligned} &\mathcal{W}(\theta_1, \ldots, \theta_k) \\ &\quad := \{(u_0, u_1, \ldots, u_k) \in \mathcal{U}(\hat{T}, \Omega^{k+1}) \mid u_i(t) = u_0(t-\theta_i) \text{ a.e. in } T \ (i = 1, \ldots, k)\} \end{aligned}$$

and consider the problem (W) of minimizing $g(x(1))$ subject to

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t)) \text{ a.e. in } T, \\ x(0) = \xi, \\ u \in \mathcal{W}(\theta_1, \ldots, \theta_k). \end{cases}$$

As Warga mentions in [11], it is a simple fact to show that (P) and (W) are equivalent. The "weak" extension of $\mathcal{W}(\theta_1, \ldots, \theta_k)$ proposed by Warga is given by

$$\mathcal{S}_{\mathrm{w}}(\theta_1, \ldots, \theta_k) := \{\mu \in \mathcal{M}(\hat{T}, \Omega^{k+1}) \mid \mathcal{P}_i\mu(t) = \mathcal{P}_0\mu(t-\theta_i) \text{ a.e. in } T \ (i = 1, \ldots, k)\}.$$

In [8] Rosenblueth and Vinter considered the problem (see the notation of section 1), which we label (RV), of minimizing $g(x(1))$ subject to

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t)) \text{ a.e. in } T, \\ x(0) = \xi, \\ u \in \mathcal{U}(\theta_1, \ldots, \theta_k). \end{cases}$$

It should be noted that, in this reformulation of the problem, ordinary controls are measurable functions defined on the interval $T = [0,1]$ and not on $\hat{T} = [-\theta_k, 1]$ as in the reformulation (W) of (P) given in [11]. As before, (P) and (RV) are equivalent (see [7] for details). The notion of "weakly relaxed controls" applied to (RV) yields the set

$$\mathcal{M}_{\mathrm{w}}(\theta_1, \ldots, \theta_k) = \{\mu \in \mathcal{M}(T, \Omega^{k+1}) \mid \mathcal{P}_i\mu(t) = \mathcal{P}_{i-1}\mu(t-\Delta_i) \text{ a.e. in } T_i \ (i = 1, \ldots, k)\}$$

and, since the three problems are equivalent, so is the question of properness of the two models of relaxed controls.

Now, in [5, 6] we studied a similar but larger space of ordinary controls to which the notion of weakly relaxed controls can also be applied. A different class of optimal control problems is derived from these spaces, and the question of properness of the weak model with respect to the larger class of ordinary controls can be posed. Consider the space of original controls

$$\begin{aligned}&\mathcal{U}'(\theta_1,\ldots,\theta_k)\\&\quad=\{(u_0,u_1,\ldots,u_k)\in\mathcal{U}(T,\Omega^{k+1})\mid u_i(t)=u_0(t-\theta_i)\text{ a.e. in }[\theta_i,1]\ (i=1,\ldots,k)\}.\end{aligned}$$

The notion of weakly relaxed controls applied to problem (R) of minimizing $g(x(1))$ subject to

$$\begin{cases}\dot{x}(t)=f(t,x(t),u(t))\text{ a.e. in }T,\\ x(0)=\xi,\\ u\in\mathcal{U}'(\theta_1,\ldots,\theta_k)\end{cases}$$

corresponds to

$$\mathcal{M}'_{\mathrm{w}}(\theta_1,\ldots,\theta_k):=\{\mu\in\mathcal{M}(T,\Omega^{k+1})\mid\mathcal{P}_i\mu(t)=\mathcal{P}_0\mu(t-\theta_i)\text{ a.e. in }[\theta_i,1]\ (i=1,\ldots,k)\},$$

which we shall call the space of *$\mathcal{R}$-weakly relaxed controls*, and the open question has been, again, if the relation $\mathcal{M}'_{\mathrm{w}}(\theta_1,\ldots,\theta_k)=\mathrm{cl}\,\mathcal{U}'(\theta_1,\ldots,\theta_k)$ holds.

Note that (R) is similar to (P) but not equivalent. The definition of $\mathcal{U}'(\theta_1,\ldots,\theta_k)$ as the space of ordinary delayed controls does not correspond to a reformulation of (P) and, as one can easily show,

$$\begin{aligned}&\mathcal{U}'(\theta_1,\ldots,\theta_k)\\&\quad=\{(u_0,u_1,\ldots,u_k)\in\mathcal{U}(T,\Omega^{k+1})\mid u_i(t)=u_{i-1}(t-\Delta_i)\text{ a.e. in }[\theta_i,1]\ (i=1,\ldots,k)\}.\end{aligned}$$

Comparing with the definition of the set $\mathcal{U}(\theta_1,\ldots,\theta_k)$, it is clear that $\mathcal{U}(\theta_1,\ldots,\theta_k)\subset\mathcal{U}'(\theta_1,\ldots,\theta_k)$, but the two sets may not coincide.

In [7] we proved an important consequence of this fact by exhibiting an element of both $\mathcal{M}_{\mathrm{w}}(\theta_1,\theta_2)$ and $\mathcal{M}'_{\mathrm{w}}(\theta_1,\theta_2)$ which belongs to the weak star closure of $\mathcal{U}'(\theta_1,\theta_2)$ but not to that of $\mathcal{U}(\theta_1,\theta_2)$. Also in [7] we showed that, for certain delays, the space of $\mathcal{R}$-weakly relaxed controls does provide a proper relaxation procedure. The result proved in [7] states that, given $0<\theta_1<\theta_2<1$ and $\Omega\subset\mathbf{R}^m$ compact, if $\theta_1\geq 1/2$, then $\mathcal{M}'_{\mathrm{w}}(\theta_1,\theta_2)=\mathrm{cl}\,\mathcal{U}'(\theta_1,\theta_2)$ and, if $\theta_1<1/2$ and θ_1 and θ_2 are commensurate, then $\mathrm{cl}\,\mathcal{U}'(\theta_1,\theta_2)$ may be strictly contained in $\mathcal{M}'_{\mathrm{w}}(\theta_1,\theta_2)$.

A new result for this model can now be obtained with a proof similar to the one of Theorem 2.1. The foregoing arguments can be applied to the conjecture stated in terms of $\mathcal{M}'_{\mathrm{w}}(\theta_1,\theta_2)$ and $\mathcal{U}'(\theta_1,\theta_2)$. If $\theta_1+\theta_2<1$, with a problem similar to the previous one, it is not difficult to see that

$$\mu(t)=\begin{cases}\frac{1}{2}\delta(0,1,1)+\frac{1}{2}\delta(1,0,0) & \text{if } t\in[0,\theta_1+\theta_2),\\ \frac{1}{2}\delta(0,0,1)+\frac{1}{2}\delta(1,1,0) & \text{if } t\in[\theta_1+\theta_2,1]\end{cases}$$

belongs to $\mathcal{M}'_{\mathrm{w}}(\theta_1,\theta_2)$ but not to $\mathrm{cl}\,\mathcal{U}'(\theta_1,\theta_2)$. We state this result.

THEOREM 3.1. *For any $0<\theta_1<\theta_2<1$ with $\theta_1+\theta_2<1$ there exists $\mu\in\mathcal{M}'_{\mathrm{w}}(\theta_1,\theta_2)$ such that $\mu\notin\mathrm{cl}\,\mathcal{U}'(\theta_1,\theta_2)$.*

REFERENCES

[1] T. Andrews, *An Existence Theory for Optimal Control Problems with Time Delays*, Ph.D. thesis, Imperial College, University of London, London, UK, 1989.

[2] J. F. Rosenblueth, *Strongly and weakly relaxed controls for time delay systems*, SIAM J. Control Optim., 30 (1992), pp. 856–866.

[3] J. F. Rosenblueth, *Proper relaxation of optimal control problems*, J. Optim. Theory Appl., 74 (1992), pp. 509–526.

[4] J. F. Rosenblueth, *Approximation of strongly relaxed minimizers with ordinary delayed controls*, Appl. Math. Optim., 32 (1995), pp. 33–46.

[5] J. F. Rosenblueth, *Relaxation of delayed controls: A review*, IMA J. Math. Control Inform., 12 (1995), pp. 181–206.

[6] J. F. Rosenblueth, *Constant relaxed controls with noncommensurate delays*, IMA J. Math. Control Inform., 13 (1996), pp. 195–209.

[7] J. F. Rosenblueth, *Certain classes of properly relaxed delayed controls*, IMA J. Math. Control Inform., to appear.

[8] J. F. Rosenblueth and R. B. Vinter, *Relaxation procedures for time delay systems*, J. Math. Anal. Appl., 162 (1991), pp. 542–563.

[9] J. F. Rosenblueth, J. Warga, and Q. J. Zhu, *On the characterization of properly relaxed delayed controls*, J. Math. Anal. Appl., 209 (1997), pp. 274–290.

[10] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[11] J. Warga, *Nonadditively Coupled Delayed Controls*, private communication, 1986.

[12] J. Warga, *A proper relaxation of controls with variable shifts*, J. Math. Anal. Appl., 196 (1995), pp. 783–793.

[13] J. Warga and Q. J. Zhu, *A proper relaxation of shifted and delayed controls*, J. Math. Anal. Appl., 169 (1992), pp. 546–561.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 359–410

REGULAR SYNTHESIS AND SUFFICIENCY CONDITIONS FOR OPTIMALITY[*]

BENEDETTO PICCOLI[†] AND HÉCTOR J. SUSSMANN[‡]

Abstract. We propose a definition of "regular synthesis" that is more general than those suggested by other authors such as Boltyanskii [SIAM J. Control Optim, 4 (1966), pp. 326–361] and Brunovský [Math. Slovaca, 28 (1978), pp. 81–100], and an even more general notion of "regular presynthesis." We give a complete proof of the corresponding sufficiency theorem, a slightly weaker version of which had been stated in an earlier article, with only a rough outline of the proof. We illustrate the strength of our result by showing that the optimal synthesis for the famous Fuller problem satisfies our hypotheses. We also compare our concept of synthesis with the simpler notion of a "family of solutions of the closed-loop equation arising from an optimal feedback law," and show by means of examples why the latter is inadequate, and why the difficulty cannot be resolved by using other concepts of solution—such as Filippov solutions, or the limits of sample-and-hold solutions recently proposed as feedback solutions by Clarke et al. [IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407]—for equations with a non-Lipschitz and possibly discontinuous right-hand side.

Key words. optimal control, regular synthesis, sufficient conditions

AMS subject classifications. 93C15, 49K15

PII. S0363012999322031

1. Introduction.

The purpose of the theory of "regular synthesis" is to turn into a rigorous, precise theorem the vague assertion that "if a collection Γ of extremals covers the whole space in a smooth enough way, then all the members of Γ are optimal." Naturally, a good definition should at the same time be stringent enough to lead to a correct proof of the sufficiency assertion, and broad enough to cover as large as possible a class of optimal control problems. In particular, it would be desirable for the theory to apply to well-known examples such as the famous "Fuller's problem" (cf. Fuller [13], Marchal [15], Zelikin–Borisov [32]), where the optimal controls have infinitely many switchings.

The three main goals of this paper are (a) to propose a new definition of regular synthesis, more general than those suggested by other authors (e.g., Boltyanskii [2], Brunovský [5], [6]), as well as an even more general notion of "regular presynthesis," (b) to give a detailed statement and proof of the corresponding sufficiency theorem, a slightly weaker version of which had been announced by one of us in an earlier article (Sussmann [31]), with only a rough sketch of the proof, and (c) to compare our definition with other concepts of synthesis such as those of Boltyanskii and Brunovský, and with the simpler notion of a "family of solutions of the closed-loop equation arising from an optimal feedback law."

Our main sufficiency result is Theorem 2.13. The key to the proof turns out to be a technical result—Theorem 2.14—on the relation between the adjoint vector of the maximum principle and the gradient of the cost function. This in turn depends on

[*] Received by the editors February 24, 1999; accepted for publication (in revised form) November 29, 1999; published electronically August 31, 2000.
http://www.siam.org/journals/sicon/39-2/32203.html

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

[‡] Department of Mathematics, Rutgers University, 110 Frelinghuysen Rd., Piscataway, NJ 08854 (sussmann@hilbert.rutgers.edu, http://www.math.rutgers.edu/~sussmann). This author was supported in part by NSF grant DMS95-00798 and AFOSR grant 0923.

results about differentiability of trajectories with respect to a parameter under weak differentiability assumptions on the dependence of the data on the parameter. These results are of interest in themselves and are presented separately in Appendix A.

Our definitions of regular presynthesis and regular synthesis are explained in section 2; cf. Definition 2.3 for the notion of "presynthesis," Definition 2.4 for that of "synthesis," and Definition 2.12 for "regularity." The strength of Theorem 2.13 is illustrated in section 3 by showing that it applies to the synthesis for Fuller's problem, where other sufficiency theorems such as those of [2] and [6] cannot be used because of the technical problem arising from the occurrence of an infinite number of switchings. (Not surprisingly, the optimality of Fuller's synthesis was originally proved without resorting to regular synthesis arguments, and using instead special symmetry properties of the problem, cf., e.g., Fuller [13], Marchal [15], Zelikin–Borisov [32].) Other definitions of synthesis—including those of Boltyanskii and Brunovský—are reviewed and compared with ours in section 4, where we also discuss in detail why the concept of "optimal synthesis" cannot just be defined by simply identifying it with that of "optimal feedback law." In particular, we explain in section 4 that if $x \to v(x)$ is an optimal feedback, then in general the concept of "solution" of the closed-loop equation $\dot{x} = f(x, v(x))$ is problematic (because $x \to f(x, v(x))$ may fail to be Lipschitz-continuous), and we compare the classical definition of solution with two other notions: (i) Filippov solutions and (ii) the limits of sample-and-hold solutions recently proposed as feedback solutions by Clarke et al. in [8], which will be referred to here as "CLSS solutions." We argue that the classical notion of solution fares better than the other two concepts, since there are situations where the optimal trajectories are not Filippov solutions of the optimal closed-loop equation, as well as cases where they are not CLSS solutions. The analysis of section 4 is supplemented in section 5 with several examples that clarify the relationship between the various definitions. Finally, a number of technical points, including some arguments pertaining to the proof of Theorem 2.13, are relegated to three appendices.

As explained in section 4, there are essentially two opposite starting points for defining the notion of a "regular" optimal synthesis. One approach, following Boltyanskii [2], Brunovský [5], [6], Cesari [7], Fleming and Rishel [11], and Sussmann [18], is to take as the primary object a smooth or "piecewise smooth" feedback control v, and then let the synthesis be the collection $Traj(v)$ of all trajectories generated by v. If this collection is too large (for example, if the closed-loop equation determined by v does not have unique solutions), then a synthesis would be a suitably defined subclass Γ of $Traj(v)$. The other strategy, proposed in Piccoli [16] and Sussmann [31], is to take a "synthesis" to be just a collection of trajectories—or, more precisely, trajectory-control pairs—not necessarily arising from a feedback control. To carry out the first strategy, Boltyanskii and Brunovský were forced to make precise sense of the notion of piecewise smoothness, which they did by requiring that the set of interest—that is, the set of initial points of optimal trajectories—be partitioned into submanifolds of various dimensions on which a smooth feedback control is specified. It turns out, however, that with this definition the feedback control by itself is not sufficient for the synthesis to be completely determined, since a discontinuous feedback control may generate too many trajectories, some of which could fail to be optimal, as shown in Example 5.3 below. So, in any case, the final outcome of the effort to implement the first strategy is a result more consistent with the second one, since one ends up with a notion of synthesis as a pair consisting of a feedback control v together with a selection Γ of a family of trajectories generated by v, suggesting that the specification of Γ is needed but that of v might not be. This leads naturally to the approach of this paper, in which v is dispensed with altogether.

An important point that ought to be further investigated but that will not be touched upon here is the relation between the geometrical approaches of [3], [4], [5], [6], [14], [17], [18], [19], [20], [21], [22], [23], [24], [26], [27], [31] and the theory of viscosity solutions (cf. Bardi and Capuzzo-Dolcetta [1], Fleming and Soner [12]). We point out that the existing results of this theory for deterministic optimal control do not appear to cover cases such as Fuller's problem. (Cf. [1], for example, where the main result is for Lagrangians that are bounded away from zero.)

2. Regular presynthesis and synthesis, and the sufficiency theorem.

We start with some definitions that will enable us to state and prove a mild generalization of the theorem whose proof was outlined in [31]. Our assumptions will be minimal, so as to achieve maximum generality.

We consider a control system,

$$\dot{x} = f(x,u), \qquad x \in \Omega, \qquad u \in U, \tag{2.1}$$

and a family $\boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{P}}(\Omega, U, f, L, \mathcal{T}, \tau) = \{\mathcal{P}(\Omega, U, f, L, \mathcal{T}, \tau, x_0)\}_{x_0 \in \Omega}$ of minimization problems, where $\mathcal{P}(\Omega, U, f, L, \mathcal{T}, \tau, x_0)$ is the problem

$$\text{minimize } \tau(x(b)) + \int_a^b L(x(t), u(t))dt \quad \text{subject to (2.1)}, \;\; x(a) = x_0, \;\; \text{and} \;\; x(b) \in \mathcal{T}. \tag{2.2}$$

(A more precise formulation of the problems $\mathcal{P}(\Omega, U, f, L, \mathcal{T}, \tau, x)$, carefully specifying the space of curves to be considered, will be given below. Since both f and L have no explicit dependence on t, we can always take $b = 0$.) We assume that

(A1) n is a positive integer and Ω—the "state space"—is an open subset of $\mathbb{R}^n$;

(A2) U—the "control set"—is a nonempty set;

(A3) $f : \Omega \times U \to \mathbb{R}^n$—the "dynamics"—and $L : \Omega \times U \to \mathbb{R}$—the "Lagrangian"—are maps;

(A4) $f(\cdot, u)$ and $L(\cdot, u)$ are maps of class $\mathcal{C}^1$ for every fixed $u \in U$;

(A5) $\mathcal{T}$—the "target"—is a nonempty subset of Ω;

(A6) $\tau : \Omega \to \mathbb{R}$ is a function of class $\mathcal{C}^1$.

We use $T_x^c S$ to denote the *contingent tangent cone* to a set S at a point $x \in S$, defined as follows: a vector v is in $T_x^c S$ if and only if there exist a sequence $\{x_j\}$ of points of S and a sequence $\{h_j\}$ of strictly positive numbers such that $h_j \to 0$ and $x_j = x + h_j v + o(h_j)$ as $j \to \infty$. We write

$$\tilde{T}_x S = T_x^c S \cap (-T_x^c S) = \{v : v \in T_x^c S \quad \text{and} \quad -v \in T_x^c S\}. \tag{2.3}$$

As in Sussmann [31], $\tilde{f}$ will denote the map $\Omega \times U \ni (x,u) \to \tilde{f}(x,u) \overset{\text{def}}{=} (f(x,u), L(x,u)) \in \mathbb{R}^{n+1}$.

If $\Omega \times \mathbb{R} \ni (x,t) \to g(x,t) \in \mathbb{R}^m$ is a time-varying vector-valued map on Ω, then $Dg(x,t)$ denotes, for each $(x,t) \in \Omega$, the Jacobian matrix at x of the map $y \to g(y,t)$, so $Dg(x,t) \in \mathbb{R}^{m \times n}$.

If $\mu : A \to B$ is a map, we use $\mathrm{Dom}(\mu)$ to indicate the domain of μ, i.e., the set A.

A *control* is a U-valued map η such that $\mathrm{Dom}(\eta)$ is a compact subinterval $[a,b]$ of $\mathbb{R}$. If η is a control, we define maps $f_\eta : \Omega \times \mathbb{R} \to \mathbb{R}^n$, $L_\eta : \Omega \times \mathbb{R} \to \mathbb{R}$, $\tilde{f}_\eta : \Omega \times \mathbb{R} \to \mathbb{R}^{n+1}$ by letting

$$\begin{array}{llll} f_\eta(x,t) = f(x,\eta(t)), & L_\eta(x,t) = L(x,\eta(t)), & \tilde{f}_\eta(x,t) = \tilde{f}(x,\eta(t)) & \text{if } t \in \mathrm{Dom}(\eta); \\ f_\eta(x,t) = 0, & L_\eta(x,t) = 0, & \tilde{f}_\eta(x,t) = 0 & \text{if } t \notin \mathrm{Dom}(\eta). \end{array}$$

A *trajectory* for a control η is an absolutely continuous map $\gamma : \mathrm{Dom}(\eta) \to \Omega$ with the property that $\dot\gamma(t) = f(\gamma(t), \eta(t))$ for almost every $t \in \mathrm{Dom}(\eta)$. If $\mathrm{Dom}(\eta) = [a, b]$ (so that $\mathrm{Dom}(\gamma) = [a, b]$ as well), and $x = \gamma(a)$, $y = \gamma(b)$, we say that γ (or the pair (γ, η)) *goes from x to y*, and that η *steers x to y*, and we write $\gamma^- = x$, $\gamma^+ = y$. We use $\mathrm{Range}(\gamma)$ to denote the range of the map γ, that is, the set $\{\gamma(t) : t \in \mathrm{Dom}(\gamma)\}$.

A control η is $\mathcal{C}^1$-*admissible* for $\tilde f$ if $\tilde f_\eta$ satisfies the following $\mathcal{C}^1$ Carathéodory conditions:

(C.i) the map $\mathrm{Dom}(\eta) \ni t \to \tilde f_\eta(x, t) \in \mathbb{R}^{n+1}$ is measurable for each $x \in \Omega$;

(C.ii) for every compact subset K of Ω there exists an integrable function $\varphi_K : \mathbb{R} \to \mathbb{R}$ such that for every $x \in K$ and $t \in \mathrm{Dom}(\eta)$

$$\|\tilde f_\eta(x, t)\| + \|D\tilde f_\eta(x, t)\| \le \varphi_K(t). \tag{2.4}$$

If η is a $\mathcal{C}^1$-admissible control for $\tilde f$, and γ is a trajectory corresponding to η, then we say that (γ, η) is a $\mathcal{C}^1$-*admissible pair* for $\tilde f$. We use $\mathrm{Adm}^{\mathcal{C}^1}(\tilde f)$ to denote the set of all $\mathcal{C}^1$-admissible pairs for $\tilde f$, and write $\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f)$, $\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$, $\mathrm{Adm}_x^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$ to denote, respectively, the set of all $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1}(\tilde f)$ such that $\mathrm{Dom}(\eta)$ is an interval of the form $[T, 0]$ with $T \le 0$, the set of those $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f)$ such that $\gamma(0) \in \mathcal{T}$, and the set of those $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$ such that $\gamma^- = x$. For a pair $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$, we use $J(\gamma, \eta)$ to denote the *cost* of (γ, η), given by

$$J(\gamma, \eta) = \tau(\gamma^+) + \int_{\mathrm{Dom}(\gamma)} L(\gamma(t), \eta(t))\, dt. \tag{2.5}$$

If $\Gamma \subseteq \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$ is an arbitrary set of $\mathcal{C}^1$-admissible pairs for $\tilde f$ ending at the target $\mathcal{T}$ at time 0, we define a function $V_\Gamma : \Omega \to \mathbb{R} \cup \{\pm\infty\}$, called the *cost function of* Γ, by letting

$$V_\Gamma(x) = \inf\{J(\gamma, \eta) : (\gamma, \eta) \in \Gamma,\ \gamma^- = x\} \tag{2.6}$$

for $x \in \Omega$. (In particular, $V_\Gamma(x) = +\infty$ if and only if there is no $(\gamma, \eta) \in \Gamma$ starting at x.) If $\Gamma = \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$, then the cost function V_Γ is the *value function* of our problem, and in that case we write $V_{\tilde f,\mathcal{T}}$ rather than $V_{\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f,\mathcal{T})}$. Naturally, $V_\Gamma \ge V_{\tilde f,\mathcal{T}}$ pointwise for every subset Γ of $\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$.

We can now formulate $\mathcal{P}(\Omega, U, f, L, \mathcal{T}, \tau, x_0)$ precisely: $\mathcal{P}(\Omega, U, f, L, \mathcal{T}, \tau, x_0)$ is the problem of finding a $(\gamma, \eta) \in \mathrm{Adm}_{x_0}^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$ such that $J(\gamma, \eta) = V_{\tilde f,\mathcal{T}}(x_0)$. Any such pair (γ, η) is said to be a *solution* of $\mathcal{P}(\Omega, U, f, L, \mathcal{T}, \tau, x_0)$. A pair $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde f, \mathcal{T})$ is *optimal* if it is a solution of $\mathcal{P}(\Omega, U, f, L, \mathcal{T}, \tau, \gamma^-)$. We use $\boldsymbol{\mathcal{P}}(\Omega, U, f, L, \mathcal{T}, \tau)$—or, simply, $\boldsymbol{\mathcal{P}}$, in any context where the meaning of Ω, U, f, L, $\mathcal{T}$, τ is clear—to denote the family of problems $\{\mathcal{P}(\Omega, U, f, L, \mathcal{T}, \tau, x_0)\}_{x_0 \in \Omega}$.

Remark 2.1. We emphasize that in the optimal control problems studied here *the trajectories have to end at the target but could pass through it before.* For example, consider the dynamical law $\dot x = 1$ in $\mathbb{R}$, with Lagrangian $L(x) \equiv -1$ and target $\mathcal{T} = \{0, 1\}$. Then the value function V is given by $V(x) = x - 1$ for $x \le 1$ and $V(x) = +\infty$ for $x > 1$. If $x < 1$, then the optimal trajectory γ_x from x to the target is given by $\gamma_x(t) = t + 1$ for $T_x \le t \le 0$, where $T_x = x - 1$. Notice that if $x < 0$, then $t = 0$ is not the first time when $\gamma_x(t)$ belongs to $\mathcal{T}$.

Given $\lambda \in \mathbb{R}_n$ (the space of row n-vectors), $\lambda^0 \in \mathbb{R}$, $x \in \Omega$, and $u \in U$, we define

$$H^c(\lambda, \lambda^0, x, u) = \langle \lambda, f(x, u)\rangle + \lambda^0 L(x, u), \tag{2.7}$$

$$H(\lambda, \lambda^0, x) = \inf\{H^c(\lambda, \lambda^0, x, u) : u \in U\}. \tag{2.8}$$

The functions $H^c : \mathbb{R}_n \times \mathbb{R} \times \Omega \times U \to \mathbb{R}$ and $H : \mathbb{R}_n \times \mathbb{R} \times \Omega \to \mathbb{R} \cup \{-\infty\}$ are known, respectively, as the *control Hamiltonian* and the *minimized Hamiltonian* of $\tilde{f}$.

Definition 2.2. *We say that the pair* $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f}, \mathcal{T})$ *is* extremal *if there exist an absolutely continuous map* $\lambda : \mathrm{Dom}(\gamma) \to \mathbb{R}_n$ *and a constant* $\lambda^0 \geq 0$ *that satisfy*

(EX1) $(\lambda^0, \lambda(t)) \neq (0, 0)$ *for some (and hence every) t;*

(EX2) $\dot{\lambda} = -\frac{\partial H^c}{\partial x}(\lambda(t), \lambda^0, \gamma(t), \eta(t))$ *for almost everywhere (a.e.)* $t \in \mathrm{Dom}(\eta)$*;*

(EX3) $H(\lambda(t), \lambda^0, \gamma(t)) = H^c(\lambda(t), \lambda^0, \gamma(t), \eta(t)) = 0$ *for a.e.* $t \in \mathrm{Dom}(\eta)$*;*

(EX4) $\langle \lambda(0), v \rangle = \lambda^0 \langle \nabla \tau(\gamma(0)), v \rangle$ *for all* $v \in \tilde{T}_{\gamma(0)}\mathcal{T}$*.*

Properties (EX1), (EX2), (EX3), (EX4) are, respectively, the *nontriviality condition*, the *adjoint equation*, the *Hamiltonian minimization condition*, and the *transversality condition.* A pair (λ, λ^0) such that (EX2) holds is called an *adjoint vector for* (γ, η). An adjoint vector for which (EX3) holds is a *minimizing adjoint vector for* (γ, η).

The maximum principle says that, under some special conditions on the target set $\mathcal{T}$, *every optimal pair* (γ, η) *is extremal.* (This is true, for example, if $\mathcal{T}$ is a submanifold of Ω of class $\mathcal{C}^1$, or a closed convex subset of Ω, or a set which is locally equivalent to a convex set via a $\mathcal{C}^1$ diffeomorphism. More generally, it is true if the cone $\tilde{T}_{\gamma(0)}\mathcal{T}$ is convex—in which case it is of course a subspace, since $v \in \tilde{T}_{\gamma(0)}\mathcal{T} \Rightarrow -v \in \tilde{T}_{\gamma(0)}\mathcal{T}$—and is an approximating cone to $\mathcal{T}$ at $\gamma(0)$. Recall that a closed convex cone C is an *approximating cone* to a set S at a point $s \in S$ if there exist a neighborhood W of 0, and a continuous map $F : W \cap C \to S$, such that $F(v) = s + v + o(||v||)$ as $v \to 0$ via values in C.)

Definition 2.3. *A presynthesis for* $\boldsymbol{\mathcal{P}}$ *is a subset* Γ *of* $\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f}, \mathcal{T})$ *such that the following holds:*

(PS) *Whenever* $(\gamma_1, \eta_1) \in \Gamma$, $(\gamma_2, \eta_2) \in \Gamma$, *and* $\gamma_1^- = \gamma_2^-$, *it follows that* $(\gamma_1, \eta_1) = (\gamma_2, \eta_2)$.

The set $\mathrm{Dom}(\Gamma) \overset{\mathrm{def}}{=} \{\gamma^- : (\gamma, \eta) \in \Gamma\}$ *is called the* domain *of* Γ. *We say that* Γ *is a* presynthesis *on a set* S *if* Γ *is a presynthesis and* $S = \mathrm{Dom}(\Gamma)$.

Clearly, giving a presynthesis on S amounts to choosing, for each $x \in S$, a $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f}, \mathcal{T})$ such that $\gamma^- = x$.

If Γ is a presynthesis such that $\mathrm{Dom}(\Gamma)$ consists of all points that can be steered to a point of $\mathcal{T}$ by means of a pair belonging to $\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f}, \mathcal{T})$, then we say that Γ is *total.*

Given a presynthesis Γ and a point $x \in \mathrm{Dom}(\Gamma)$, we will always use $(\gamma_x^\Gamma, \eta_x^\Gamma)$—or, simply, (γ_x, η_x) when the meaning of Γ is clear from the context—to denote the unique pair $(\gamma, \eta) \in \Gamma$ such that $\gamma^- = x$, and will write $T_x = \min \mathrm{Dom}(\gamma_x)$, so $T_x \leq 0$ and $\mathrm{Dom}(\gamma_x) = \mathrm{Dom}(\eta_x) = [T_x, 0]$.

Definition 2.4. *A presynthesis on a set* S *is* memoryless *if whenever* $x \in S$ *and* $t \in \mathrm{Dom}(\eta_x)$ *it follows that* $y = \gamma_x(t)$ *belongs to* S *and* η_y *is the restriction of* η_x *to the interval* $[t, 0]$. *A* synthesis *is a memoryless presynthesis.*

Definition 2.5. *If each pair of a presynthesis* Γ *is optimal (resp., extremal), then we say that* Γ *is* optimal *(resp.,* extremal*).*

In particular, a presynthesis Γ for $\boldsymbol{\mathcal{P}}(\Omega, U, f, L, \mathcal{T}, \tau)$ is a total optimal presynthesis if and only if $V_\Gamma \equiv V_{\tilde{f},\mathcal{T}}$.

As explained in the introduction, we wish to propose a definition of "regular synthesis" that will give rise to a sufficiency theorem. The theorem will say that a total extremal regular synthesis that satisfies the appropriate boundary conditions is optimal. We will actually define the more general notion of "regular presynthesis," and prove the sufficiency theorem in this broader setting, so the desired optimality result will be true in particular when the presynthesis is a true synthesis.

The definition will say that a "regular presynthesis" is a presynthesis that satisfies suitable regularity assumptions. So we begin by describing these regularity assumptions in detail.

Definition 2.6. *We call a function* $V : \Omega \to \mathbb{R} \cup \{+\infty\}$ weakly upper semicontinuous (w.u.s.c) *at a point* $x \in \Omega$ *if* $\liminf_{y\to x} \limsup_{z\to y} V(z) \le V(x)$.

Equivalently, V is w.u.s.c. at x if there exist sequences $\{x_j\}$, $\{\varepsilon_j\}$, and $\{\delta_j\}$ such that $x_j \in \Omega$, $\varepsilon_j > 0$, and $\delta_j > 0$ for all j, and $x_j \to x$ and $\delta_j \to 0$ as $j \to \infty$, having the property that $V(y) \le V(x) + \delta_j$ whenever $\|y - x_j\| < \varepsilon_j$.

Definition 2.7. *Suppose we are given a Lipschitz-continuous vector field* X *on* Ω *and a function* $V : \Omega \to \mathbb{R} \cup \{+\infty\}$. *We say that* V *has the* NDJ ("no downward jumps") property along X *if the following holds:*

(NDJ) If $[a,b] \ni t \to \gamma(t) \in \Omega$ is an integral curve of X, then $\liminf_{h\downarrow 0} V(\gamma(t-h)) \le V(\gamma(t))$ for every $t \in]a,b]$.

Definition 2.8. *We say that* $V : \Omega \to \mathbb{R} \cup \{+\infty\}$ *satisfies the* weak continuity conditions *for the control system* (2.1) *if* V *is lower semicontinuous and w.u.s.c. at every* $x \in \Omega$ *and has the NDJ property along the vector field* $x \to f(x,u)$ *for every* $u \in U$.

We now define the "weak differentiability conditions" for Γ. In this definition, we let $\rho_{\tilde{f},\Gamma,\bar{x},T}(v)$ be the integrable $\mathbb{R}^{n+1}$-valued function on $[T,0]$ (see definition below) given by

$$\rho_{\tilde{f},\Gamma,\bar{x},T}(v)(t) = \tilde{f}_{\eta_{\bar{x}+v}}(\gamma_{\bar{x}}(t),t) - \tilde{f}_{\eta_{\bar{x}}}(\gamma_{\bar{x}}(t),t)$$

and use the symbol $\rho_{\tilde{f},\Gamma,\bar{x},T}$ to denote the correspondence $v \to \rho_{\tilde{f},\Gamma,\bar{x},T}(v)$, regarded as a map into the space $\mathrm{Bor}([T,0],\mathbb{R}^{n+1})$ of $\mathbb{R}^{n+1}$-valued Borel measures on $[T,0]$. (Recall that $\mathrm{Bor}([T,0],\mathbb{R}^{n+1})$ is the dual of the space $C^0([T,0],\mathbb{R}^{n+1})$ of continuous $\mathbb{R}^{n+1}$-valued functions on $[T,0]$.)

Definition 2.9. *A presynthesis* $\Gamma = \{(\gamma_x,\eta_x) : x \in S\}$ *for* $\boldsymbol{\mathcal{P}}(\Omega,U,f,L,\mathcal{T},\tau)$ *is said to be* (f,L)-differentiable *at a point* $\bar{x} \in \Omega$ *if there exist* (a) *a neighborhood* N *of* $\bar{x}$ *in* Ω *such that* $N \subseteq \mathrm{Dom}(\Gamma)$, (b) *an interval* $[T,0]$ *such that* $\mathrm{Dom}(\eta_x) \subseteq [T,0]$ *for all* $x \in N$, *and* (c) *an* $\bar{\varepsilon} > 0$, *for which*

(DC1) *there are integrable functions* $\psi_\varepsilon : [T,0] \to \mathbb{R}$, *for* $\varepsilon \in]0,\bar{\varepsilon}]$, *with the property that* $\lim_{\varepsilon\to 0} \int_T^0 \psi_\varepsilon(t)\,dt = 0$, *for which the inequality*

$$\left\|\tilde{f}_{\eta_x}(y,t) - \tilde{f}_{\eta_x}(\gamma_{\bar{x}}(t),t) - D\tilde{f}_{\eta_{\bar{x}}}(\gamma_{\bar{x}}(t),t)\cdot(y-\gamma_{\bar{x}}(t))\right\| \le \psi_\varepsilon(t).(||y-\gamma_{\bar{x}}(t)|| + ||x-\bar{x}||) \tag{2.9}$$

holds for every $y \in \Omega$, $x \in N$, $t \in [T,0]$ *such that* $\|y - \gamma_{\bar{x}}(t)\| \le \varepsilon$ *and* $||x-\bar{x}|| \le \varepsilon$;

(DC2) *the map* $\rho_{\tilde{f},\Gamma,\bar{x},T}$ *is weak*-differentiable at* $v = 0$ *in the sense that, for every continuous function* $\alpha : [T,0] \to \mathbb{R}$, *the map* $\mathbb{R}^n \ni v \to \int_T^0 \alpha(t).\rho_{\tilde{f},\Gamma,\bar{x},T}(v)(t)\,dt \in \mathbb{R}^{n+1}$ *is differentiable at* $v = 0$.

Definition 2.10. *A subset* A *of* Ω *is* thin *if there exist* A_1, A_2 *such that* $A = A_1 \cup A_2$, A_1 *is a finite or countable union of connected* $\mathcal{C}^1$ *submanifolds of positive codimension, and* $\mathcal{H}^{n-1}(A_2) = 0$, *where* $\mathcal{H}^k$ *denotes* k*-dimensional Hausdorff measure.*

If S is a subset of Ω, we use $\mathring{S}$ to denote the interior of S in Ω.

DEFINITION 2.11. *If n, Ω, U, f, L, $\mathcal{T}$, τ are such that assumptions* (A1)–(A6) *hold, Γ is a presynthesis for $\boldsymbol{\mathcal{P}}(\Omega, U, f, L, \mathcal{T}, \tau)$, and $S = \mathrm{Dom}(\Gamma)$, we say that Γ has the* interior approximation property *if the following hold:*

(IAP) *If* $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$ is such that $\gamma(t) \in S$ for all $t \in \mathrm{Dom}(\gamma)$, then there exists a sequence $\{(\gamma_j, \eta_j)\}_{j=1}^{\infty}$ in $\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$ such that

(IAP.1) $\gamma_j(t) \in \mathring{S}$ *for every* $t \in \mathrm{Dom}(\gamma_j)$*;*

(IAP.2) $V_\Gamma(\gamma_j^+) \to V_\Gamma(\gamma^+)$*;*

(IAP.3) $\gamma_j^- \to \gamma^-$*;*

(IAP.4) $\int_{\mathrm{Dom}(\gamma_j)} L(\gamma_j(t), \eta_j(t))\, dt \to \int_{\mathrm{Dom}(\gamma)} L(\gamma(t), \eta(t))\, dt.$

We are now ready to present our definition of "regular presynthesis."

DEFINITION 2.12. *Let n, Ω, U, f, L, $\mathcal{T}$, τ be such that assumptions* (A1)–(A6) *hold. Let Γ be a presynthesis for $\boldsymbol{\mathcal{P}}(\Omega, U, f, L, \mathcal{T}, \tau)$. We say that Γ is* regular *if*

(a) *the associated cost function V_Γ satisfies the weak continuity conditions;*

(b) *there exists a thin subset A of Ω such that Γ is (f, L)-differentiable at all points of* $\mathrm{Dom}(\Gamma)\backslash A$.

With this definition, our main sufficiency theorem (a weaker version of which was already stated in [31], with only an outline of the proof) says the following.

THEOREM 2.13. *Let n, Ω, U, f, L, $\mathcal{T}$, τ be such that assumptions* (A1)–(A6) *hold. Let Γ be an extremal regular presynthesis for $\boldsymbol{\mathcal{P}}(\Omega, U, f, L, \mathcal{T}, \tau)$ and let $S =$* $\mathrm{Dom}(\Gamma)$*. Then we have the following:*

(a) *The "dynamic programming inequality"*

$$V_\Gamma(\gamma^-) \le \int_{\mathrm{Dom}(\gamma)} L(\gamma(t), \eta(t))\, dt + V_\Gamma(\gamma^+) \tag{2.10}$$

holds for every $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$ *such that* $\mathrm{Range}(\gamma) \subseteq \mathring{S}$.

(b) *If V_Γ satisfies the boundary condition*

$$V_\Gamma(x) \le \tau(x) \quad \text{for every} \quad x \in \mathcal{T}, \tag{2.11}$$

then $V_\Gamma(\gamma^-) \le J(\gamma, \eta)$ for all $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f}, \mathcal{T})$ such that $\mathrm{Range}(\gamma) \subseteq \mathring{S}$.

(c) *If Γ has the interior approximation property, then* (2.10) *holds for all* $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$ *such that* $\mathrm{Range}(\gamma) \subseteq S$.

(d) *If Γ has the interior approximation property and satisfies* (2.11)*, then $V_\Gamma(\gamma^-) \le J(\gamma, \eta)$ for all $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f}, \mathcal{T})$ such that* $\mathrm{Range}(\gamma) \subseteq S$.

(e) *If Γ is total and satisfies* (2.11)*, then Γ is optimal.*

Proof. We first establish the relation between the differential of the cost function V_Γ and the adjoint fields of vectors along the extremal trajectories of the presynthesis. This result is important in its own right, so we state it as a separate theorem.

THEOREM 2.14. *Let n, Ω, U, f, L, $\mathcal{T}$, τ be such that assumptions* (A1)–(A6) *hold, and let $\Gamma = \{(\gamma_x, \eta_x) : x \in S\}$ be a presynthesis. Let $\bar{x} \in S$ be such that Γ is (f, L)-differentiable at $\bar{x}$. Then*

(a) *the function V_Γ is differentiable at $\bar{x}$;*

(b) *if the pair $(\gamma_{\bar{x}}, \eta_{\bar{x}})$ is extremal, then every minimizing adjoint vector (λ, λ^0) for $(\gamma_{\bar{x}}, \eta_{\bar{x}})$ satisfies the identity*

$$\lambda^0 (\nabla V_\Gamma)(\bar{x}) = \lambda(T_{\bar{x}}).$$

In particular, $\lambda^0 \neq 0$, and (λ, λ^0) is unique up to multiplication by a positive constant.

Proof of Theorem 2.14. We apply Theorem A.11 in Appendix A, taking P to be $\mathbb{R}^n$ and letting $f_x = f_{\eta_x}$, $L_x = L_{\eta_x}$, $\tilde{f}_x = \tilde{f}_{\eta_x}$. Then, if $y \in \Omega$ and $x \in N$, we have $f_x(y,t) = f(y, \eta_x(t))$, $L_x(y,t) = L(y, \eta_x(t))$, $\tilde{f}_x(y,t) = \tilde{f}(y, \eta_x(t))$ if $t \in [T_x, 0]$, and $f_x(y,t) = 0$, $L_x(y,t) = 0$, $\tilde{f}_x(y,t) = 0$ if $t \notin [T_x, 0]$.

We let $\hat{V}_\Gamma(x) = \int_{T_x}^0 L(\gamma_x(t), \eta_x(t))\,dt$, so $\hat{V}_\Gamma$ is the "Lagrangian part" of our cost function. We extend the curve γ_x to the interval $[T, 0]$ by letting $\gamma_x(t) = x$ for $T \leq t < T_x$, so γ_x still is an integral curve of f_{η_x} on $[T, 0]$, and $\hat{V}_\Gamma(x) = \int_T^0 L_x(\gamma_x(t), t)\,dt$.

The map $p \to \bar{x}_p$ of Theorem A.11 is the identity map, which is certainly differentiable at $\bar{x}$. So Theorem A.11 tells us that the "endpoint map" $x \to \mathcal{E}(x) \overset{\mathrm{def}}{=} \gamma_x(0)$ and the function $\hat{V}_\Gamma$ are differentiable at $\bar{x}$. This implies that the function V_Γ is differentiable at $\bar{x}$, proving Theorem 2.14(a).

Assume now that $(\gamma_{\bar{x}}, \eta_{\bar{x}})$ is extremal and (λ, λ^0) is a minimizing adjoint vector for $(\gamma_{\bar{x}}, \eta_{\bar{x}})$. We extend the adjoint vector λ to $[T, 0]$ by letting $\lambda(t) = \lambda(T_{\bar{x}})$ for $T \leq t < T_{\bar{x}}$. We then define

$$H_x^c(p, p^0, y, t) = \langle p, f_x(y,t) \rangle + p^0 L_x(y,t).$$

Then the adjoint equation $\dot{\lambda}(t) = -\nabla_y H_{\bar{x}}^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t)$ holds a.e. on $[T, 0]$. Moreover,

$$H_x^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t) \geq H_{\bar{x}}^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t) = 0 \text{ for a.e. } t \in [T, 0]. \tag{2.12}$$

(To see this, we first observe that $H_{\bar{x}}^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t) = 0$ a.e. on $[T, 0]$, because (a) the equality holds a.e. when $t \in [T_x, 0]$ thanks to (EX3), and (b) it holds everywhere when $t \in [T, T_x[$, because in this case $f_{\bar{x}}(\gamma_{\bar{x}}(t), t) = 0$ and $L_{\bar{x}}(\gamma_{\bar{x}}(t), t) = 0$. Next, we show that, if $x \in N$, then $H_x^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t) \geq 0$. The inequality follows from (EX3) for almost every t when $t \geq T_{\bar{x}}$ and $t \geq T_x$, because in this case $H_x^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t) = H^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), \eta_x(t)) \geq H(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t)) = 0$ for almost every t. When $t < T_x$, $H_x^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t)$ vanishes, because $f_x(\gamma_{\bar{x}}(t), t) = 0$ and $L_x(\gamma_{\bar{x}}(t), t) = 0$. Finally, we must consider the possibility that $T_x \leq t < T_{\bar{x}}$. For this purpose, we observe that, if $u \in U$ is arbitrary, then $H^c(\lambda(s), \lambda^0, \gamma_{\bar{x}}(s), u) \geq H(\lambda(s), \lambda^0, \gamma_{\bar{x}}(s)) \geq 0$ for almost every $s \in [T_{\bar{x}}, 0]$. Since $s \to H^c(\lambda(s), \lambda^0, \gamma_{\bar{x}}(s), u)$ is continuous, we conclude that $H^c(\lambda(s), \lambda^0, \gamma_{\bar{x}}(s), u) \geq 0$ for *all* $s \in [T_{\bar{x}}, 0]$. In particular, $H^c(\lambda(T_{\bar{x}}), \lambda^0, \bar{x}, u) \geq 0$. Since λ is constant on $[T, T_{\bar{x}}]$, we see that $H^c(\lambda(s), \lambda^0, \bar{x}, u) \geq 0$ for all $s \in [T, T_{\bar{x}}]$ and all $u \in U$. Now, if $T_x \leq t < T_{\bar{x}}$, then $H_x^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t) = H^c(\lambda(t), \lambda^0, \bar{x}, \eta_x(t)) \geq 0$, and (2.12) is proved.)

From Theorem A.11, the identity

$$\begin{aligned}
&\lambda(0).D\mathcal{E}(\bar{x}).v + \lambda^0 \nabla \hat{V}_\Gamma(\bar{x}).v - \lambda(T_{\bar{x}}) \cdot v \\
&\qquad = \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} \int_T^0 \Big(H_{\bar{x}+\varepsilon v}^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t) - H_{\bar{x}}^c(\lambda(t), \lambda^0, \gamma_{\bar{x}}(t), t) \Big)\,dt
\end{aligned} \tag{2.13}$$

follows, for every vector v. Then (2.12) tells us that the right-hand side of inequality (2.13) is nonnegative. So

$$\lambda(0).D\mathcal{E}(\bar{x}).v + \lambda^0 \nabla \hat{V}_\Gamma(\bar{x}).v - \lambda(T_{\bar{x}}) \cdot v \geq 0 \qquad \text{for all } v \in \mathbb{R}^n. \tag{2.14}$$

Given v, the curve $r \to \gamma_{\bar{x}+rv}(0)$ is defined for r in a neighborhood of 0 in $\mathbb{R}$, and is differentiable at 0 with derivative $D\mathcal{E}(\bar{x}).v$. On the other hand, this curve is contained in $\mathcal{T}$, so the vector $w = D\mathcal{E}(\bar{x}).v$ belongs to $\tilde{T}_{\gamma_{\bar{x}}(0)}\mathcal{T}$. It then follows from the transversality condition (EX4) that $\langle\lambda(0), w\rangle = \lambda^0\langle\nabla\tau(\gamma(0)), w\rangle$. Therefore

$$\lambda^0(\nabla\tau(\gamma(0)).D\mathcal{E}(\bar{x}).v + \nabla\hat{V}_\Gamma(\bar{x}).v) - \lambda(T_{\bar{x}})\cdot v \geq 0 \qquad \text{for all } v \in \mathbb{R}^n. \tag{2.15}$$

Since the left-hand side of (2.15) is linear in v, it has to vanish for all v. So

$$\lambda^0\nabla V_\Gamma(\bar{x}).v = \lambda^0(\nabla\tau(\gamma(0)).D\mathcal{E}(\bar{x}).v + \nabla\hat{V}_\Gamma(\bar{x}).v) = \lambda(T_{\bar{x}})\cdot v \qquad \text{for all } v \in \mathbb{R}^n. \tag{2.16}$$

Therefore $\lambda^0\nabla V_\Gamma(\bar{x}) = \lambda(T_{\bar{x}})$.

In particular, λ^0 cannot vanish, because $\lambda^0 = 0$ would imply $\lambda(T_{\bar{x}}) = 0$, contradicting (EX1). So we may, after multiplication by a positive constant, assume that $\lambda^0 = 1$, and then $\nabla V_\Gamma(\bar{x}) = \lambda(T_{\bar{x}})$. This proves, in particular, the uniqueness of $(\lambda(\cdot), \lambda^0)$ up to multiplication by a positive constant, since the Cauchy problem for the adjoint equation has unique solutions. □

Continuation of the proof of Theorem 2.13. We first prove Theorem 2.13(a). For any given $x \in S \setminus A$, let λ_x be the unique λ such that $(\lambda, 1)$ is a minimizing adjoint vector along (γ_x, η_x). (The existence and uniqueness of λ_x follows from Theorem 2.14.) The minimization condition (EX4) of maximum principle then implies that, for any given $u \in U$, the inequality

$$\langle\lambda_x(t), f(\gamma_x(t), u)\rangle + L(\gamma_x(t), u) \geq 0$$

holds for almost all $t \in [T_x, 0]$. Since $t \to \langle\lambda_x(t), f(\gamma_x(t), u)\rangle + L(\gamma_x(t), u)$ is continuous on $[T_x, 0]$, the inequality is valid at time T_x. In view of Theorem 2.14,

$$\langle\nabla V_\Gamma(x), f(x, u)\rangle + L(x, u) \geq 0. \tag{2.17}$$

Now consider an arbitrary $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$, with domain $[T, 0]$, such that $\mathrm{Range}(\gamma) \subseteq \mathring{S}$. We want to prove that (2.10) holds. It follows from Lemma 4.1 of [28] that there exists a sequence of admissible pairs $(\gamma_j, \eta_j) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$, with domain $[T, 0]$, such that $\gamma_j(0) = \gamma(0)$, η_j is piecewise constant, $\gamma_j \to \gamma$ uniformly on $[T, 0]$, and $\int_T^0 L(\gamma_j(t), \eta_j(t))\,dt \to \int_T^0 L(\gamma(t), \eta(t))\,dt$.

Since $\mathrm{Range}(\gamma) \subseteq \mathring{S}$, we can clearly assume that $\mathrm{Range}(\gamma_j) \subseteq \mathring{S}$. Since V_Γ is lower semicontinuous, it follows that, if $V_\Gamma(\gamma_j(T)) + \int_T^0 L(\gamma_j(t), \eta_j(t))\,dt \leq V_\Gamma(\gamma(0))$ for each j, then $V_\Gamma(\gamma(T)) + \int_T^0 L(\gamma(t), \eta(t))\,dt \leq V_\Gamma(\gamma(0))$. Hence it suffices to assume that the control η is piecewise constant.

If (2.10) holds for two pairs $(\gamma_i, \eta_i) \in \mathrm{Adm}^{\mathcal{C}^1}(\tilde{f})$ such that $\mathrm{Dom}(\gamma_1) = [a, b]$, $\mathrm{Dom}(\gamma_2) = [b, c]$, and $\gamma_1(b) = \gamma_2(b)$ for some a, b, c, then it also holds for the concatenated pair (γ, η) defined in an obvious way. It therefore suffices to prove (2.10) for all $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1}(\tilde{f})$ corresponding to a constant control. Since L does not depend on time we can assume $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$ and η constant, so we make this assumption from now on.

Let $u \in U$ be such that $\eta(t) = u$ for $t \in [T, 0]$. For each $y \in \Omega$, let $t \to \zeta^y(t)$ be the trajectory of η that satisfies the terminal condition $\zeta^y(0) = y$. Let W be an open subset of Ω such that $\gamma(0) \in W$, $\zeta^y(T)$ is defined for all $y \in W$, and $\zeta^y([T, 0])$ is contained in $\mathring{S}$ whenever $y \in W$. We will show that

(#) for almost every $y \in W$ the set $B^y = \{t \in [T,0] : \zeta^y(t) \in A\}$ is finite or countable.

Since A is thin, we can write $A = A_1 \cup A_2$, where $A_1 = \bigcup_j M_j$ and $\{M_j\}_{j\in J}$ is a finite or countable family of connected submanifolds of Ω of codimension $d_j > 0$, while $\mathcal{H}^{n-1}(A_2) = 0$. After replacing each M_j by a finite or countable family of open submanifolds of M_j, we may assume that the M_j are embedded.

Define $\tilde{W} = W \times]\,T, 0\,[$, and let Φ be the map $\tilde{W} \ni (y,t) \to \zeta^y(t) \in \Omega$. Then Φ is a $\mathcal{C}^1$ submersion, because for each fixed t the partial map $y \to \Phi(y,t)$ is a $\mathcal{C}^1$ diffeomorphism. Therefore each set $\tilde{M}_j = \Phi^{-1}(M_j)$ is an embedded submanifold of $\tilde{W}$ of codimension d_j, and $\mathcal{H}^n(\Phi^{-1}(A_2)) = 0$.

Let π be the projection $\tilde{W} \ni (y,t) \to y \in W$. Let S_j be the set of points $s \in \tilde{M}_j$ such that the restriction $\pi\lceil\tilde{M}_j$ of π to $\tilde{M}_j$ is not regular at s. (Recall that, if S_1, S_2 are $\mathcal{C}^1$ manifolds, then a $\mathcal{C}^1$ map $F : S_1 \to S_2$ is *regular* at $s \in S_1$ if the differential $d\pi(s)$ maps the tangent space T_sS_1 surjectively onto $T_{F(s)}S_2$.) Then $\mathcal{H}^n(\pi(\Phi^{-1}(A_2))) = 0$ (see, for example, Federer [9]), and $\mathcal{H}^n(\pi(S_j)) = 0$ for each j, by Sard's theorem. So the "bad" set $\mathcal{B} = \pi(\Phi^{-1}(A_2)) \cup (\bigcup_j \pi(S_j))$ has Lebesgue measure zero.

Now let $y \in W \backslash \mathcal{B}$. Then $\zeta^y(t) \notin A_2$ if $T < t < 0$. (Otherwise, if $\zeta^y(t) \in A_2$, $T < t < 0$, it would follow that $(y,t) \in \Phi^{-1}(A_2)$, so $y \in \pi(\Phi^{-1}(A_2))$, and then $y \in \mathcal{B}$.) Also, we claim that the set $\{t \in [T,0] : \zeta^y(t) \in A_1\}$ is at most countable. To see this, it suffices to show that, for each j, the set $E_j = \{t : T < t < 0\,,\, \zeta^y(t) \in M_j\}$ is finite or countable. Fix j, and suppose $t \in E_j$. The manifold $\tilde{M}_j$ has codimension d_j in $\tilde{W}$. Since $d_j > 0$, the dimension ν_j of $\tilde{M}_j$ is $\leq n$. Moreover, $d\pi(y,t)$ maps $T_{y,t}\tilde{M}_j$ surjectively onto $\mathbb{R}^n$ (the tangent space to Ω at y), because $y \notin \mathcal{B}$. So $\nu_j = n$, and $d\pi(y,t)$ is injective on $T_{y,t}\tilde{M}_j$. Since $d\pi(y,t)(\frac{\partial}{\partial t})$ is obviously equal to 0, it follows that $\frac{\partial}{\partial t} \notin T_{y,t}\tilde{M}_j$. Since $\tilde{M}_j$ is embedded, $(y,t') \notin \tilde{M}_j$ if $0 < |t'-t| \leq \varepsilon$, provided that ε is sufficiently small. Therefore t is an isolated point of E_j. So E_j is a discrete subset of the open interval $]\,T, 0\,[$. Therefore E_j is finite or countable, as desired.

We have therefore proved that, if $y \in W \backslash \mathcal{B}$, then $\zeta^y(t) \notin A_2$ for all $t \in]T,0[$, and the set $\{t \in [T,0] : \zeta^y(t) \in A_1\}$ is at most countable. Since $A = A_1 \cup A_2$, and $\mathcal{B}$ has measure zero, (#) follows.

In view of the weak upper semicontinuity of V_Γ, there exist x_j, $\varepsilon_j > 0$, $\delta_j > 0$ such that $x_j \to \gamma(0)$ and $\delta_j \to 0$ as $j \to \infty$, and $V_\Gamma(y) \leq V_\Gamma(\gamma(0)) + \delta_j$ whenever $\|y - x_j\| \leq \varepsilon_j$. Choose the ε_j's so that $\|y - x_j\| \leq \varepsilon_j$ implies $y \in W$. Then, by (#), we can choose for each j a y_j such that $\|y_j - x_j\| \leq \varepsilon_j$ and the bad set B^{y_j} is at most finite or countable. Since $\zeta^{y_j}(t) \in \mathring{S}$ for every $t \in [T,0]$, (2.17) implies that the function $[T,0] \ni t \to \varphi_j(t) = V_\Gamma(\zeta^{y_j}(t)) + \int_T^t L(\zeta^{y_j}(s), \eta(s))ds \in \mathbb{R}$ is differentiable at every t, except for a finite or countable number of values, with a nonnegative derivative. Therefore φ_j satisfies assumption (a) of Lemma B.1 (Appendix B). Moreover, the lower semicontinuity of V_Γ ensures assumption (b) of Lemma B.1 for φ_j, and the NDJ condition ensures assumption (c) of Lemma B.1. Hence we can apply Lemma B.1 to φ_j, and conclude that $\varphi_j(T) \leq \varphi_j(0)$. So

$$V_\Gamma(\zeta^{y_j}(T)) \leq \int_T^0 L(\zeta^{y_j}(t), \eta(t))dt + V_\Gamma(y_j) \leq \int_T^0 L(\zeta^{y_j}(t), \eta(t))dt + V_\Gamma(\gamma(0)) + \delta_j.$$

If we let $j \to \infty$, we find, using the lower semicontinuity of V_Γ, that (2.10) holds for (γ, η), as desired. So (2.10) holds when η is constant and, as explained earlier, this completes the proof of (a).

To prove (b), we pick $(\gamma, \eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f}, \mathcal{T})$ such that γ is contained in $\mathring{S}$, and apply (a) to conclude that $V_\Gamma(\gamma^-) \leq \int_{\mathrm{Dom}(\gamma)} L(\gamma(t),\eta(t))dt + V_\Gamma(\gamma(0))$. Then (2.11) implies the inequality $V_\Gamma(\gamma(0)) \leq \tau(\gamma(0))$, so $V_\Gamma(\gamma^-) \leq J(\gamma,\eta)$ and (b) is proved.

To prove (c), we pick $(\gamma,\eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$ such that γ is contained in S, and find pairs (γ_j,η_j) such that conditions (IAP.1)–(IAP.4) hold. Then (a) implies that the inequality $V_\Gamma(\gamma_j^-) \leq \int_{\mathrm{Dom}(\gamma_j)} L(\gamma_j(t),\eta_j(t))dt + V_\Gamma(\gamma_j(0))$ holds. Using (IAP.2), (IAP.3), (IAP.4), and the lower semicontinuity of V_Γ, we conclude that (2.10) holds, completing the proof of (c).

The proof of (d) is identical to that of (b), using (c) instead of (a).

Finally, to prove (e), we assume that Γ is total and prove that it satisfies the interior approximation property. So now S is the set of all points of Ω that can be steered to a point of $\mathcal{T}$ by a pair belonging to $\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$. Let $(\gamma,\eta) \in \mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f})$ be a trajectory-control pair such that γ is entirely contained in S. Then $\gamma(0) \in S$, so $\gamma(0) \in \mathrm{Dom}(\Gamma)$ and then $V_\Gamma(\gamma(0)) < +\infty$. The weak upper semicontinuity of V_Γ implies that there exist x_j, $\varepsilon_j > 0$, $\delta_j > 0$ such that $x_j \to \gamma(0)$ and $\delta_j \to 0$ as $j \to \infty$, and $V_\Gamma(y) \leq V_\Gamma(\gamma(0)) + \delta_j$ whenever $\|y - x_j\| \leq \varepsilon_j$. It then follows that $V_\Gamma(x_j) \to V_\Gamma(\gamma(0))$, since V_Γ is lower semicontinuous. Let $\mathrm{Dom}(\gamma) = [T,0]$. Let γ_j be the maximally defined trajectory for the control η that satisfies $\gamma_j(0) = x_j$. Then γ_j is defined on $[T,0]$ if j is large enough, and conditions (IAP.2), (IAP.3), and (IAP.4) hold with $\eta_j = \eta$. We may then assume that γ_j is defined on $[T,0]$ for all j. To prove that (IAP.1) holds as well, we observe that $y \in S$ whenever $\|y - x_j\| \leq \varepsilon_j$, because in that case $V_\Gamma(y) < +\infty$, since $V_\Gamma(y) \leq V_\Gamma(\gamma(0)) + \delta_j$, and $V_\Gamma(\gamma(0)) < +\infty$ because $\gamma(0) \in S$. Let ζ^y be the trajectory for the control η that satisfies $\zeta^y(0) = y$. Then for each j there exists a $\tilde{\varepsilon}_j$ such that $0 < \tilde{\varepsilon}_j \leq \varepsilon_j$, with the property that the curve ζ^y is defined on $[T,0]$ for all y such that $\|y - x_j\| \leq \tilde{\varepsilon}_j$. Moreover, it is clear that y is reachable from $\zeta^y(t)$ for every $t \in [T,0]$, so $\zeta^y([T,0]) \subseteq S$ if $\|y - x_j\| \leq \tilde{\varepsilon}_j$, because $y \in S$. Now, if $T \leq t \leq 0$, the map $y \to \zeta^y(t)$ is a $\mathcal{C}^1$ diffeomorphism from the neighborhood $\{x : \|x - x_j\| < \tilde{\varepsilon}_j\}$ of x_j onto a neighborhood $N_j(t)$ of $\gamma_j(t)$. Clearly, $N_j(t) \subseteq S$, so $\gamma_j(t) \in \mathring{S}$. This completes the proof of (e). □

Remark 2.15. The function V_Γ is w.u.s.c. everywhere if and only if it is w.u.s.c. at all the points of the target $\mathcal{T}$. Indeed, let $x \in \Omega$. If $x \notin S$, then $V_\Gamma(x) = +\infty$, and the conclusion follows. Next, suppose $x \in S$, and let (γ_x,η_x) be its associated admissible pair. Assume that V_Γ is w.u.s.c. at $\gamma_x(0)$. Let y_j, ε_j, δ_j be such that $y_j \to \gamma_x(0)$, $\varepsilon_j > 0$, $\delta_j > 0$, $\delta_j \to 0$, and $V_\Gamma(y) \leq V_\Gamma(\gamma_x(0)) + \delta_j$ whenever $\|y - y_j\| \leq \varepsilon_j$. Let ζ^{y_j} be the trajectory corresponding to the control η_x and such that $\zeta^{y_j}(0) = y_j$. Since $y_j \to \gamma_x(0)$, we may assume, by picking j large enough, that ζ^{y_j} is defined on $[T_x,0]$ for all j, and let $x_j = \zeta^{y_j}(T_x)$. Let $\tilde{\varepsilon}_j$ be such that $0 < \tilde{\varepsilon}_j < \frac{1}{j}$ and, if $\|z - x_j\| \leq \tilde{\varepsilon}_j$, then the trajectory θ^z of η_x with initial condition $\theta^z(T_x) = z$ is defined on $[T_x,0]$ and satisfies $\|\theta^z(0) - y_j\| < \varepsilon_j$. Then, using the same reasoning as in the proof of Theorem 2.13, we can show that

$$V_\Gamma(z) \leq \int_{T_x}^0 L(\theta^z(t),\eta_x(t))\,dt + V_\Gamma(\gamma_x(0)) + \delta_j \tag{2.18}$$

whenever $\|z - x_j\| \leq \tilde{\varepsilon}_j$. (First approximate η_x by piecewise constant controls, η^k, and let $\theta^{z,k}$ be the corresponding trajectories with terminal condition $\theta^{z,k}(0) = \theta^z(0)$. Then for each k use statement (#) of the proof of Theorem 2.13 to approximate the terminal point $\theta^z(0)$ by points $w^{z,k,i}$ with the property that, if $\tilde{\theta}^{z,k,i}$ is the trajectory for η^k for which $\tilde{\theta}^{z,k,i}(0) = w^{z,k,i}$, it follows that $V_\Gamma(\tilde{\theta}^{z,k,i}(T_x)) \leq \int_{T_x}^0 L(\tilde{\theta}^{z,k,i}(t),\eta^k(t))\,dt + V_\Gamma(w^{z,k,i})$. Since $\|\theta^z(0) - y_j\| < \varepsilon_j$, the $w^{z,k,i}$ satisfy

$\|w^{z,k,i} - y_j\| < \varepsilon_j$ if i is large enough, so $V_\Gamma(\tilde{\theta}^{z,k,i}(T_x)) \leq \int_{T_x}^0 L(\tilde{\theta}^{z,k,i}(t), \eta^k(t))\, dt + V_\Gamma(\gamma_x(0)) + \delta_j$ when i is large enough. If we let $i \to \infty$ and then let $k \to \infty$, and use the lower semicontinuity of V_Γ, we find that (2.18) holds.)

Let $\tilde{\delta}_j = \sup\{\| \int_{T_x}^0 L(\theta^z(t), \eta_x(t))\, dt - \int_{T_x}^0 L(\gamma_x(t), \eta_x(t))\, dt\| : \|z - x_j\| \leq \tilde{\varepsilon}_j\}$. Then $\tilde{\delta}_j \to 0$, because $x_j \to x$ and $\tilde{\varepsilon}_j \to 0$. On the other hand, $V_\Gamma(z) \leq V_\Gamma(x) + \delta_j + \tilde{\delta}_j$ whenever $\|z - x_j\| \leq \tilde{\varepsilon}_j$. So V_Γ is w.u.s.c. at x.

3. The Fuller phenomenon. In this section we show that the famous minimization problem considered by Fuller in [13]—and extensively studied by other authors, e.g., Zelikin–Borisov [32]—has a regular optimal synthesis in the sense of our Definitions 2.4 and 2.12.

Consider the system

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = u, \qquad |u| \leq 1 \tag{3.1}$$

and the family $\hat{\boldsymbol{\mathcal{P}}} = \{\hat{\mathcal{P}}_x\}_{x \in \mathbb{R}^2}$ of minimization problems:

$$\hat{\mathcal{P}}_x : \qquad \text{minimize} \int_0^{+\infty} x_1^2(t)\, dt \text{ subject to } \quad (x_1(0), x_2(0)) = x. \tag{3.2}$$

We proceed to describe the optimal synthesis for this problem, determined by Fuller, referring the reader to Zelikin–Borisov [32, Chap. 2] for further details. First define the "switching locus"

$$\zeta = \{(x_1, x_2) : |x_1| = Cx_2^2,\ x_1 x_2 \leq 0\},$$

where C is a constant whose precise definition will be given later. Write $\zeta = \zeta^+ \cup \zeta^- \cup \{(0,0)\}$, where $\zeta^\pm = \{(x_1, x_2) \in \zeta : \pm x_1 > 0\}$. Then the piecewise smooth curve ζ divides the plane into two regions A^- and A^+ consisting, respectively, of the points that lie above and below ζ. We define a discontinuous feedback control $k : \mathbb{R}^2 \to [-1, 1]$ by letting $k(x_1, x_2) = -1$ if $(x_1, x_2) \in A^- \cup \zeta^-$, $k(x_1, x_2) = 1$ if $(x_1, x_2) \in A^+ \cup \zeta^+$, and $k(0,0) = 0$.

Then for every $x \in \mathbb{R}^2$ there exists a unique solution $\hat{\gamma}_x : [0, \infty[\, \to \mathbb{R}^2$ of the Cauchy problem $\dot{x}_1 = x_2$, $\dot{x}_2 = k(x_1, x_2)$, $(x_1(0), x_2(0)) = x$. The corresponding open-loop control $\hat{\eta}_x$ is related to $\hat{\gamma}_x$ by $\hat{\eta}_x(t) = +1$ when $\hat{\gamma}_x(t) \in A^+$, $\hat{\eta}_x(t) = -1$ when $\hat{\gamma}_x(t) \in A^-$, and $\hat{\eta}_x(t) = 0$ when $\hat{\gamma}_x(t) = 0$.

The switchings occur when $\hat{\gamma}(t) \in \zeta$. The curve $\hat{\gamma}_x$ reaches the origin in a finite time $\hat{T}_x$ and then stays there. Therefore $\hat{\eta}_x(t) = 0$ when $\hat{\gamma}_x(t) = 0$, which happens if and only if $t \geq \hat{T}_x$ (see Figure 1). Therefore the admissible pairs $(\hat{\gamma}_x, \hat{\eta}_x)$ actually solve the family $\boldsymbol{\mathcal{P}} = \{\mathcal{P}_x\}_{x \in \mathbb{R}^2}$ of minimization problems

$$\mathcal{P}_x : \quad \text{minimize} \int_a^b x_1^2(t) dt \text{ subject to } (x_1(a), x_2(a)) = x \text{ and } (x_1(b), x_2(b)) = (0,0), \tag{3.3}$$

i.e., the problems of reaching the origin in finite time, subject to the dynamical constraints (3.1), and minimizing the cost given by (3.3).

So far, C was an arbitrary positive constant. It turns out that, if C is chosen to be equal to C^*, where C^* is the unique positive root of the polynomial $x^4 + x^2/12 - 1/18$, then the pairs $(\hat{\gamma}_x, \hat{\eta}_x)$ are extremal. (This is proved in [32]. On page 26, it is shown that C must be chosen so that $(1 - 2C)^{1/2}(1 + 2C)^{-1/2} = \mu$, where μ is such that, if $\sigma = \mu + \mu^{-1}$, then $\sigma^2 - 3\sigma - 6 = 0$. These equations easily imply $C^4 + C^2/12 - 1/18 = 0$.)

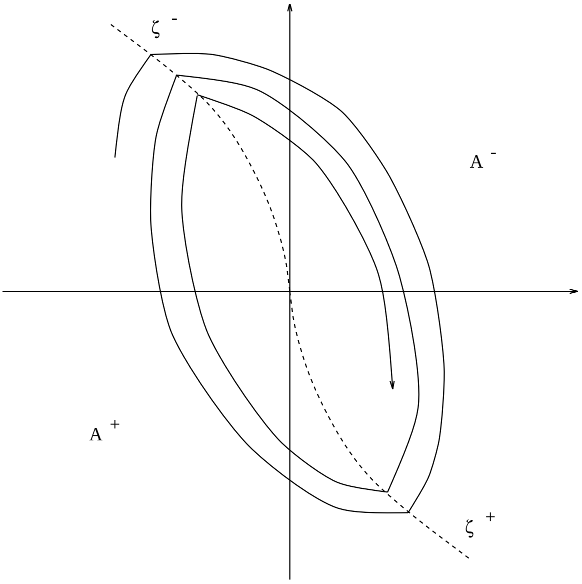


FIG. 1.

We now choose $C = C^*$, so, in particular, $0.4 < C < 0.5$. Then $\{\hat{\gamma}_x\}_{x\in\mathbb{R}^2}$ is a family of extremals reaching the target, and one would like to prove that the $\hat{\gamma}_x$ are optimal. The sufficiency theorems of [2] and [6] do not apply, since the controls $\hat{\eta}_x$ have an infinite number of switchings. For this reason, the optimality of the synthesis obtained by Fuller is usually established by other means. For example, one can prove optimality using the *dilation symmetry properties* of the problem. Precisely, there exists a one parameter family $\Delta = \{\Delta_\rho\}_{\rho>0}$ of maps from $\mathbb{R}^2$ to $\mathbb{R}^2$ such that, if (γ, η) is an admissible pair having cost c, and we let $\gamma_\rho(t) = \Delta_\rho(\gamma(\frac{t}{\rho}))$ and $\eta_\rho(t) = \eta(\frac{t}{\rho})$, then (γ_ρ, η_ρ) is an admissible pair having cost $\rho^5 c$. It then follows that the Δ_ρ map optimal trajectories to optimal trajectories.

The proof of optimality using the dilation symmetry is given, for example, in [32], and will not be repeated here. It will turn out, however, that the dilations Δ_ρ will play a role in our arguments, so we recall their definition and one simple identity that will be useful later:

(1) By definition,

$$\Delta_\rho(x_1, x_2) = (\rho^2 x_1, \rho x_2).$$

(2) It follows from (1), in particular, that

$$\Delta_\rho(x_1, x_2) - x = (\rho - 1)\,(2x_1, x_2) + (\rho - 1)^2\,(x_1, 0)$$

for all ρ, x_1, x_2.

In order to apply Theorem 2.13, we first make a time translation. We define $T_x = -\hat{T}_x$, and then let

$$\gamma_x(t) = \hat{\gamma}_x(t + \hat{T}_x)\,, \qquad \eta_x(t) = \hat{\eta}_x(t + \hat{T}_x)$$

for $T_x \leq t \leq 0$. Then $\Gamma = \{(\gamma_x, \eta_x)\}_{x\in\mathbb{R}^2}$ is a memoryless family of $\mathcal{C}^1$-admissible extremal pairs such that γ_x is defined on $[T_x, 0]$, $\gamma_x(T_x) = x$, and $\gamma_x(0) = 0$.

We want to prove that Γ is an optimal synthesis. For this purpose, we will verify the hypotheses of Theorem 2.13, and in particular the (f, L)-differentiability of Γ at every point $x \in \mathbb{R}^2 \backslash \zeta$. Since it is very easy to check property (DC1), we will verify only (DC2).

As in [32, p. 26], we let

$$\mu = \sqrt{\frac{1-2C}{1+2C}}.$$

Then $0 < \mu < 1$.

Fix a point $\bar{x} = (\bar{x}_1, \bar{x}_2)$ belonging to $\mathbb{R}^2 \backslash \{(0,0)\}$. Then there exist $s_0(\bar{x})$ and $s_1(\bar{x})$ such that $T_{\bar{x}} = s_0(\bar{x}) < s_1(\bar{x}) \leq \mu s_0(\bar{x}) < 0$, and

$$\eta_{\bar{x}} = \pm \sum_{k=0}^{+\infty} (-1)^k \chi_{_{I_k(\bar{x})}},$$

where

(a) $I_0(\bar{x}) = [s_0(\bar{x}), s_1(\bar{x})]$;

(b) $I_k(\bar{x}) = [\mu^{k-1} s_1(\bar{x}), \mu^k s_1(\bar{x})]$ for $k = 1, 2, \ldots$;

(c) $\chi_{_S}$ is the indicator function of a set S;

and

(d) the $\pm$ sign depends on $\bar{x}$ as follows: it is $+$ if $\bar{x} \in A^+ \cup \zeta^+$ and $-$ if $\bar{x} \in A^- \cup \zeta^-$.

Let T be such that $T < 0$ and the interval $[T, 0]$ contains $\mathrm{Dom}(\eta_x)$—i.e., $T \leq T_x$—for x near $\bar{x}$. To prove (DC2), we must pick a point $\bar{x} \in \mathbb{R}^2 \setminus \zeta$ and a continuous function $\alpha : [T, 0] \to \mathbb{R}$, and show that the map

$$\mathbb{R}^2 \ni x \longrightarrow J^\alpha(x) \overset{\text{def}}{=} \int_T^0 \alpha(t).\tilde{f}_{\eta_x}(\gamma_{\bar{x}}(t), t)\, dt \in \mathbb{R}^3 \tag{3.4}$$

is differentiable at $x = \bar{x}$. We will do this by showing that, for two linearly independent directions $v_1(x)$, $v_2(x)$, depending smoothly on x, the directional derivatives $D_{v_1(x)} J^\alpha$ and $D_{v_2(x)} J^\alpha$ exist and are continuous with respect to x. We choose $v_1(x)$ to be the direction of the curve $\rho \to \Delta_\rho(x)$ at $\rho = 1$, and take $v_2(x) = f(x, \eta_x(T_x))$. The determinant of $v_1(x)$ and $v_2(x)$ is then equal to $\pm(2x_1 - x_2^2)$ (the sign being as in (d)), which never vanishes on $A^+ \cup A^-$, since $C < 1/2$. So $v_1(x)$ and $v_2(x)$ are linearly independent for all $x \in A^+ \cup A^-$. We will just determine the directional derivatives at $\bar{x}$, and the result will make it obvious that they depend continuously on $\bar{x}$, so differentiability will follow.

Let us assume, for simplicity, that $\bar{x} \in A^+$. (The case when $\bar{x} \in A^-$ is similar.) We first differentiate J^α in the direction of $v_1(\bar{x})$. If $\rho > 0$, we have (letting $\bar{s}_0 = s_0(\bar{x})$, $\bar{s}_k = \mu^{k-1} s_1(\bar{x})$ for $k = 1, 2, \ldots$, $I_k = [\bar{s}_k, \bar{s}_{k+1}]$, and $I_k^\rho = [\rho \bar{s}_k, \rho \bar{s}_{k+1}]$ for $k = 0, 1, \ldots$, and writing $\bar{x}_\rho = \Delta_\rho(\bar{x})$)

$$T_{_{\bar{x}_\rho}} = \rho T_{\bar{x}}, \qquad \eta_{_{\bar{x}_\rho}} = \sum_{k=0}^{+\infty} (-1)^k \chi_{_{I_k^\rho}},$$

and

$$\tilde{f}_{\eta_{\bar{x}_\rho}}(\gamma_{\bar{x}}(t), t) - \tilde{f}_{\eta_{\bar{x}}}(\gamma_{\bar{x}}(t), t) = \begin{pmatrix} 0 \\ \eta_{_{\bar{x}_\rho}}(t) - \eta_{\bar{x}}(t) \\ 0 \end{pmatrix} + R_\rho(t), \tag{3.5}$$

where $R_\rho(t)$ is a term arising because of the possibility that $T_{\bar x} \neq T_{\bar x_\rho}$.

Then, if $\rho \geq 1$ and $\rho - 1$ is sufficiently small,

$$\begin{aligned}
\int_T^0 (\eta_{\bar x_\rho}(t) - \eta_{\bar x}(t))\alpha(t)\,dt &= \int_{\rho\bar s_0}^0 \sum_{k=0}^{+\infty}(-1)^k\Big(\chi_{I_k^\rho}(t) - \chi_{I_k}(t)\Big)\alpha(t)\,dt \\
&= \sum_{k=0}^{+\infty}(-1)^k\left(\int_{\rho\bar s_k}^{\rho\bar s_{k+1}}\alpha(t)\,dt - \int_{\bar s_k}^{\bar s_{k+1}}\alpha(t)\,dt\right) \\
&= \sum_{k=0}^{+\infty}(-1)^k\left(\int_{\rho\bar s_k}^{\bar s_k}\alpha(t)\,dt - \int_{\rho\bar s_{k+1}}^{\bar s_{k+1}}\alpha(t)\,dt\right) \\
&= \int_{\rho\bar s_0}^{\bar s_0}\alpha(t)\,dt + 2\sum_{k=1}^{+\infty}(-1)^k\left(\int_{\rho\bar s_k}^{\bar s_k}\alpha(t)\,dt\right).
\end{aligned}$$

The numbers $(\rho-1)^{-1}\|\int_{\rho\bar s_k}^{\bar s_k}\alpha(t)\,dt\|$ are bounded by $\mu^{k-1}|\bar s_1|.\|\alpha\|_{L^\infty}$, which is the general term of a convergent series, since $\mu < 1$. So we can divide by $\rho - 1$, let $\rho \longrightarrow 1$, and take the limit of each term separately. We then conclude that

$$\lim_{\rho\downarrow 1}\frac{1}{\rho-1}\int_T^0(\eta_{\bar x_\rho}(t) - \eta_{\bar x}(t))\alpha(t)dt = -\bar s_0\alpha(\bar s_0) - 2\bar s_1\sum_{k=1}^{+\infty}(-1)^k\mu^{k-1}\alpha(\mu^{k-1}\bar s_1). \tag{3.6}$$

We now find the limit of $(\rho-1)^{-1}\int_T^0\alpha(t).R_\rho(t)\,dt$ as $\rho \downarrow 1$. Clearly, $T_{\bar x_\rho} = \rho T_{\bar x} < T_{\bar x}$ if $\rho > 1$. So

$$\int_T^0 \alpha(t).R_\rho(t)\,dt = \int_{\rho T_{\bar x}}^{T_{\bar x}}\alpha(t).\begin{pmatrix}\gamma_{\bar x_\rho}^2(t)\\ 0\\ \left(\gamma_{\bar x_\rho}^1(t)\right)^2\end{pmatrix}dt,$$

where $\gamma_{\bar x_\rho}^i(t)$, for $i = 1, 2$, is the ith component of $\gamma_{\bar x_\rho}(t)$. Therefore

$$\lim_{\rho\downarrow 1}(\rho-1)^{-1}\int_T^0\alpha(t).R_\rho(t)\,dt = -T_{\bar x}\alpha(T_{\bar x}).\begin{pmatrix}\bar x_2\\ 0\\ \bar x_1^2\end{pmatrix}. \tag{3.7}$$

Similar computations show that (3.6) and (3.7) hold as well for the limits as $\rho \uparrow 1$. So the directional derivative of J^α at $x = \bar x$ in the direction of $v_1(\bar x)$ exists and is given by the sum of the right-hand sides of (3.6) and (3.7), which depend continuously on $\bar x$, since $\bar s_0$, $\bar s_1$, and $T_{\bar x}$ (which is none other than $\bar s_0$) depend continuously on $\bar x$ as long as $\bar x \in A^+$.

A much simpler argument verifies the existence of the derivative of J^α in the direction of $v_2(\bar x)$—i.e., of $f(\bar x, 1)$—whose value turns out to be the vector with components $\alpha(T_{\bar x})\bar x_2$, $\alpha(T_{\bar x})$, and $\alpha(T_{\bar x})\bar x_1^2$. Once again, this expression depends continuously on $\bar x$, so the proof of (DC2) is complete.

The other assumptions of Theorem 2.13 are easily checked. So the theorem applies, and the optimality of the synthesis described above follows.

4. Comparison with other definitions of synthesis. The purpose of this section is to compare our concept of regular presynthesis with other definitions of synthesis that either have been proposed by other authors or are sufficiently natural to be worth considering as possible ways to define this notion. For simplicity, we will only consider the case of a single-point target set $\mathcal{T}$, and take this point to be the origin. (Both Boltyanskii and Brunovský, whose definitions will be discussed in detail, do the same, so this will facilitate the comparison.) Moreover, we will consider only nonnegative Lagrangians, so a trajectory containing a loop can never be optimal, and an optimal trajectory necessarily terminates when it hits the origin.

Perhaps the simplest conceivable notion of a "regular synthesis" is that of a "feedback control." In principle, one could define a "feedback control law" on a set S to be a map $v : S \to U$. If v is such that the vector field $S \ni x \to f(x, v(x)) \in \mathbb{R}^n$ is Lipschitz-continuous, and we assume for simplicity that S is open, then v gives rise to unique maximally defined trajectories $\hat{\gamma}_x$ starting at each $x \in S$, and corresponding open-loop controls $\hat{\eta}_x$, given by $\hat{\eta}_x(t) = v(\hat{\gamma}_x(t))$. If we make the additional assumption that every trajectory $\hat{\gamma}_x$ obtained in this way reaches the target at a time $\hat{T}_x$, define $T_x = -\hat{T}_x$, and let $\gamma_x(t) = \hat{\gamma}(t+\hat{T}_x)$ and $\eta_x(t) = \hat{\eta}(t+\hat{T}_x)$ for $T_x \leq t \leq 0$, then we will have constructed a synthesis $\Gamma = \{(\gamma_x, \eta_x)\}_{x \in S}$—in the sense of Definition 2.4—with domain S.

The main drawback of such a definition is that, as is well known, for most reasonable optimal control problems there does not exist an optimal feedback that renders the map $x \to f(x, v(x))$ Lipschitz-continuous or even continuous. So it is absolutely essential to allow "discontinuous feedback laws," and when this is done one immediately runs into the problems of

(a) the lack of a truly satisfactory notion of solution

and

(b) the lack of good theorems guaranteeing existence of solutions.

Difficulty (b) can be handled in at least two ways, namely, by

(A) incorporating into the definition requirements that imply existence and uniqueness of solutions of the closed-loop equation $\dot{x} = f(x, v(x))$;

(B) incorporating into the definition requirements that imply existence—but not necessarily uniqueness—of solutions of the closed-loop equation, and adding to the specification of v a definite prescription for choosing a solution when uniqueness fails, so that the "feedback law" is no longer the closed-loop control v alone, but the *pair* (v, Γ), where $\Gamma = \{(\gamma_x, \eta_x)\}$ is a family that selects one solution of the closed-loop equation for each initial condition x.

Moreover, whether we choose (A) or (B), one has to be precise about the concept of "solution." Here we will consider three such concepts, namely,

(1) classical solutions (i.e., absolutely continuous curves $t \to x(t)$ having the property that the equality $\dot{x}(t) = f(x(t), v(x(t)))$ holds for almost every t),

(2) Filippov solutions (cf. [10]),

(3) "CLSS solutions," that is, the "feedback solutions" defined by Clarke et al. in [8].

The examples of section 5 will show that the concepts of CLSS solution and Filippov solution are not adequate. Indeed, Example 5.4 shows an optimal synthesis whose trajectories are not Filippov solutions of the optimal closed-loop equation. In this example, the optimal trajectories are CLSS solutions, but the optimal closed-loop equation has many other CLSS solutions that are not optimal (cf. Remark 5.5). In Example 5.6, we exhibit an optimal control problem whose optimal trajectories are not CLSS solutions of the optimal closed-loop equation.

This leaves the notion of classical solution as the only viable candidate for the concept of solution to be used in the definition of optimal synthesis. Example 5.3

shows that it may happen that the set of classical solutions of the optimal closed-loop equation contains, in addition to all the optimal trajectories, some other arcs that are not optimal.

It then follows that (A) is not an adequate way of handling difficulty (b), and we are left with (B). This was indeed the strategy followed by Boltyanskii in [2], Brunovský in [6], and Sussmann in [18]. In all these cases, a "regular synthesis" is defined by first specifying a—not necessarily continuous—feedback control law $x \to v(x)$. The feedback v is supposed to be "piecewise smooth" in a technical sense that, among other things, guarantees the existence of trajectories for every initial condition. Uniqueness is not assumed, but the specification of v is supplemented with a prescription for selecting a trajectory when nonuniqueness occurs. So, a "synthesis" in the sense of [2], [6], and [18] is more than just an optimal feedback: it is really a pair (v, Γ) of the kind discussed in (B).

We now review the three concepts of regular synthesis proposed in [2], [6], and [18], starting with Brunovský's definition as stated in [6], and then explaining how to modify this idea to obtain the alternative formulations suggested by Boltyanskii in [2] and Sussmann in [18]. Since the notations used in [2], [6], and [18] are different, we will give a unified account of the three definitions using a single set of symbols, and indicating which notations of [2] and [6] they correspond to. Furthermore, our accounts of [2] and [6] will be slightly modified versions of the text of the published papers, correcting what we believe are some minor imprecisions or typographical errors, and introducing some additional notations of our own for extra clarity.

Both Boltyanskii and Brunovský work with a system defined on an open subset Ω of a finite-dimensional real space (called X in [2], equal to $\mathbb{R}^n$ in [6]). Both assume that the dynamical behavior of the controlled system is given by a law $\dot{x} = f(x, u)$, where U is a subset of $\mathbb{R}^m$, and the maps f and L are of class $\mathcal{C}^1$, in the sense that they can be extended to maps of class $\mathcal{C}^1$ on an open subset of $\Omega \times \mathbb{R}^m$ that contains $\Omega \times clos_{\mathbb{R}^m}(U)$. Both assume that the synthesis is defined on an open subset S of Ω (called V in [2], G in [6]).

Brunovský's definition uses the concept of a stratification, so we review this notion first.

Definition 4.1. *Given a $k \in \{1, 2, \ldots\} \cup \{+\infty, \omega\}$, and a manifold M of class $\mathcal{C}^k$, a $\mathcal{C}^k$* stratification *in M of a subset S of M is a partition $\mathcal{P}$ of S into nonempty connected embedded submanifolds of M of class $\mathcal{C}^k$, such that $\mathcal{P}$ is locally finite in M (i.e., every compact subset of M intersects finitely many members of $\mathcal{P}$) and the following "frontier axiom" holds:*

(FA) *If $P_1, P_2 \in \mathcal{P}$, $P_1 \neq P_2$ (so that $P_1 \cap P_2 = \emptyset$), and $P_1 \cap clos_M(P_2) \neq \emptyset$, then $P_1 \subseteq clos_M(P_2)$ and $\dim(P_1) < \dim(P_2)$.*

A $\mathcal{C}^k$-stratified subset *of M is a pair $(S, \mathcal{P})$ having the property that $S \subseteq M$ and $\mathcal{P}$ is a $\mathcal{C}^k$ stratification of S in M.*

We are now ready to present Brunovský's definition, inserting some comments of our own—labeled "BP&HS"—in square brackets.

Definition 4.2. *Let $S \subseteq \Omega$ be an open subset such that the origin belongs to S. A* Brunovský regular synthesis on S for the control problem (2.1), (2.2), with target the origin, *is a* 6*-tuple $\Xi = (\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \Pi, \Sigma, v)$ (called $(\mathcal{S}, \mathcal{S}_1, \mathcal{S}_2, \Pi, \Sigma, v)$ in* [6]*) such that*

(Br.1) *$\mathcal{P}$ is the union of $\{\{0\}\}$ and a locally finite (in S) partition $\mathcal{P}\backslash\{\{0\}\}$ of $S \setminus \{0\}$ into nonempty connected embedded $\mathcal{C}^1$ submanifolds of S (called "cells"),*

(Br.2) *$\mathcal{P}\backslash\{\{0\}\}$ is the disjoint union of $\mathcal{P}_1$ (the set of "type* I *cells") and $\mathcal{P}_2$ (the set of "type* II *cells"),* (Br.3) $v : S \to U$ *(the "closed loop control"),* $\Pi : \mathcal{P}_1 \to \mathcal{P}$*, and* $\Sigma : \mathcal{P}_2 \to \mathcal{P}_1$ *are maps such that the following properties are satisfied:*

(Br.3.A) *If* $\mathcal{P}' = \{P \in \mathcal{P} : \dim P < n\,,\, P \neq \{0\}\}$*, and* $S' = \bigcup\{P \in \mathcal{P}'\}$*, then* $(S', \mathcal{P}')$ *is a stratified subset of* S *of dimension* $< n$*.* [BP&HS: *This turns out to be equivalent to the much simpler statement that* $\mathcal{P}$ is a $\mathcal{C}^1$ stratification of S*. The proof is somewhat delicate and will be given in full detail in Appendix* C.]

(Br.3.B) *The function* v *is of class* $\mathcal{C}^1$ *on each cell.*

(Br.3.C) *If* $P_1 \in \mathcal{P}_1$*, then* $f(x, v(x)) \in T_x P_1$ *(the tangent space to* P_1 *at* x*) for every* $x \in P_1$*. In addition, for each* $x \in P_1$*, if we let* ξ_x *be the maximally defined solution of the initial value problem*

$$\dot{\xi} = f(\xi, v(\xi))\,,\ \ \xi(0) = x\,,\ \ \xi \in P_1, \tag{4.1}$$

and define $t_x = \sup \mathrm{Dom}(\xi_x)$*, then the limit* $\xi_x(t_x-) \stackrel{\mathrm{def}}{=} \lim_{t\uparrow t_x} \xi_x(t)$ *exists (in* Ω*) and belongs to* $\Pi(P_1)$*.* [BP&HS: *The limit* $\xi_x(t_x-)$ *cannot belong to* P_1*, because if it did then it would be equal to the limit of* $\xi_x(t)$ *in* P_1 *(because* P_1 *is embedded), so* $\xi_x(t)$ *would have a limit in* P_1 *as* $t \uparrow t_x$*, and then* ξ_x *would be extendable to a solution of* (4.1) *on an interval containing* $[0, t_x + \varepsilon\,[$ *for some positive* ε*, contradicting the choice of* t_x*. It then follows that* $P_1 \neq \Pi(P_1) \subseteq clos(P_1)$ *and, moreover,* $\dim(\Pi(P_1)) < \dim(P_1)$ *for all* $P_1 \in \mathcal{P}_1$*. Indeed, if* $P_1 \in \mathcal{P}_1$*, then we can pick* $x \in P_1$ *and conclude that* $\xi_x(t_x-)$ *belongs to* $\Pi(P_1)\backslash P_1$*. So* $\Pi(P_1) \cap clos(P_1) \neq \emptyset$ *and* $\Pi(P_1) \neq P_1$*. Therefore* $\Pi(P_1) \subseteq clos(P_1)$ *and* $\dim(\Pi(P_1)) < \dim(P_1)$*, since* $P_1 \in \mathcal{P}$*,* $\Pi(P_1) \in \mathcal{P}$*, and* $\mathcal{P}$ *is a stratification.*]

(Br.3.D) *If* $P_2 \in \mathcal{P}_2$*, then the control* v *is continuous on* $P_2 \cup \Sigma(P_2)$*, and for each* $x \in P_2$ *there exists a unique curve* $\xi_x : [0, t_x\,[\to \Omega$ *such that the restriction* $\xi_x \lceil\,]0, t_x\,[$ *is a maximally defined integral curve of the vector field* $f(\cdot, v(\cdot))$ *on* $\Sigma(P_2)$*, and* $\xi_x(0) = x$*.* [BP&HS: *This implies that* $\Sigma(P_2) \neq P_2$*,* $P_2 \subseteq clos(\Sigma(P_2))$*, and* $\dim(\Sigma(P_2)) > \dim(P_2)$ *for every* $P_2 \in \mathcal{P}_2$*.*]

(Br.3.E) *On every cell* P*,* $x \to t_x$ *is a continuously differentiable function, and* $(t, x) \to \xi_x(t)$*,* $(t, x) \to u_x(t) \stackrel{\mathrm{def}}{=} v(\xi_x(t))$ *are continuously differentiable maps on the set*

$$E(P) \stackrel{\mathrm{def}}{=} \{(t, x) : x \in P\,,\, t \in [0, t_x]\} \tag{4.2}$$

in the sense that they can be prolonged to maps of class $\mathcal{C}^1$ *on some open subset of* $\mathbb{R} \times P$ *containing* $E(P)$*.*

(Br.3.F) *For every* $x \in S\backslash\{0\}$*, if we let* $\tilde{\xi}_x$ *denote the curve obtained in an obvious way by piecing together the trajectories on every single cell, and write* $\tilde{\eta}_x(t) = v(\tilde{\xi}_x(t))$*, then the admissible pair* $(\tilde{\xi}_x, \tilde{\eta}_x)$ *ends at the origin after passing from one cell to another a finite number of times, and is extremal.*

(Br.3.G) *The cost function* $V^{\Xi} \stackrel{\mathrm{def}}{=} V_{\Gamma(\Xi)}$ *corresponding to the synthesis* $\Gamma(\Xi) = \{(\tilde{\xi}_x, \tilde{\eta}_x)\}_{x \in S}$ *is continuous.*

Boltyanskii's definition, as given in [2], includes an extra ingredient, namely, a subset N of S where the synthesis is allowed to have a more singular behavior. The definition is formulated in terms of a sequence $\mathbf{P} = (P^0, \ldots, P^n)$ of subsets of S such that

$$P^0 \subseteq P^1 \subseteq \cdots \subseteq P^n = S, \tag{4.3}$$

and the object that we called $\mathcal{P}$ here (that is, the set of "cells") is the union

$$\mathcal{P} = \bigcup_{i=0}^{n} \mathcal{P}^i, \tag{4.4}$$

where

(Bo.*) $\mathcal{P}^i$, for $i \geq 0$, is the set of all connected components of $P^i \setminus (P^{i-1} \cup N)$, where $P^{-1} = \emptyset$,

so now $\mathcal{P}$ is a partition of $(S \setminus N) \cup \{0\}$ rather than of S. Boltyanskii does not explicitly require that this partition be locally finite in S. He asks only that the sets P^i and N be "piecewise smooth" (that is, locally finite unions of "curvilinear polyhedra," i.e., of sets that are $\mathcal{C}^1$-diffeomorphic to closed bounded polyhedra in finite-dimensional Euclidean spaces), but this does not imply that $\mathcal{P}$ is locally finite. (In fact, in section 5, Example 5.2, we exhibit a Boltyanskii regular synthesis whose set of cells is not locally finite.)

In addition, Boltyanskii imposes some extra requirements, which imply some of the Brunovský conditions. In order to facilitate the comparison, we will include in our definition of a Boltyanskii synthesis all the Brunovský conditions that necessarily follow from the Boltyanskii assumptions, even though this will introduce a number of redundancies.

Definition 4.3. *A* Boltyanskii regular synthesis *on the open subset S of the open set $\Omega \subseteq \mathbb{R}^n$, for the control problem* (2.1), (2.2), *with target the origin, is an 8-tuple $\Xi = (\mathbf{P}, \mathcal{P}_1, \mathcal{P}_2, \Pi, \Sigma, v, N, \Xi^N)$ such that*

(Bo.1) $\mathbf{P} = (P^0, \dots, P^n)$ *is a sequence of piecewise smooth subsets of S such that* (4.3) *holds.*

(Bo.2) *N is a piecewise smooth subset of S such that* $\dim(N) < n$.

(Bo.3) *If $\mathcal{P}$—the set of "cells"—is defined by* (4.4), *with the $\mathcal{P}^i$ defined by* (*Bo.*$*$), *then all the cells are embedded submanifolds of S, of class $\mathcal{C}^1$, such that* $\dim(P) = i$ *whenever* $P \in \mathcal{P}^i$. [BP&HS: *The cells are obviously connected, by definition, and form a partition of* $(S \setminus N) \cup \{0\}$. *But $\mathcal{P}$ need not be locally finite and a fortiori $\mathcal{P}$ need not be a stratification.*]

(Bo.4) $\dim(\Pi(P)) = \dim(P) - 1$ *whenever* $P \in \mathcal{P}_1$.

(Bo.5) *The only zero-dimensional cell is* $\{0\}$.

(Bo.6) Ξ^N *is a family of pairs in* $\mathrm{Adm}^{\mathcal{C}^1,0}(\tilde{f}, \{0\})$ *such that* (a) *if $(\xi, \eta) \in \Xi^N$, then $\xi^- \in N$,* (b) *for every $x \in N$ there is a $(\xi, \eta) \in \Xi^N$ such that $\xi^- = x$, and* (c) *if (ξ_1, η_1) and (ξ_2, η_2) belong to Ξ^N, and $\xi_1^- = \xi_2^-$, then $J(\xi_1, \eta_1) = J(\xi_2, \eta_2)$.*

(Bo.7) *Conditions* (Br.2) *and* (Br.3) *of Brunovský's definition hold, with the following modifications:* (a) *v is defined on $S \setminus N$ rather than on S,* (b) *the stratification requirement* (Br3.A) *is dropped, and* (c) *$V^{\Xi}(x)$ is defined to be equal to $J(\tilde{\xi}_x, \tilde{\eta}_x)$ if $x \in S \setminus N$, and to $J(\xi, \eta)$ if $x \in N$ and (ξ, η) is any member of Ξ^N such that $\xi^- = x$.*

(Bo.8) *Whenever the trajectory $\tilde{\xi}_x$ enters a new cell, it does so "at a nonzero angle." (That is, more precisely: whenever $P \in \mathcal{P}_1$ and $x \in P$, the tangent vector to $\tilde{\xi}_x$ at time t_x is not tangent to* $\Pi(P)$ *at $\tilde{\xi}_x(t_x)$.)*

It is then clear that the two concepts of a regular synthesis defined by Boltyanskii and Brunovský are not comparable. A Brunovský synthesis can fail to be a Boltyanskii synthesis because

(a) it violates the "nonzero angle" condition (Bo.8), as in the example given by Brunovský in [6],

or

(b) it violates the dimension condition (Bo.4), as in our Example 5.4 below,

or

(c) it violates condition (Bo.5), as in Example 5.1 below.

On the other hand, a Boltyanskii synthesis can fail to be a Brunovský synthesis because, for example, the set of cells could fail to be locally finite, as in our Example 5.2. Also, the Brunovský definition, as stated in [6], does not allow for an extra "singular set" N.

The uniqueness requirement of (Br.3.D) is rather strong. In [18] Sussmann proposed an even less restrictive formulation, as we now explain. For a $P \in \mathcal{P}_2$, we not only specify a cell $\Sigma(P) \in \mathcal{P}_1$ as in (Br.3.D), but also give a continuous "exiting map" $\mathcal{E}_P$, defined on the set $\{(x,t) : x \in P, 0 \le t < \varepsilon(x)\}$ and with values in $P \cup \Sigma(P)$. Here $\varepsilon : P \to \mathbb{R}$ is a continuous strictly positive function, and the map $\mathcal{E}_P$ is required to be such that, for every $x \in P$, (a) $\mathcal{E}_P(x,0) = x$, and (b) the map $]0, \varepsilon(x)[\ni t \to \mathcal{E}_P(x,t)$ takes values in $\Sigma(P)$ and is an integral curve of $f(\cdot, v(\cdot))$. So in [18] a *synthesis* is not just a 6-tuple $\Gamma = (\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \Pi, \Sigma, v)$ as in the Brunovský definition, but a 7-tuple $\Gamma = (\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \Pi, \Sigma, v, \mathcal{E})$, where the extra ingredient is the family $\mathcal{E} = \{\mathcal{E}_P\}_{P \in \mathcal{P}_2}$ of exiting maps.

Remark 4.4. The map $\mathcal{E}_P$ can be thought of as way of selecting in a continuous fashion, for each $x \in P$, an initial piece $]0, \varepsilon(x)[\ni t \to \mathcal{E}_P(x,t) \in \Sigma(P)$ of an integral curve of the feedback vector field $\Sigma(P) \ni y \to f(y, v(y)) \in T_y\Sigma(P)$ such that $\lim_{t \downarrow 0} \mathcal{E}_{P,x}(t) = x$. Naturally, if (Br.3.D) was satisfied we would just choose this curve to be the restriction of ξ_x to $]0, t_x[$.

It is not hard to produce a concept of regular synthesis that contains as special cases the three definitions of [2], [6], and [18] but does not differ too much from any of them. This can be done by removing from the definitions of [2], [6], and [18] conditions that are not really needed, such as the local finiteness requirements and Boltyanskii's conditions (Bo.4), (Bo.5), and (Bo.8). One can also eliminate the requirement that the control space U be a subset of some Euclidean space $\mathbb{R}^m$, and take U to be an arbitrary set with no extra structure, provided only that all the conditions involving differentiability of the feedback control $x \to v(x)$ are replaced by a differentiability requirement for the map $x \to \tilde{f}(x, v(x))$. One possible result is the following concept, that we will call "BB-regular synthesis," using the letters BB to stand for "Boltyanskii and Brunovský."

Definition 4.5. *A* BB-regular synthesis *on the open subset S of Ω, for the control problem* (2.1), (2.2), *with target* $\{0\}$*, is a* 9*-tuple* $\Xi = (\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \Pi, \Sigma, \mathcal{E}, v, N, \Xi^N)$ *such that*

(BB.1) $0 \in N \subseteq S$*, and* $\mathcal{P}$ *is a finite or countable partition of* $S \backslash N$ *into nonempty connected embedded submanifolds of* S *of class* $\mathcal{C}^1$ *(called "cells");*

(BB.2) $\mathcal{P}$ *is the disjoint union of* $\mathcal{P}_1$ *(the set of "type* I *cells") and* $\mathcal{P}_2$ *(the set of "type* II *cells"),*

(BB.3) $v : \bigcup\{P : P \in \mathcal{P}_1\} \to U$ *(the "closed loop control"),* $\Pi : \mathcal{P}_1 \to \mathcal{P} \cup \{\{0\}\}$*, and* $\Sigma : \mathcal{P}_2 \to \mathcal{P}_1$ *are maps, and* $\mathcal{E}$ *is a family* $\{\mathcal{E}_P\}_{P \in \mathcal{P}_2}$*, such that the following properties are satisfied:*

(BB.3.A) *If* $P \in \mathcal{P}_1$*, then the map* $P \ni x \to \tilde{f}(x, v(x))$ *is of class* $\mathcal{C}^1$*, and such that* $f(x, v(x))$ *belongs to* T_xP *(the tangent space to* P *at* x*) for every* $x \in P$*.*

(BB.3.B) *If* $P \in \mathcal{P}_2$*, then* $\Sigma(P) \in \mathcal{P}_1$*, and* $\mathcal{E}_P$ *is a map that assigns to each* $x \in P$ *a maximally defined integral curve* $\mathcal{E}_{P,x}$ *of the vector field* $\Sigma(P) \ni y \to f(y, v(y)) \in T_y\Sigma(P)$ *in such a way that* $0 = \inf \mathrm{Dom}(\mathcal{E}_{P,x})$ *and* $\lim_{t\downarrow 0} \mathcal{E}_{P,x}(t) = x$*.*

(BB.3.C) *For* $x \in P \in \mathcal{P}_2$*, define a map* $\xi_x : \{0\} \cup \mathrm{Dom}(\mathcal{E}_{P,x}) \to \Omega$ *by letting* $\xi_x(0) = x$ *and* $\xi_x(t) = \mathcal{E}_{P,x}(t)$ *for* $t \in \mathrm{Dom}(\mathcal{E}_{P,x})$*. For* $x \in P_1 \in \mathcal{P}_1$*, let* ξ_x *be the maximally defined solution of the initial value problem* (4.1)*. Define* $t_x = \sup \mathrm{Dom}(\xi_x)$ *for* $x \in P \in \mathcal{P}$*. Then for every cell* $P \in \mathcal{P}$ *the function* $P \ni x \to t_x$ *is of class* $\mathcal{C}^1$*, and the maps* $\hat{E}(P) \ni (t, x) \to \xi_x(t) \in \Omega$ *and* $\hat{E}(P) \ni (t, x) \to \tilde{f}(\xi_x(t), v(\xi_x(t))) \in \mathbb{R}^{n+1}$ *have continuously differentiable extensions to a neighborhood of* $E(P)$ *in* $\mathbb{R}\times P$*, where* $E(P)$ *is the set defined by* (4.2)*, and* $\hat{E}(P) = \{(t, x) \in E(P) : 0 < t < t_x\}$*.*

(BB.3.D) *If* $x \in P \in \mathcal{P}_1$*, then the limit* $\xi_x(t_x-) = \lim_{t\uparrow t_x} \xi_x(t)$*—which exists because of* (BB.3.C)*—belongs to* $\Pi(P)$*.*

(BB.4) *Boltyanskii's condition* (Bo.6) *holds.*

(BB.5) *Brunovský's condition* (Br.3.F) *holds for* $x \in S \setminus N$*.*

(BB.6) *Brunovský's condition* (Br.3.G) *holds, with* V^{Ξ} *defined as in* (Bo.7.c)*.*

It is then easy to show that if Ξ is a regular synthesis in the sense of Boltyanskii [2], Brunovský [6], or Sussmann [18], then it is possible to associate to Ξ in a natural way a BB-regular synthesis in the sense of Definition 4.5. Moreover, a BB-regular synthesis gives rise in a natural way to a regular presynthesis $\Gamma(\Xi)$ in the sense of our Definition 2.12, which is a synthesis if $N = \{0\}$.

Our Definition 2.12—a slightly different version of which was already proposed in [31]—is even more general, because of the following:

(a) We do not require that V_Γ be continuous, and we assume only the weak continuity conditions.

(b) We do not require that the domain of the synthesis be an open set. This is quite important because one wants the theory to apply to systems that are not locally controllable, and in those cases one usually wants to take S to be the set $\hat{S}$ of all points of Ω that can be steered to the target, and in general $\hat{S}$ is not open.

(c) We do not require the existence of a partition into "cells." From our point of view, such a partition is only needed to guarantee the weak differentiability condition (b) of Definition 2.12.

(d) Even when a good partition into "cells" exists, satisfying the most restrictive conditions of both definitions (that is, Brunovský's stratification condition (Br.3.A) and Boltyanskii's conditions (Bo.4), (Bo.5), and (Bo.8)), it may still happen that the crucial "finite number of steps" condition (Br.3.F) is violated, as in Fuller's example, but our requirements for a regular presynthesis are still met.

5. Examples.

We now present several examples of optimal regular syntheses to illustrate the differences between the various definitions.

One example showing how our notion of regular synthesis is more general than those of Boltyanskii and Brunovský was already discussed in section 3, where we showed that our theory applies to Fuller's problem while the other ones clearly do not, since they do not allow the trajectories of the synthesis to have infinitely many switchings. Another important example, relevant for the comparison of the Boltyanskii and Brunovský definitions, is the one given by Brunovský in [6], showing how the "nonzero angle" condition (Bo.8) can be violated.

We now show by means of an example how Boltyanskii's condition (Bo.5) can be violated. In principle, it is quite easy to produce examples of syntheses that satisfy all of Boltyanskii's conditions except for (Bo.5). It turns out, however, that in most such examples one can construct a new synthesis where (Bo.5) holds, for example, by incorporating the zero-dimensional cells in the singular set N, or by drawing a one-dimensional arc A_q through every q such that $\{q\}$ is a cell and making the A_q's cells of the new synthesis. To produce an example where no synthesis can possibly satisfy (Bo.5) more work is needed, but it can be done, as we now show.

Example 5.1. Let

$$S_1 = \{(x,0) : x \leq 0\}\,, \qquad S_2 = \{(x,x) : 0 \leq x \leq 1\}\,,$$
$$S_3 = \{(x,2x-1) : x \geq 1\}\,, \qquad S_4 = \{(x,0) : x \geq 0\}\,.$$

Let $\Lambda_1 = S_1 \cup S_2 \cup S_3$, $\Lambda_2 = S_1 \cup S_4$. Let A be the set of points (x,y) such that $x \geq 0$ and either $0 \leq y \leq x \leq 1$ or $0 \leq y \leq 2x-1$, and let B be the set of those (x,y) such that either $y \leq 0$, or $x \leq 0$, or $0 \leq x \leq 1$ and $x \leq y$, or $x \geq 1$ and $y \geq 2x-1$. Let $\varphi, \psi : \mathbb{R}^2 \to \mathbb{R}$ be smooth nonnegative functions such that $\varphi(x,y) = 0$ if and only if $(x,y) \in A$, and $\psi(x,y) = 0$ if and only if $(x,y) \in B$.

We consider the minimum time-optimal control problem in $\mathbb{R}^2$ with target $\{(0,0)\}$ and dynamics given by

$$\dot{x} = \frac{u_1}{1+\varphi(x,y)}\,, \qquad \dot{y} = \frac{u_2}{1+\psi(x,y)}\,,$$

where the control constraint is $|u_1| + |u_2| \leq 1$.

If $q = (\bar{x},\bar{y})$ is any point in $\mathbb{R}^2$, and $[a,b] \ni t \to \xi(t) = (x(t),y(t)) \in \mathbb{R}^2$ is any trajectory from q to the target, corresponding to a control $[a,b] \ni t \to \eta(t) = (u_1(t),u_2(t))$, then

$$b - a \geq \int_a^b \Big(|u_1(t)| + |u_2(t)|\Big)\,dt \geq \int_a^b \Big(|\dot{x}(t)| + |\dot{y}(t)|\Big)\,dt \geq |\bar{x}| + |\bar{y}|\,.$$

Moreover, a trajectory ξ_q from q to the origin whose cost is exactly $|\bar{x}| + |\bar{y}|$ can be constructed as follows:

1. If q belongs to Λ_1, then the point $\xi_q(t)$ moves along Λ_1 towards the origin, with a control vector $\eta_q(t) = (u_1(t),u_2(t))$ such that $|u_1(t)| + |u_2(t)| = 1$ (so that $\eta_q(t) = (1,0)$ if $q \in S_1$ and $\eta_q(t) = (-\frac{1}{2},-\frac{1}{2})$ if $q \in S_2$; if $q \in S_3$, then $\eta_q(t) = (-\frac{1}{3},-\frac{2}{3})$ until $\xi_q(t) = (1,1)$, and $\eta_q(t) = (-\frac{1}{2},-\frac{1}{2})$ from then on).
2. If $q \in A$, then $\xi_q(t)$ moves horizontally to the left (with control $(u_1,u_2) = (-1,0)$) until it reaches Λ_1, and then follows the trajectory described in item 1.
3. If $q \in B$ and $\bar{y} > 0$, then $\xi_q(t)$ moves vertically down (with control $(u_1,u_2) = (0,-1)$) until it reaches Λ_1, and then follows the trajectory described in item 1.
4. If $q \in B$ and $\bar{y} < 0$, then $\xi_q(t)$ moves vertically up (with control $(u_1,u_2) = (0,1)$) until it reaches Λ_2, and then moves horizontally towards the origin with control $(1,0)$ or $(-1,0)$.

The family $\Xi = \{\xi_q\}_{q\in\mathbb{R}^2}$ is an optimal synthesis. If we define P^0 to be the two-point set consisting of $(0,0)$ and $(1,1)$, and let $P^1 = \cup_{i=1}^4 S_i$, $P^2 = \mathbb{R}^2$, and $N = \emptyset$, then all the conditions of a Boltyanskii regular synthesis are satisfied (with 17 cells), except for (Bo.5). Moreover, it is easy to show that *this problem does not admit a regular* *synthesis that satisfies all the Boltyanskii conditions, including* (Bo.5). (The proof is as follows. The optimal trajectories for our problem are clearly unique for all initial points $q = (\bar{x}, \bar{y})$. So for any optimal synthesis the trajectories are the ξ_q. Assume we have a regular synthesis satisfying all of Boltyanskii's conditions. Then the point $(1, 1)$ cannot belong to the singular set N, because the definition of a Boltyanskii synthesis implies that every marked trajectory starting at a point of $\mathbb{R}^2 \setminus N$ is entirely contained in $\mathbb{R}^2 \setminus N$, so the set $U = \{(x, y) : x \geq 1\,,\, y \geq 1\}$ must be entirely contained in N—because if $q \in U$, then ξ_q goes through $(1, 1)$—and this contradicts the fact that the dimension of N is at most one (as below). Now, let C be the cell containing $(1, 1)$. Since every optimal trajectory going through $(1, 1)$ has a discontinuous velocity at $(1, 1)$, the cell C cannot be of type I. In particular, C is not two-dimensional, so (Bo.5) implies that it is one-dimensional. It follows from the definition of a Boltyanskii synthesis that the direction $\nu(q)$ of the marked trajectory ξ_q at its starting point q must be a continuous function of q as long as $q \in C$. But $\nu(1, 1) = (-\frac{1}{2}, -\frac{1}{2})$, and there is no direction along which $\nu(q)$ is continuous as $q \to (1, 1)$, so it is impossible for C to be one-dimensional.)

In our second example, we present a Boltyanskii optimal synthesis which is not a Brunovský synthesis, because the set of cells is not locally finite.

Example 5.2. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^\infty$ such that (a) $\psi(x) = 0$ whenever $x \leq 0$, (b) $0 \leq \psi(x) < x$ when $x > 0$, (c) the set of zeros of ψ is the union of $\{0\}$ and the points of a decreasing infinite sequence $\{x_k\}_{k=1}^\infty$ having an accumulation point at 0, (d) for every $k \geq 1$ there exists $\bar{x}_k \in]x_{k+1}, x_k[$ such that ψ is strictly increasing on $[x_{k+1}, \bar{x}_k]$ and strictly decreasing on $[\bar{x}_k, x_k]$ and (e) ψ is strictly increasing on $[x_1, +\infty[$. Let

$$\begin{aligned}
&A = \{(x, y) \in \mathbb{R}^2 : x \geq 0 \ \text{ and } \ -\psi(x) \leq y \leq \psi(x)\},\\
&B = \{(x, y) \in \mathbb{R}^2 : x \leq |y|\},\\
&E = \{(x, y) \in \mathbb{R}^2 : x \geq 0 \text{ and } (\psi(x) \leq y \leq x \text{ or } -x \leq y \leq -\psi(x))\}.
\end{aligned}$$

Let L be the y axis. Then A, B, E, and L are closed subsets of $\mathbb{R}^2$. Let $\sigma : \mathbb{R}^2 \to \mathbb{R}$ be a function of class $\mathcal{C}^\infty$ such that $\sigma \equiv 0$ on $A \cup B$ and $\sigma > 0$ on $\mathbb{R}^2 \setminus (A \cup B)$. Let $\tau : \mathbb{R}^2 \to \mathbb{R}$ be a function of class $\mathcal{C}^\infty$ such that $\tau \equiv 0$ on $E \cup L$ and $\tau > 0$ on $\mathbb{R}^2 \setminus (E \cup L)$.

Consider the optimal control problem in $\mathbb{R}^2$ with target $\{(0, 0)\}$, in which the dynamics is given by

$$\dot{x} = u_1, \qquad \dot{y} = u_2\,,$$

the control constraints are $u_1 \in \{-1, 0, 1\}$ and $|u_1| + |u_2| = 1$, and the cost functional to be minimized is

$$J = \int_a^b \Big(u_1(t)^2\Big(1 + \sigma(x(t), y(t))\Big) + u_2(t)^2\tau(x(t), y(t))\Big)\, dt.$$

If $q = (\bar{x}, \bar{y}) \in \mathbb{R}^2$, and $[a, b] \ni t \to \xi(t) = (x(t), y(t))$ is any trajectory going from q to the origin, then $J(\xi) \geq \int_a^b u_1(t)^2\, dt = \int_a^b |u_1(t)|\, dt \geq |\int_a^b \dot{x}(t)\, dt| = |\bar{x}|$. On the other hand, there exists a trajectory ξ_q for which $J(\xi_q) = |\bar{x}|$. To see this, observe that purely vertical motion—i.e., $|u_2| = 1$ and $u_1 = 0$—costs nothing as long as it takes place in $E \cup L$, while horizontal motion—i.e., $|u_1| = 1$ and $u_2 = 0$—has a cost per unit time equal to 1 on $A \cup B$. If $q \in B$, then we can move horizontally towards the y axis—using $u_1 = -\mathrm{sgn}(\bar{x})$—and then move vertically along the y axis and end up at the origin, thereby obtaining a trajectory whose cost is exactly $|\bar{x}|$. If $q \in E$, then we can move vertically up or down—using $u_2 = \mathrm{sgn}(\bar{y})$—until we hit the boundary of B. Once we are in B, we follow the optimal trajectory already described for points of B, and end up producing a trajectory from q to the origin whose cost is $|\bar{x}|$. If $q \in A$ and $\bar{y} = 0$, then the constant control $u_1 = -1$, $u_2 = 0$ steers q to the origin with cost $\bar{x}$. If $q \in A$ and $\bar{y} \neq 0$, then we can use the constant control $u_1 = -1$, $u_2 = 0$ until we reach a point $(\hat{x}, \hat{y})$ such that $|\hat{y}| = \psi(\hat{x})$. The cost of this is clearly $\bar{x} - \hat{x}$, and $(\hat{x}, \hat{y})$ belongs to E, so we know how to go from $(\hat{x}, \hat{y})$ to the origin with cost $\hat{x}$. So we can go from q to the origin with cost $\bar{x}$.

Therefore the value function V for this problem is given by $V(x, y) = |x|$. Moreover, we have given an explicit description of a family $\Xi = \{\xi_q\}_{q\in\mathbb{R}^2}$ which is a total optimal synthesis.

We now show that Ξ *is the family of trajectories of a Boltyanskii regular synthesis whose set of cells is not locally finite.* To see this, we define N to be the set $[0, \infty[\, \times \{0\}$, so N is clearly piecewise smooth. We take $P^0 = \{(0, 0)\}$, and let P^1 be the union of the two coordinate axes, the two half-lines $L_1 = \{(x, x) : x \geq 0\}$ and $L_2 = \{(x, -x) : x \geq 0\}$, and the graphs of ψ and $-\psi$. Finally, we let $P^2 = \mathbb{R}^2$. It is then clear that P^0, P^1, and P^2 are piecewise smooth. The connected components of $P^1 \setminus (P^0 \cup N)$ are the open half-lines

$$\begin{aligned}
&\Lambda_1 = \{(x, x) : x > 0\}\,, &&\Lambda_2 = \{(0, y) : y > 0\}\,, &&\Lambda_3 = \{(x, 0) : x < 0\}\,,\\
&\Lambda_4 = \{(0, y) : y < 0\}\,, &&\Lambda_5 = \{(x, -x) : x > 0\}
\end{aligned}$$

and the arcs

$$\begin{aligned}
A_0^+ &= \{(x, \psi(x)) : x > x_1\}\,, & A_0^- &= \{(x, -\psi(x)) : x > x_1\}\,,\\
A_k^+ &= \{(x, \psi(x)) : x_{k+1} < x < x_k\}\,, & A_k^- &= \{(x, -\psi(x)) : x_{k+1} < x < x_k\}\,.
\end{aligned}$$

The connected components of $P^2 \setminus (P^1 \cup N)$ are the sets

$$\begin{aligned}
&U_1 = \{(x, y) : x > 0\,,\, \psi(x) < y < x\}\,, &&U_2 = \{(x, y) : 0 < x < y\}\,,\\
&U_3 = \{(x, y) : x < 0 < y\}\,, &&U_4 = \{(x, y) : x < 0\,,\, y < 0\}\,,\\
&U_5 = \{(x, y) : x > 0\,,\, y < -x\}\,, &&U_6 = \{(x, y) : x > 0\,,\, -x < y < -\psi(x)\}
\end{aligned}$$

and the bounded regions

$$\begin{gathered}
W_0^+ = \{(x, y) : x > x_1\,,\, 0 < y < \psi(x)\}\,,\ \ W_0^- = \{(x, y) : x > x_1\,,\, 0 > y > -\psi(x)\}\,,\\
W_k^+ = \{(x, y) : x_{k+1} < x < x_k\,,\, 0 < y < \psi(x)\}\,,\\
W_k^- = \{(x, y) : x_{k+1} < x < x_k\,,\, 0 > y > -\psi(x)\}\,.
\end{gathered}$$

The arcs A_0^+, A_0^-, A_k^+, and A_k^- and the half-lines Λ_1, Λ_5 are declared to be type II cells, and we choose

$$\Sigma(A_0^+) = \Sigma(A_k^+) = U_1\,, \quad \Sigma(A_0^-) = \Sigma(A_k^-) = U_6\,, \quad \Sigma(\Lambda_1) = U_2\,, \quad \Sigma(\Lambda_5) = U_5\,.$$

The remaining cells (that is, Λ_2, Λ_3, Λ_4, the U_i, W_0^+, W_0^-, the W_k^+ and the W_k^-) are type I cells. The feedback v is defined by letting

$$v \equiv \begin{cases} (-1, 0) & \text{on} \quad Z\,,\\ (1, 0) & \text{on} \quad U_3 \cup U_4\,,\\ (0, 1) & \text{on} \quad U_1 \cup \Lambda_4\,,\\ (0, -1) & \text{on} \quad U_6 \cup \Lambda_2\,,\end{cases}$$

where

$$Z = (N \setminus \{(0,0)\}) \bigcup (W_0^+ \cup W_0^-) \bigcup (\cup_k W_k^+) \bigcup (\cup_k W_k^-) \bigcup U_2 \bigcup U_5 .$$

The definitions of Π and Ξ^N are the obvious ones. It is then clear that all the conditions of the Boltyanskii definition are satisfied, but the set of cells is not locally finite.

Example 5.3. We now present an example of a synthesis that satisfies all the Brunovský conditions except for the continuity requirement (Br.3.G). This example will, in addition, exhibit the phenomenon of nonuniqueness of trajectories for the closed-loop equation arising from the optimal feedback, thereby providing a concrete illustration of the reasons for including as an extra ingredient a selection of trajectories, i.e., for following strategy (B) of section 4.

We construct and study in detail the time-optimal synthesis for the planar system:

$$\dot{q} = F(q) + u\,G(q)\,, \qquad |u| \le 1, \tag{5.1}$$

where

$$q \overset{\text{def}}{=} \begin{pmatrix} x \\ \\ y \end{pmatrix} \in \mathbb{R}^2\,, \qquad F(q) \overset{\text{def}}{=} \begin{pmatrix} 1 - \frac{y}{2} \\ \\ \frac{x+1}{2} \end{pmatrix}, \qquad G(q) \overset{\text{def}}{=} \begin{pmatrix} -\frac{y}{2} \\ \\ \frac{x+1}{2} \end{pmatrix}, \tag{5.2}$$

and the target is the origin of $\mathbb{R}^2$.

The trajectories of (5.1) corresponding to the constant control $u \equiv -1$ are straight horizontal lines going from left to right, while those corresponding to $u \equiv +1$ are circles centered at the point $(-1, 1)$, running counterclockwise. The optimal synthesis, to be determined below, is shown in Figure 2.

An arc $\gamma : [a, b] \to \mathbb{R}^2$ is *simple* if γ is an injective map. Clearly, all optimal trajectories are simple.

Since $G(q) = 0$ only for $q = (-1, 0)$, it is clear that a trajectory $\gamma : [a, b] \to \mathbb{R}^2$ uniquely determines the corresponding control $u(\cdot)$, unless the set $\{t : \gamma(t) = (-1, 0)\}$

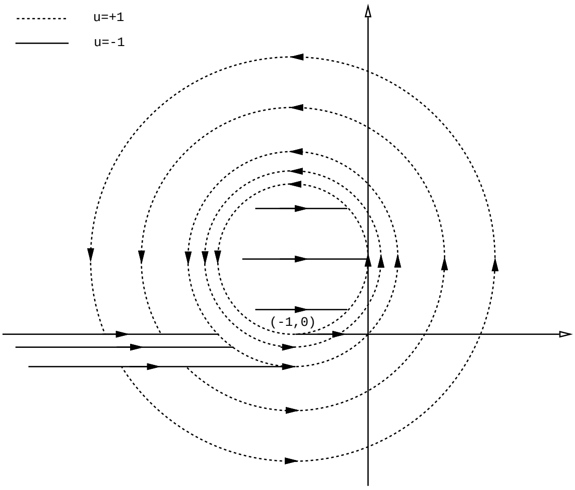


FIG. 2.

is of positive measure. In particular, for simple trajectories γ—the only ones that will be of interest to us—the control is uniquely determined by γ.

As in previous sections, we use $\mathbb{R}_k$ to denote the space of real row k-vectors. If Ω is an open subset of $\mathbb{R}^k$ and $X, Y : \Omega \to \mathbb{R}^k$ are vector fields of class C^1, then $[X, Y]$ will denote the Lie bracket of X and Y, so $[X, Y] = (DY).X - (DX).Y$.

If X, Y are $\mathbb{R}^2$-valued functions on a subset of $\mathbb{R}^2$, then (X, Y) will denote the 2×2 matrix with columns X, Y. In particular, following Sussmann [25], we define two real-valued functions Δ_A, Δ_B on $\mathbb{R}^2$, by letting

$$\Delta_A = \det(F, G) \qquad \Delta_B = \det(G, [F, G]) \tag{5.3}$$

where det stands for determinant. From the results of [25], we know that every extremal trajectory of (5.1) is a finite concatenation of "bang" and "singular" arcs. An arc is a *bang arc* if it is a B_- arc—i.e., a trajectory for the control $u = -1$—or a B_+ arc—i.e., a trajectory for $u = 1$. Singular arcs lie in the set of zeros of the function Δ_B. In our case,

$$\Delta_A = \frac{x+1}{2} \quad \text{and} \quad \Delta_B = -\frac{y}{4}, \tag{5.4}$$

so the only singular arcs are those entirely contained in the x axis, and these correspond to the control -1, so they actually are bang. So *all the extremals of our problem are bang-bang, that is, finite concatenations of bang arcs.*

If $\Delta_A(q) \neq 0$, then the vectors $F(q)$ and $G(q)$ are linearly independent, and we can define the numbers $f(q)$ and $g(q)$ as the coefficients of the linear combination:

$$[F, G](q) = f(q)F(q) + g(q)G(q). \tag{5.5}$$

Let Ω_A, Ω_B be, respectively, the sets $\{q : \Delta_A(q) \neq 0\}$, $\{q : \Delta_B(q) \neq 0\}$, and define $\Omega_{AB} = \Omega_A \cap \Omega_B$. It was proved in [25] that, if γ is an extremal trajectory and $\gamma(t) \in \Omega_{AB}$ for every t, then γ is bang-bang with at most one switching, and this switching can only be from the control value -1 to the value $+1$ if $f > 0$, and from $+1$ to -1 if $f < 0$. In our case, it follows from (5.4) that

$$\Omega_A = \{(x, y) : x \neq -1\}, \qquad \Omega_B = \{(x, y) : y \neq 0\}, \tag{5.6}$$

so Ω_{AB} is the union of four open quadrants $\Omega_{AB,i}$, $i = 1, 2, 3, 4$, defined by

$$\begin{aligned} \Omega_{AB,1} &= \{(x, y) : x > -1\,,\, y > 0\}, \\ \Omega_{AB,2} &= \{(x, y) : x < -1\,,\, y > 0\}, \\ \Omega_{AB,3} &= \{(x, y) : (x, -y) \in \Omega_{AB,2}\}, \\ \Omega_{AB,4} &= \{(x, y) : (x, -y) \in \Omega_{AB,1}\}. \end{aligned} \tag{5.7}$$

Moreover,

$$[F, G] = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix} = \frac{y}{2(x+1)} F + \frac{2-y}{2(x+1)} G, \tag{5.8}$$

so $f > 0$ on $\Omega_{AB,1} \cup \Omega_{AB,3}$ and $f < 0$ on $\Omega_{AB,2} \cup \Omega_{AB,4}$. So the control switchings are from -1 to 1 on $\Omega_{AB,1} \cup \Omega_{AB,3}$ and from 1 to -1 on $\Omega_{AB,2} \cup \Omega_{AB,4}$.

To determine the optimal synthesis, we will first construct a "sufficient family" of trajectories for our problem, i.e., a collection $\mathcal{F}$ of simple trajectories such that every optimal trajectory ending at the target is in $\mathcal{F}$. We will then find, for every initial point q, a $\gamma_q \in \mathcal{F}$ such that if $\delta \in \mathcal{F}$ goes from q to 0 and $\delta \neq \gamma_q$, then δ is not optimal. Since the existence of optimal trajectories from any initial point follows by elementary arguments, we will be able to conclude that γ_q is optimal.

Suppose that $\gamma : [a,b] \to \mathbb{R}^2$ is time-optimal, $\gamma(b) = (0,0)$, $a < b$, and γ arises from an open-loop control $t \to u(t)$. Then γ is simple, and is a concatenation of finitely many maximal bang arcs. Since γ is an extremal, we can fix a nontrivial minimizing adjoint vector λ along γ and a constant $\lambda_0 \geq 0$ for which (EX4) holds. Define the *switching function*

$$\varphi(t) = \lambda(t) \cdot G(\gamma(t)).$$

The Hamiltonian minimization condition implies that

$$\varphi(t) \neq 0 \implies u(t) = -sgn(\varphi(t)). \tag{5.9}$$

Let $\delta : [c,d] \to \mathbb{R}^2$ be a maximal B_- piece of γ, so $c < d$ and δ is of the form:

$$\delta(t) = (x_0 + t, y_0). \tag{5.10}$$

For a.e. $t \in [c,d]$, we have $\dot{\lambda} = -\lambda \cdot D(F - G) = 0$. So $\lambda = (\lambda_1, \lambda_2)$ is constant, $(\lambda_1, \lambda_2) \neq (0,0)$, and

$$\varphi(t) = -\lambda_1 \frac{y_0}{2} + \lambda_2 \frac{x_0 + t + 1}{2} \tag{5.11}$$

is a linear function of t. Since a linear function cannot have more than one zero unless it vanishes identically, we conclude that $a < c < d < b$ can only be true if $\lambda_2 = y_0 = 0$. If $d = b$, then of course $y_0 = 0$, since $\gamma(b) = (0,0)$. So one of the following two possibilities occurs:

(I) δ is contained in the x axis,
(II) $a = c$.

This shows that all the maximal B_- pieces of γ are entirely contained in the x axis, with the only possible exception that the domain $[a,b]$ of γ may contain an initial segment $[a,d]$ such that $\gamma([a,d])$ is contained in a horizontal line for which $y \neq 0$.

Let us now analyze the adjoint equation for a maximal B_+ piece $\delta : [c,d] \to \mathbb{R}^2$ of γ, such that $c < d$. Clearly, δ is of the form

$$x(t) = -1 + r\cos(t+\theta)\,, \qquad y(t) = 1 + r\sin(t+\theta)\,, \qquad c \leq t \leq d, \tag{5.12}$$

where r, θ are constants, $r > 0$, and $\theta \in [0, 2\pi]$. (The possibility that $r = 0$ is obviously excluded, because γ is time-optimal and $c < d$.) The components λ_1, λ_2 of λ satisfy $\dot{\lambda}_1 = -\lambda_2$ and $\dot{\lambda}_2 = \lambda_1$, so there exist constants $A > 0$, $\theta_1 \in [0, 2\pi]$, such that

$$\lambda_1(t) = A\cos(t+\theta_1)\,, \qquad \lambda_2(t) = A\sin(t+\theta_1), \qquad c \leq t \leq d. \tag{5.13}$$

Therefore, up to multiplication by a positive constant,

$$\varphi(t) = -\cos(t+\theta_1) + r\sin(\theta_1 - \theta). \tag{5.14}$$

Now suppose that one of the two endpoints c, d of the domain of δ is a switching time and the corresponding point in $\mathbb{R}^2$ lies in the x axis. That is, we assume that $h \in \{c, d\}$ is such that $\varphi(h) = 0$ and $y(h) = 0$. Then

$$r\sin(h+\theta) = -1, \tag{5.15}$$

so

$$\begin{aligned}\varphi(t) &= -\cos(t+\theta_1) + r(\sin(h+\theta_1)\cos(h+\theta) - \cos(h+\theta_1)\sin(h+\theta)) \\ &= -\cos(t+\theta_1) + \cos(h+\theta_1) + r\sin(h+\theta_1)\cos(h+\theta).\end{aligned} \tag{5.16}$$

In particular,

$$0 = \varphi(h) = r\sin(h+\theta_1)\cos(h+\theta), \tag{5.17}$$

so $\varphi(t) = -\cos(t+\theta_1)+\cos(h+\theta_1)$. If $r \neq 1$, then (5.15) implies $0 \geq \sin(h+\theta) \neq -1$, so $\cos(h+\theta) \neq 0$. But then (5.15) implies $\sin(h+\theta_1) = 0$, so $\cos(h+\theta_1) = \pm 1$. The possibility that $\cos(h+\theta_1) = 1$ is excluded, for in that case φ would be > 0 throughout $]\,c, d\,[$, contradicting the minimization condition. So $\varphi(t) = -1 - \cos(t+\theta_1)$. This shows that $\varphi(t) < 0$ for $t \in [c,d] \setminus \{h\}$, unless $d - c = 2\pi$. But if $d - c = 2\pi$, then γ would not be simple. So $d - c < 2\pi$, and we have shown that the other endpoint h' of $[c,d]$ cannot be a switching point, and therefore $h' \in \{a, b\}$.

It then follows that one of the following possibilities must occur:

(i) $d = b$;

(ii) $a = c < d < b$ and $\gamma(d)$ lies in the x axis;

(iii) $a < c < d < b$ and $\gamma(d) = (-1, 0)$.

Indeed, if $d < b$ but $\gamma(d)$ is not in the x axis, then we could let δ' be the maximal B_- piece of γ that starts at time d, and our analysis of the B_- pieces of γ, applied to δ', would show that δ' cannot in fact occur, since δ' is not an initial segment of γ and is not contained in the x axis. So, if $d < b$, then $\gamma(d)$ lies in the x axis, and d is a switching time of γ. Our previous argument shows that $c = a$, unless $r = 1$. But if $r = 1$, and $\gamma(d)$ is in the x axis, then $\gamma(d) = (-1, 0)$. So $d < b$ implies that either (ii) or (iii) must hold.

We now combine the results about the two types of pieces and provide a complete description of the optimal trajectories ending at the origin. Let us use L, L_0, C to denote, respectively, the trajectory types "B_- and not contained in the x axis," "B_- and contained in the x axis," and "B_+." (The symbols L and C stand for "line" and "circle.") Let us use $T_1T_2\ldots T_m$ to denote the trajectory type "T_1 followed by T_2 $\ldots$ followed by T_m."

With this notation, we now show that *a nontrivial optimal trajectory γ ending at the origin is of one of the following seven types*: L_0, C, L_0C, LC, CL_0, LCL_0, CL_0C. First, concatenations of six or more types are ruled out as follows: any such concatenation would have to contain at least three different C pieces; at most one of them can satisfy (i) and at most one can satisfy (ii); if both of these possibilities occurred, then we would have a concatenation of the form $C * C * C$, i.e., five pieces rather than six; so there must be two different pieces satisfying (iii), and then γ would go at least twice through the point $(-1, 0)$, and would not be simple. Next, all concatenations containing an L_0CL_0 sequence are excluded, because if γ is any such concatenation, then the C piece that lies between the two L_0s would have to satisfy (iii), so its last switching would have to happen at $(-1, 0)$; then the C piece would have to be an arc of the circle with center $(-1, 1)$ and radius 1, which intersects the x axis only at $(-1, 0)$; then the other switching would also happen at $(-1, 0)$, and γ would not be simple. A five-piece concatenation must be CL_0CL_0C, LCL_0CL_0, or $L_0CL_0CL_0$, because every B_- piece must satisfy (I) or (II), so it cannot be an L unless it is the first piece. So all such concatenations contain an L_0CL_0 piece and are therefore excluded. A four-piece concatenation must be CL_0CL_0, L_0CL_0C, or

LCL_0C. The types CL_0CL_0 and L_0CL_0C are ruled out because they contain an L_0CL_0 triple. The type LCL_0C is excluded because the first C would have to satisfy (iii), so the switching from C to L_0 would have to happen at $(-1, 0)$; then the L_0 piece following the switching has to be a segment going from $(-1, 0)$ to a point $(\alpha, 0)$ and α must satisfy $-1 < \alpha < 0$, for if α was ≥ 0, then the L_0 piece would already reach the origin before switching to the last C; and if $-1 < \alpha < 0$, then the C arc starting at $(\alpha, 0)$ does not go through the origin. So all four-piece concatenations are excluded. Of the three-piece possibilities, CLC, LCL, and L_0CL are excluded because an L piece can occur only at the beginning of the sequence, and we also know that L_0CL_0 is excluded. So only CL_0C and LCL_0 are left. Of the two- and one-piece concatenations, CL and L are obviously excluded, and we are left with L_0, C, CL_0, L_0C, and LC.

The LCL_0 trajectories are further restricted by the condition that the switching from C to L_0 must happen at the point $(-1, 0)$. Let us call those LCL_0 trajectories that satisfy this extra condition "good."

Let $\mathcal{F}$ be the set of all simple trajectories that end at the origin and are either of one of the six types L_0, C, L_0C, LC, CL_0, CL_0C, or of type LCL_0 and good. We have shown that every optimal trajectory belongs to $\mathcal{F}$. It will turn out that the optimal synthesis involves all seven types listed above, but not all trajectories of one of these types are optimal, so we need a finer analysis.

To begin with, observe that $\dot{x} = 1 - y$ along a B_+ arc, so $\dot{x} > 1$ along a B_+ arc as long as $y < 0$. Therefore, if $q_1 = (x_1, y_1)$ and $q_2 = (x_2, y_2)$ satisfy $y_1 = y_2 \leq 0$, $x_1 < x_2$, and $x_1 + x_2 = -2$, so that q_1 and q_2 can be joined both by a C arc and an L (or L_0) arc, then the C arc is contained in the lower half-plane, and is faster than the L or L_0 arc. In particular, if we let $L_0(\beta)$ denote, for $\beta > 0$, the L_0 arc going from $(-\beta, 0)$ to $(0, 0)$, we see that $L_0(\beta)$ *is not optimal if* $1 < \beta \leq 2$, because the piece of L_0 going from $(-\beta, 0)$ to $(-2 + \beta, 0)$ can be replaced by a faster type C arc. Then the principle of optimality implies that $L_0(\beta)$ *is not optimal if* $1 < \beta$, because if $\beta > 2$, then $L_0(\beta)$ contains $L_0(2)$, which is not optimal.

Next we show that $L_0(1)$ *is optimal.* It suffices to notice that there is no trajectory from $(-1, 0)$ to $(0, 0)$ of any of the types C, LC, LCL_0, and exactly one trajectory of each of the types L_0, L_0C, CL_0, and CL_0C, but the last three are not simple, so $L_0(1)$ is left as the only possible candidate. It then follows from the principle of optimality that $L_0(\beta)$ *is optimal for* $0 \leq \beta \leq 1$.

We now study the optimal trajectories that are of one of the types C, L_0C, LC, and CL_0C. Let $\gamma : [a, b] \to \mathbb{R}^2$ be such a trajectory, and take $b = 0$, so $\gamma(0) = (0, 0)$.

If γ is a C arc, then $a > -2\pi$, for otherwise γ would contain a full loop. For $0 < a < 2\pi$, we let $C(a)$ be the C arc $\gamma : [-a, 0] \to \mathbb{R}^2$ such that $\gamma(0) = (0, 0)$. It will be shown below that the arcs $C(a)$ are all optimal.

Next suppose that γ is LC. Since the C piece satisfies (5.12), we must have $r = \sqrt{2}$ and $\theta = \frac{7\pi}{4}$. Let c be the time when the switching from L to C occurs, so $-2\pi < c < 0$, for if $c \leq -2\pi$, then γ would not be simple. On $[c, 0]$, the switching function φ is given, up to multiplication by a positive constant, by

$$\varphi(t) = \sqrt{2} \sin\Big(\theta_1 - \frac{7\pi}{4}\Big) - \cos(t + \theta_1) = \sin\theta_1 + \cos\theta_1 - \cos(t + \theta_1).$$

So

$$\varphi(t) = (1 + \sin t) \sin\theta_1 + (1 - \cos t). \cos\theta_1.$$

Suppose that $\sin\theta_1 = 0$. Then $\cos\theta_1 = \pm 1$, so $\cos\theta_1 = -1$, because $\varphi(t) < 0$ when t

is negative and close to 0, since $u(t) = +1$ for such t. So $\varphi(t) = \cos t - 1$, and then $\varphi(c) \neq 0$, since $-2\pi < c < 0$. Since $\varphi(c) = 0$, we conclude that $\sin\theta_1 \neq 0$. We can then define

$$\psi(t) = \frac{1+\sin t}{1-\cos t}\,, \qquad \nu = \frac{\cos\theta_1}{\sin\theta_1}\,.$$

Then ψ is well defined and smooth on $[c,0\,[$, and $\psi(t)$ goes to $+\infty$ as $t \uparrow 0$. Moreover, φ can be expressed as $\varphi(t) = \sin\theta_1.(1-\cos t).(\nu+\psi(t))$ for $t \in [c,0\,[$. Therefore $\sin\theta_1 < 0$, and $\psi(c) = -\nu$, since $1-\cos c \neq 0$, because $-2\pi < c < 0$. Clearly, $\psi(t) \geq 0$ for all $t \in]-2\pi, 0\,[$, and ψ attains the value 0 on $]-2\pi, 0\,[$ at $t = -\frac{\pi}{2}$ and nowhere else. A simple calculation shows that ψ is strictly increasing on $]-\frac{\pi}{2}, 0\,[$. If $\nu > 0$, then $\psi + \nu$ never vanishes on $]-2\pi, 0[$, so $\varphi(c) \neq 0$. So $\nu \leq 0$. If $\nu = 0$, then $\psi(c) = -\nu$ implies $c = -\frac{\pi}{2}$. If $\nu < 0$, then ψ must take the value $-\nu$ at some point $t(\nu)$ such that $-\frac{\pi}{2} < t(\nu) < 0$, because $\psi(-\frac{\pi}{2}) = 0$ and $\psi(0-) = +\infty$. Since ψ is strictly increasing on $]-\frac{\pi}{2}, 0\,[$, we see that $t(\nu)$ is unique, and φ changes sign at $t(\nu)$. If $c < t(\nu)$, we would contradict the fact that φ has constant sign on $[c, 0]$. On the other hand, it is impossible that $c > t(\nu)$, because then the equation $\psi(t) = -\nu$ would have two solutions on $]-\frac{\pi}{2}, 0\,[$. So $c = t(\nu)$, and we have proved that $-\frac{\pi}{2} \leq c < 0$. If $-\frac{\pi}{4} < c$, then $\gamma(c) \in \Omega_{AB,4}$, and the switching is not permitted. So the only possibility left to us is $-\frac{\pi}{2} \leq c \leq -\frac{\pi}{4}$. If $c = -\frac{\pi}{2}$, then we really have an L_0C arc rather than an LC arc. So the only LC arcs that could be optimal are those for which $-\frac{\pi}{2} < c \leq -\frac{\pi}{4}$.

For $0 < \sigma < 2\pi$, $s > 0$, let $LC(s,\sigma)$ denote the trajectory defined on the interval $[-s-\sigma, 0]$ that ends at the origin and corresponds to a $u = -1$ control on $[-s-\sigma, -\sigma]$ followed by a $u = 1$ control on $[-\sigma, 0]$. We have shown that *$LC(s,\sigma)$ is not optimal unless $\frac{\pi}{4} \leq \sigma < \frac{\pi}{2}$.* We now show that *if $\frac{\pi}{4} \leq \sigma < \frac{\pi}{2}$, then $LC(s,\sigma)$ is optimal.* To see this, let $q = (x,y)$ be the starting point of $LC(s,\sigma)$, for $s > 0$, $\frac{\pi}{4} \leq \sigma < \frac{\pi}{2}$. Then $1-\sqrt{2} \leq y < 0$. It is clear that $LC(s,\sigma)$ is the only simple LC trajectory from q to $(0,0)$. Also, it is easy to see that there are no trajectories from q to $(0,0)$ of the types L_0, C, L_0C, or good LCL_0. There is exactly one CL_0 trajectory, but it is easy to see that the time $\hat{\tau}$ along this trajectory is larger than the time τ along $LC(s,\sigma)$. (Actually, $\tau = s + \sigma < |x| < \hat{\tau} - \frac{3\pi}{2}$.) Finally, there is exactly one CL_0C trajectory γ, which can also be ruled out. (Let $\bar{x} = -1 - \sqrt{2-(y-1)^2}$, so $\bar{q} = (\bar{x}, y)$ is the point where $LC(s,\sigma)$ switches from L to C. Then both $LC(s,\sigma)$ and γ go through $\bar{q}$ and coincide from that point on. So it suffices to compare the parts of $LC(s,\sigma)$ and γ up to $\bar{q}$. The time along $LC(s,\sigma)$ up to $\bar{q}$ is $s = \bar{x} - x$ and that along γ is $> \sqrt{(1+x)^2+(1-y)^2} - \sqrt{2} + \frac{3\pi}{2}$. So we have to show that $\alpha > 0$, where

$$\begin{aligned}\alpha &= \sqrt{(1+x)^2+(1-y)^2} - \sqrt{2} + \frac{3\pi}{2} - (\bar{x}-x)\\ &= \sqrt{(1+x)^2+(1-y)^2} - \sqrt{2} + \frac{3\pi}{2} + x + 1 + \sqrt{2-(y-1)^2}.\end{aligned}$$

Then, using the fact that $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$ when $a \geq 0$ and $b \geq 0$, we get the inequalities

$$\begin{aligned}\alpha &\geq \sqrt{(1+x)^2+2} - \sqrt{2} + \frac{3\pi}{2} + x + 1\\ &> \sqrt{(1+x)^2} + x + 1\\ &= |1+x| + 1 + x \geq 0,\end{aligned}$$

as desired.)

For $s > 0$, we let $L_0C(s)$ denote the trajectory that consists of a B_- arc during time s followed by a B_+ arc during time $\frac{\pi}{2}$, and ending at the origin. Then it is easily shown that *$L_0C(s)$ is optimal for every $s > 0$.* (Indeed, if $q = (x, 0)$ is the starting point of $L_0C(s)$, for $s > 0$, then $L_0C(s)$ is the only simple L_0C trajectory from q to $(0, 0)$, and there are no simple trajectories from q to $(0, 0)$ of the types C, LC, CL_0, LCL_0, or CL_0C. There is an L_0 arc, namely, $L_0(|x|)$, but we know that this arc is not optimal because $|x| > 1$.)

Now define

$$R_1 = \{(0,0)\}\,, \qquad R_2 =]-1,0[\times\{0\}\,, \qquad R_3 = \{(-1,0)\}.$$

For $q = (0, 0)$, let γ_q be the trajectory γ with domain $\{0\}$ such that $\gamma(0) = (0, 0)$. If $q = (x, 0)$ belongs to $R_2 \cup R_3$, let $\gamma_q = L_0(|x|)$. Then the γ_q, for $q \in R_1 \cup R_2 \cup R_3$, are optimal.

For $r > 0$, let $D(r)$ be the open disc with center $(-1, 1)$ and radius r, and let $\partial D(r)$ denote the boundary of $D(r)$. Let

$$R_4 = D(1)\,, \qquad R_5 = \partial D(1)\backslash\{(-1,0)\}.$$

Then every point $q \in R_4$ can be joined to $(0, 0)$ by a unique good LCL_0 trajectory. We use γ_q to denote this trajectory. Then *the arcs γ_q, for $q \in R_4$, are optimal.* (To see this, observe that if $q \in R_4$, then the only trajectories other than γ_q that are of one of our seven types and go from q to the origin are either LCL_0 but not good or not simple, or LC with a switching from L to C at a time c that violates the requirement that $-\frac{\pi}{2} < c$. So all the alternative candidates are excluded, and γ_q must be the optimal trajectory.)

A similar argument shows that *if $q \in R_5$ and γ_q is the unique CL_0 trajectory going from q to $(0, 0)$, then γ_q is optimal.*

Next, let

$$R_6 = D(\sqrt{2})\backslash(R_2 \cup R_3 \cup R_4 \cup R_5).$$

If $q \in R_6$, then the obvious candidate for optimality is the arc γ_q of type CL_0 obtained by following the C arc that starts at q until R_2 is reached, and then continuing along R_2 towards the origin. We now prove that *the arcs γ_q, $q \in R_6$, are optimal.* To see this, observe that if $q = (x, y) \in R_6$, then there are no simple arcs from q to $(0, 0)$ of the types C, L_0C, CL_0C. If $y \neq 0$, then there are no L_0 arcs either. If $y = 0$, then there is an L_0 arc, but it is $L_0(\beta)$ for $\beta = |x| > 1$, so it is not optimal. If $y \neq 0$, then there is an LC arc, but (i) if $y < 0$, then this arc switches from L to C in $\Omega_{AB,4}$, so it is not optimal; (ii) if $y > 0$, then the arc is $LC(s, \sigma)$ for some $\sigma \in]\,5\pi/4, 2\pi\,[$, so it is not optimal either. If $y \leq 0$, then γ_q is the only simple CL_0 arc, and if $y > 0$, then there are two such arcs, but the other one is not optimal, because it contains $L_0(\beta)$ for some $\beta > 1$. So we have ruled out all the alternative candidates to γ_q other than good LCL_0 arcs. We now exclude this possibility too, by doing it first for the case when $y \neq 2$. If $y \leq 0$ or $y > 2$ or $x \geq -1$, then there is no good LCL_0 arc. If $0 < y < 2$ and $x < -1$, then there is a good LCL_0 arc, but it switches from L to C at a point in $\Omega_{AB,2}$, so this arc is not optimal. So now all the alternative possibilities have been ruled out if $q = (x, y) \in R_6$ is such that $y \neq 2$ or $y = 2$ and $x > -1$. So γ_q is optimal for all such q's. Suppose now that $y = 2$ and $x < -1$. Then there is a good LCL_0 arc γ, which switches from L to C at the point $p = (-1, 2)$. This arc can be ruled out in a number of ways, of which the following qualitative argument appears to be the shortest. Let γ be defined on $[T,0]$. Let α be a small positive number—actually, any α such that $0<\alpha<\pi$ will do—and let $\tilde{\gamma}:[T-\alpha,0]\to\mathbb{R}^2$ be the trajectory such that $\tilde{\gamma}(t)=\gamma(t)$ for $t\in[T,0]$ and $\tilde{\gamma}$ is a B_+ arc on $[T-\alpha,T]$. Let $(\tilde{x},\tilde{y})=\tilde{q}=\tilde{\gamma}(T-\alpha)$. Then $\tilde{q}$ also belongs to R_6, but it's not true that $\tilde{y}=2$ and $\tilde{x}<-1$. So we already know that $\gamma_{\tilde{q}}$ is the unique optimal arc from $\tilde{q}$ to the origin. Since γ_q is a cofinal piece of $\gamma_{\tilde{q}}$, the principle of optimality implies that γ_q is optimal as well. If γ was optimal, then γ_q and γ would go from q to 0 in the same time, so $\gamma_q(T)=\gamma_{\tilde{q}}(T)=q$. Let $[\tilde{T},0]$ be the domain of $\gamma_{\tilde{q}}$, and define $\delta:[\tilde{T},0]\to\mathbb{R}^2$ by letting δ agree with $\gamma_{\tilde{q}}$ on $[\tilde{T},T]$ and with γ on $[T,0]$. Then δ is an optimal arc from $\tilde{q}$ to 0, and δ is $CLCL_0$. This, however, is impossible, because we know that $CLCL_0$ is excluded. So we have now completed the proof that *for each $q\in R_6$, γ_q is the unique optimal arc from q to the origin.*

Next, we define five sets R_7, R_8, R_9, R_{10}, R_{11} by letting

$$\begin{aligned}
R_7 &= \partial D(\sqrt{2})\cap\Big(\mathbb{R}\times\,]\,0,+\infty\,[\,\Big),\\
R_8 &= \Big\{(-2,0)\Big\},\\
R_9 &= \Big\{(x,y)\in\partial D(\sqrt{2}):-2<x<-1,\ y<0\Big\},\\
R_{10} &= \Big\{(-1,1-\sqrt{2})\Big\},\\
R_{11} &= \Big\{(x,y)\in\partial D(\sqrt{2}):-1<x<0,\ y<0\Big\}.
\end{aligned}$$

Then the sets R_1, R_7, R_8, R_9, R_{10}, and R_{11} are pairwise disjoint and

$$\partial D(\sqrt{2})=R_1\cup R_7\cup R_8\cup R_9\cup R_{10}\cup R_{11}.$$

For $q\in R_7\cup R_8\cup R_9\cup R_{10}\cup R_{11}$ there is a unique trajectory γ_q of type C going from q to $(0,0)$. The optimality of γ_q for $q\in R_7\cup R_8\cup R_9\cup R_{10}\cup R_{11}$ is proved by considerations similar to those used for R_6.

We let

$$\begin{aligned}
R_{12} &= \,]-\infty,-2\,[\,\times\{0\},\\
R_{13} &= \{(x,y):1-\sqrt{2}<y<0\,,\ x<-1\,,\ (x+1)^2+(1-y)^2>2\},\\
R_{14} &= \,]-\infty,-1\,[\,\times\{1-\sqrt{2}\}.
\end{aligned}$$

Then $R_{13}\cup R_{14}$ is exactly the set of starting points q of the trajectories $LC(s,\sigma)$, for $s>0$, $\frac{\pi}{4}\le\sigma<\frac{\pi}{2}$, and we know that for each $q\in R_{13}\cup R_{14}$ the $LC(s,\sigma)$ trajectory is unique and optimal. We define γ_q to be this trajectory. A similar fact is true for R_{12}, with $L_0C(s)$ rather than $LC(s,\sigma)$ in the role of γ_q.

Finally, we let

$$R_{15}=\mathbb{R}^2\backslash(R_1\cup\cdots\cup R_{14}).$$

For $q\in R_{15}$, we let γ_q be the unique simple CL_0C trajectory from q to $(0,0)$. We show that γ_q *is optimal for all $q\in R_{15}$*. Let $q\in R_{15}$, and let $\gamma\in\mathcal{F}$ be an extremal trajectory starting at q and verifying $\gamma(0)=0$. It is then obvious that γ is not of type L_0, C, or L_0C. The case LC is impossible, because in that case γ would have a switching at a time c such that $c<-\pi/2$. If γ is of type CL_0, then it contains an $L_0(\beta)$ piece, with $\beta>1$, and we know that any such piece is nonoptimal, so γ is not optimal. Finally, if γ is of type LCL_0 and good, then $\gamma(t) = p = (x_p, y_p) \in R_5$ for some t. It is clear that $y_p > 0$. If $x_p > -1$, then γ is not simple, so this possibility is excluded. If $x_p \leq -1$, then we can pick a positive ε such that $\hat{q} = \gamma(t - \varepsilon)$ is in R_6. Then the restriction $\hat{\gamma}$ of γ to the interval $[t - \varepsilon, 0]$ is a good LCL_0 trajectory from $\hat{q}$ to 0. Since $\hat{q} \in R_6$, we know that $\hat{\gamma}$ is not optimal. So γ is not optimal either, by the principle of optimality.

We have now completely analyzed the minimum time problem with target set $\{(0,0)\}$ for every initial condition q, and proved in all possible cases that γ_q is the unique solution whose terminal time is 0. So the synthesis $\Gamma = \{\gamma_q\}_{q \in \mathbb{R}^2}$ shown in Figure 2 is optimal, and is the unique optimal presynthesis. The partition $\mathcal{P} = (R_1, \ldots, R_{15})$ is a C^1—and, actually, C^ω—stratification of $\mathbb{R}^2$.

Let us define the feedback control $(x, y) \to v(x, y)$ by letting

$$v = \begin{cases} 0 & \text{on} \quad R_1\,, \\ -1 & \text{on} \quad R_2 \cup R_3 \cup R_4 \cup R_{12} \cup R_{13} \cup R_{14}\,, \\ +1 & \text{on} \quad R_5 \cup R_6 \cup R_7 \cup R_8 \cup R_9 \cup R_{10} \cup R_{11} \cup R_{15}\,. \end{cases}$$

We let

$$\begin{aligned} \mathcal{P}_1 &= \{R_2, R_4, R_5, R_6, R_7, R_9, R_{11}, R_{12}, R_{13}, R_{14}, R_{15}\}, \\ \mathcal{P}_2 &= \{R_3, R_8, R_{10}\}. \end{aligned}$$

We then define

$$\begin{array}{llll} \Pi(R_2) = R_1\,, & \Pi(R_4) = R_5\,, & \Pi(R_5) = R_3\,, & \Pi(R_6) = R_2\,, \\ \Pi(R_7) = R_8\,, & \Pi(R_9) = R_{10}\,, & \Pi(R_{11}) = R_1\,, & \Pi(R_{12}) = R_8\,, \\ \Pi(R_{13}) = R_9\,, & \Pi(R_{14}) = R_{10}\,, & \Pi(R_{15}) = R_{12}\,, & \Sigma(R_3) = R_2\,, \\ \Sigma(R_8) = R_9\,, & \Sigma(R_{10}) = R_{11}\,. & & \end{array}$$

Then $(\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \Pi, \Sigma, v)$ is a regular synthesis that satisfies all the conditions of Brunovský's definition, except for the continuity requirement (Br.3.G) on the cost function V_Γ. On the other hand, it is easy to verify that V_Γ *satisfies the weak continuity conditions.* So $(\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \Pi, \Sigma, v)$ gives rise to a "regular synthesis" in the sense of our definition.

Moreover, it is clear that *the* $+1$ *trajectory starting from the point* $(-1, 0)$ *is an admissible classical trajectory for the optimal feedback control but is not optimal.* So this example illustrates a very important positive aspect of the Boltyanskii–Brunovský's definition, namely, the role of the "instantaneous exit map" Σ. To understand how Σ matters, let us suppose we did not specify Σ, and allowed all the classical trajectories of the discontinuous feedback v. Then we could, for example, start from the type II cell $R_3 = \{(-1, 0)\}$ and follow the trajectory corresponding to the control $+1$ up to time 2π, returning to the starting point, after which we could go to the origin using the control -1. The curve γ_1 defined in this way is also an admissible classical trajectory of v, but it is clear that γ_1 is not time optimal. Even worse, we can define a trajectory γ_2 of v, with domain $[0, +\infty[$, corresponding to the constant control $+1$ and starting from $(-1, 0)$. This trajectory neither reaches the origin in finite time nor approaches it as time goes to infinity. Notice that the trajectories γ_1 and γ_2 satisfy the equation $\dot{\gamma}(t) = f(\gamma(t), v(\gamma(t)))$ for every t in their domain, so they satisfy the equation corresponding to the discontinuous feedback v in the classical sense. However, if we regard it as part of the specification of our synthesis that Σ has to be given as well, then the difficulty disappears.

Example 5.4. We now exhibit an optimal synthesis in Brunovský's sense that does not satisfy condition (Bo.4) of Boltyanskii's definition. We consider the minimum time problem for the system

$$\dot{x} = u, \qquad \dot{y} = v, \qquad \dot{z} = \frac{w}{\sqrt{1+x^2+y^2}},$$

with target $\{(0,0,0)\}$ and control constraint $(u,v,w) \in U$, where U is the set of those $(u,v,w) \in \mathbb{R}^3$ such that $(u^2+v^2)^{1/2} + |w| \leq 1$, i.e., the solid of revolution obtained by rotating the square $\{(u,w) : |u| + |w| \leq 1\}$ in the u,w-plane about the w axis. Define the sets

$$\begin{aligned} R_0 &= \{(0,0,0)\}, & R_3 &= \{(x,y,z) : x^2+y^2 > 0,\, z = 0\}, \\ R_1 &= \{(x,y,z) : x = y = 0,\, z > 0\}, & R_4 &= \{(x,y,z) : x^2+y^2 > 0,\, z > 0\}, \\ R_2 &= \{(x,y,z) : x = y = 0,\, z < 0\}, & R_5 &= \{(x,y,z) : x^2+y^2 > 0,\, z < 0\}. \end{aligned}$$

We then define a feedback control law $(u,v,w) : \mathbb{R}^3 \to U$ by specifying its restriction $R_i \ni (x,y,z) \to (u_i(x,y,z), v_i(x,y,z), w_i(x,y,z)) \in \mathbb{R}^3$ for every R_i as follows:

$$\begin{aligned} & u_0 = v_0 = w_0 = 0, & & u_3 = u_4 = u_5 = -\frac{x}{\sqrt{x^2+y^2}}, \\ & u_1 = v_1 = 0,\ w_1 = -1, & & v_3 = v_4 = v_5 = -\frac{y}{\sqrt{x^2+y^2}}, \\ & u_2 = v_2 = 0,\ w_2 = 1, & & w_3 = w_4 = w_5 = 0. \end{aligned} \tag{5.18}$$

If we follow the trajectories of this feedback starting from a point $(x,y,z) \in R_3 \cup R_4 \cup R_5$, we move horizontally towards the z axis with speed 1, and then, if we are not yet at the origin, we move to the origin along the z axis with speed 1. This obviously gives rise to a synthesis Γ, and it is possible to prove that Γ is optimal by applying Theorem 2.13. However, in this case it is easy to verify directly that the associated cost function

$$V_\Gamma = |z| + \sqrt{x^2+y^2}$$

is the value function of our problem. Indeed, since V_Γ is the cost function arising from the synthesis Γ, all we need is to show that the cost $b-a$ of any trajectory $[a,b] \ni t \to \xi(t) = (x(t), y(t), z(t))$ from a point $\bar{q} = (\bar{x}, \bar{y}, \bar{z})$ to the origin is bounded below by $V_\Gamma(\bar{q})$.

Assume first that $x(t)^2 + y(t)^2 > 0$ for all $t \in]\,a,b\,[$, and let $t \to (u(t), v(t), w(t))$ be the corresponding control. Write $\rho(t) = \sqrt{u(t)^2 + v(t)^2}$. Then an elementary calculation shows that

$$\begin{aligned} \frac{d}{dt}\Big(V_\Gamma(\xi(t))\Big) &= \frac{u(t).x(t) + v(t).y(t)}{\sqrt{x(t)^2+y(t)^2}} + \operatorname{sgn}(z(t)).\frac{w(t)}{\sqrt{1+x(t)^2+y(t)^2}} \\ &\geq -\rho(t) - |w(t)| \geq -1. \end{aligned}$$

Therefore, since the function $t \to V_\Gamma(\xi(t))$ is Lipschitz, and hence absolutely continuous, we can conclude that

$$V_\Gamma(\bar{q}) = -\int_a^b \frac{d}{dt}\Big(V_\Gamma(\xi(t))\Big)\,dt \leq b - a.$$

Now, if $\xi : [a, b] \to \mathbb{R}^3$ is an arbitrary trajectory from $\bar{q}$ to the origin, then ξ can be approximated by a sequence of trajectories $\xi_j : [a_j, b_j] \to \mathbb{R}^3$ from $\bar{q}$ to $(0, 0, 0)$ such that $\xi_j(t)$ is not in the z axis for $a_j < t < b_j$. Then $b_j - a_j \geq V_\Gamma(\bar{q})$ for all j, so $b - a \geq V_\Gamma(\bar{q})$. Therefore Γ is an optimal synthesis.

Notice that in this case $\Pi(R_4) = R_1$, $\Pi(R_5) = R_2$, and $\Pi(R_3) = R_0$. So the optimal synthesis Γ does not satisfy the conditions of Boltyanskii's definition, because (Bo.4) is violated, since the dimension of the exit manifold from R_i is equal to $\dim(R_i) - 2$ for $i = 3, 4, 5$.

Remark 5.5. In Example 5.4 the optimal pair (γ_q, η_q) starting at q is unique for every q, and is a classical solution of the closed-loop equation corresponding to the optimal feedback law defined by (5.18).

On the other hand, it is easy to see that *the optimal trajectories are not Filippov solutions of the optimal closed-loop equation.* Finally, we point out that, *although the optimal trajectories are CLSS solutions, it is not true that all CLSS solutions of the optimal closed-loop equation are optimal trajectories.* For example, if the initial condition is a point q which is not in the z axis and also not in the plane $z = 0$, then most CLSS solutions $t \to \gamma(t)$ will never switch at a time t such that $\gamma(t)$ lies in the z axis, and will therefore be entirely contained in a plane $z =$ constant, without ever reaching the target. □

Example 5.6. We now exhibit an example of an optimal synthesis for which it is not true that all the optimal trajectories are CLSS solutions of the optimal closed-loop equation.

Consider the optimal control problem in $\mathbb{R}$, with control space $U = \mathbb{R}$, dynamics given by $\dot{x} = u$, Lagrangian $L = (u^4 - x^2)^2$, and target $\{1\}$. The value function V is given by $V(x) \equiv 0$. The feedback control law v defined by

$$v(x) = \operatorname{sgn}(1 - x).\sqrt{|x|}$$

leads to trajectories $\gamma_x : [T_x, 0] \to \mathbb{R}$, corresponding to open-loop controls $\eta_x : [T_x, 0] \to \mathbb{R}$, given by the formulas

$$T_x = 2.\operatorname{sgn}(1 - x).\Big(\operatorname{sgn}(x).\sqrt{|x|} - 1\Big)\,,$$

$$\gamma_x(t) = \begin{cases} \dfrac{(2+t)^2}{4} & \text{if} \quad x \leq 1 \quad \text{and} \quad \max(T_x, -2) \leq t \leq 0\,, \\ -\dfrac{(2+t)^2}{4} & \text{if} \quad x \leq 1 \quad \text{and} \quad T_x \leq t \leq -2\,, \\ \dfrac{(2-t)^2}{4} & \text{if} \quad x \geq 1 \quad \text{and} \quad T_x \leq t \leq 0\,, \end{cases}$$

$$\eta_x(t) = \begin{cases} \dfrac{1}{2}|2 + t| & \text{if} \quad x \leq 1 \quad \text{and} \quad T_x \leq t \leq 0\,, \\ \dfrac{1}{2}(2 - t) & \text{if} \quad x \geq 1 \quad \text{and} \quad T_x \leq t \leq 0\,. \end{cases}$$

It is easy to see that the family $\Gamma = \{(\gamma_x, \eta_x)\}_{x\in\mathbb{R}}$ is a regular synthesis in the sense of our definition. Moreover, although the optimal trajectories are not unique, it is not hard to show that Γ *is the only optimal synthesis for our problem.* (Indeed, suppose that $\tilde{\Gamma} = \{(\tilde{\gamma}_x, \tilde{\eta}_x)\}_{x\in\mathbb{R}}$ was another optimal synthesis, and let $\operatorname{Dom}(\gamma_x) = [\tilde{T}_x, 0]$ for $x \in \mathbb{R}$. Let $x \in \mathbb{R}$. Then the cost of $(\tilde{\gamma}_x, \tilde{\eta}_x)$ must be zero, so $|\tilde{\eta}_x(t)| = \sqrt{|\tilde{\gamma}_x(t)|}$ for a.e. t. Suppose $x < 1$. Then $\tilde{\gamma}_x(t) < 1$ for all $t \in [\tilde{T}_x, 0\,[$, because if $\tilde{\gamma}_x(t) \geq 1$ for some $t \in [\tilde{T}_x, 0\,[$, then the map $\tilde{\gamma}_x$ would not be injective, and this would contradict the fact that Γ is a synthesis. If $\tilde{\eta}_x(t) < 0$ for all t in a subset E of $[\tilde{T}_x, 0]$ of positive measure, then once again the map $\tilde{\gamma}_x$ would not be injective. So $\tilde{\eta}_x(t) \geq 0$ a.e., and then $\tilde{\eta}_x(t) = v(\tilde{\gamma}_x(t))$ for a.e. t. A similar argument shows that $\tilde{\eta}_x(t) = v(\tilde{\gamma}_x(t))$ for a.e. t if $x > 1$. So the $\tilde{\gamma}_x$ are solutions of the equation $\dot{\gamma} = v(\gamma)$. This equation has unique solutions, except for the possibility that, starting at some $x < 0$, a trajectory reaching the origin may stay there for an arbitrary length of time before it resumes its motion to the right. This possibility, however, is clearly excluded for a trajectory of a synthesis. So, for each x, $\tilde{\gamma}_x$ is the unique solution of $\dot{\gamma} = v(\gamma)$ that starts at x, ends at 1 at time 0, and is one-to-one. Therefore $\tilde{\gamma}_x = \gamma_x$ for all x, so $\tilde{\Gamma} = \Gamma$.)

If we let $f(x,u) = u$, then we have established that the arcs γ_x are classical solutions of the closed-loop equation $\dot{x} = f(x, v(x))$. On the other hand, *the trajectory γ_0 corresponding to the initial condition $x = 0$ is not a CLSS solution of the closed-loop equation $\dot{x} = f(x, v(x))$* because, on an interval $[a,b]$, the only CLSS solution γ of this equation with initial condition $\gamma(a) = 0$ is $\gamma(t) \equiv 0$. $\square$

Appendix A. Differentiation of trajectories with respect to a parameter. The purpose of this appendix is to prove a general theorem on differentiation with respect to a parameter p, taking values in a normed space P, of a family of trajectories $x_p : [a,b] \to \mathbb{R}^n$ of time-varying vector fields f_p. We will state and prove the theorem under minimal hypotheses, much weaker than what is actually needed for the main results of this paper, because this result has other applications, such as the theory of envelopes (cf. Sussmann [29], [30]), where the more general statement is useful.

If $\dot{x}_p(t) = f_p(x_p(t), t)$, then formal differentiation with respect to p at $p = p_0$ in the direction of a vector v yields the variational equation

$$\text{(A.1)} \qquad \dot{y}_{p_0,v}(t) = \frac{\partial f_{p_0}}{\partial x}(x_{p_0}(t), t).y_{p_0,v}(t) + \lim_{\varepsilon \downarrow 0} \frac{W_{p_0+\varepsilon v}(t) - W_{p_0}(t)}{\varepsilon},$$

where $y_{p_0,v}(t)$ is the directional derivative of $x_p(t)$ at p_0 in the v-direction, and $W_p(t) = f_p(x_{p_0}(t), t)$.

The technical problem that will concern us here is to make (A.1) rigorous even when the limit in the right-hand side only exists in a weak sense.

For example, suppose that $n = 1$ and we are looking at the control system $\dot{x} = u$, and a family of "bang-bang" controls u_s depending on $s \in [0, \infty[$, such that $u_s(t) = 1$ for $t < s$, and $u_s(t) = -1$ for $t \geq s$. Let $x_s : [0, \infty[\to \mathbb{R}$ be the solution of $\dot{x} = u_s$, $x(0) = 0$. Then $x_s(t) = t$ for $t \leq s$, and $x_s(t) = 2s - t$ for $t \geq s$. Given any $T > 0$, the limit $\lim_{\varepsilon \to 0} \frac{u^T_{s+\varepsilon} - u^T_s}{\varepsilon}$—where u^T_s is the restriction of u_s to $[0,T]$—does not exist pointwise but exists in the weak* sense—regarding the u_s as members of the dual of the space of continuous functions on $[0,T]$—and equals $2\delta_s$, where δ_s is the Delta function at s. If we let $y_s(t)$ be the derivative of $x_s(t)$ with respect to s—so $y_s(t) = y_{s,1}(t)$—then y_s satisfies the differential equation $\dot{y}_s(t) = 2\delta_s(t)$, which makes better sense in integrated form: $y_s(t) = 2\chi_{[s,\infty[}(t)$. This says, in particular, that $y_s(T) = 2$ for $s < T$, and $y_s(T) = 0$ for $s > T$.

As this example shows, it can happen that the variational equation is a linear equation of the form $\dot{y}(t) = A(t)y(t) + w(t)$ where the "input" w is the formal derivative of a function W which is not necessarily absolutely continuous or even continuous. So our first task will be to study the solutions of linear systems with "generalized inputs" w.

Let $-\infty < a < b < \infty$, and let n be a positive integer. If $A \in L^1([a,b],\mathbb{R}^{n\times n})$, and $\bar{Y} \in \mathbb{R}^n$, then it is well known that the initial value problem

$$\dot{Y}(t) = A(t)\cdot Y(t) + w(t)\,, \qquad Y(a) = \bar{Y} \tag{A.2}$$

has a unique solution for every initial condition and every map $w \in L^1([a,b],\mathbb{R}^n)$. Since (A.2) can also be written formally as

$$dY(t) = A(t)\cdot Y(t)\,dt + w(t)\,dt\,, \qquad Y(a) = \bar{Y}, \tag{A.3}$$

we will use $Y_{A,w\,dt,\bar{Y}}$ to denote the unique solution of (A.2).

As explained before, we want to solve (A.2) for "inputs" w more general than integrable functions. The appropriate class of inputs turns out to be that of *formal derivatives of bounded measurable functions.* Precisely, we let $BM([a,b],\mathbb{R}^n)$ denote the space of all bounded Lebesgue measurable functions $W : [a,b] \to \mathbb{R}^n$, and use $BM_0([a,b],\mathbb{R}^n)$ to denote the set of those W in $BM([a,b],\mathbb{R}^n)$ such that $W(a) = 0$. (In $BM([a,b],\mathbb{R}^n)$ we do *not* identify two functions that are equal a.e.)

We endow $BM([a,b],\mathbb{R}^n)$ with the norm

$$||W||_{BM} \overset{\text{def}}{=} \sup\{||W(t)|| : t \in [a,b]\}. \tag{A.4}$$

Then $BM([a,b],\mathbb{R}^n)$ is a Banach space, and $BM_0([a,b],\mathbb{R}^n)$ is a closed subspace of $BM([a,b],\mathbb{R}^n)$. The space $C^0([a,b],\mathbb{R}^n)$ of continuous maps $W : [a,b] \to \mathbb{R}^n$ is of course a closed subspace of $BM([a,b],\mathbb{R}^n)$, but $L^\infty([a,b],\mathbb{R}^n)$ is *not* a subspace of $BM([a,b],\mathbb{R}^n)$, because in $L^\infty([a,b],\mathbb{R}^n)$ we identify functions that are equal a.e., but in $BM([a,b],\mathbb{R}^n)$ we do not.

If $W \in BM_0([a,b],\mathbb{R}^n)$ is absolutely continuous, and $w = \dot{W}$, so that, formally, $dW = w\,dt$, then we can rewrite (A.2) formally as

$$dY(t) = A(t)\cdot Y(t)\,dt + dW(t) \qquad Y(a) = \bar{Y}. \tag{A.5}$$

A solution of (A.5) is then a continuous map $t \to Y(t)$ such that

$$Y(t) = \bar{Y} + \int_a^t A(s)\cdot Y(s)\,ds + W(t) \qquad \text{for all} \quad t \in [a,b]. \tag{A.6}$$

Now (A.6) makes sense for arbitrary $Y \in BM([a,b],\mathbb{R}^n)$, $W \in BM_0([a,b],\mathbb{R}^n)$, even if W is not absolutely continuous. So we turn it into a definition: for a pair $(W,\bar{Y}) \in BM_0([a,b],\mathbb{R}^n)\times\mathbb{R}^n$, a *solution* of (A.5) is a map $Y \in BM([a,b],R^n)$ such that (A.6) holds. It then follows that

(EU) *for every $W \in BM_0([a,b],\mathbb{R}^n)$ and every $\bar{Y} \in \mathbb{R}^n$ there exists a unique solution Y of* (A.5).

Indeed, uniqueness is trivial, since the difference $Y = Y^1 - Y^2$ of two solutions is a solution of the homogeneous problem (i.e., of (A.2) with $w \equiv 0$ and $\bar{Y} = 0$), so $Y^1 - Y^2 \equiv 0$. To prove existence, we observe that (A.6) is equivalent to

(A.7)

$$Y(t) - W(t) = \bar{Y} + \int_a^t A(s)\cdot(Y(s) - W(s))\,ds + \int_a^t A(s)\cdot W(s)\,ds \qquad \text{for all} \quad t \in [a,b],$$

i.e., to $dZ = AZ\,dt + AW\,dt$, $Z(a) = \bar{Y}$, with $Z = Y - W$. Since $AW \in L^1$, because $A \in L^1$ and W is bounded and measurable, the equation $\dot{Z}(t) = A(t)\cdot Z(t) + A(t)W(t)$

has a unique solution Z^* for which $Z^*(a) = \bar{Y}$. If we let $Y(t) = Z^*(t) + W(t)$, and recall that $W(a) = 0$, we see that Y is the desired solution of (A.5).

We will use $Y_{A,dW,\bar{Y}}$ to denote the solution of (A.5), whose existence and uniqueness has just been established. In addition to proving (EU), the preceding argument also gives an explicit formula for $Y_{A,dW,\bar{Y}}$, namely,

$$Y_{A,dW,\bar{Y}} = W + Y_{A,AW\,dt,\bar{Y}}. \tag{A.8}$$

Clearly, the function $Y_{A,AW\,dt,\bar{Y}}$ is absolutely continuous, so the regularity properties of $Y_{A,dW,\bar{Y}}$—modulo absolutely continuous functions—are exactly the same as those of W. (For example, Y is continuous, right-continuous, or left-continuous at a point $t \in [a,b]$ if and only if W is; Y is of bounded variation if and only if W is; Y is absolutely continuous if and only if W is.)

Gronwall's inequality, applied to (A.7), tells us that

$$||Y_{A,dW,\bar{Y}}(t) - W(t)|| \leq e^{||A||_{L^1}}(||\bar{Y}|| + ||A||_{L^1}||W||_{BM}). \tag{A.9}$$

Therefore

$$||Y_{A,dW,\bar{Y}}||_{BM} \leq ||W||_{BM} + e^{||A||_{L^1}}(||\bar{Y}|| + ||A||_{L^1}||W||_{BM}). \tag{A.10}$$

On the other hand, (A.6) also implies that

$$||W||_{BM} \leq ||\bar{Y}|| + ||Y_{A,dW,\bar{Y}}||_{BM}(1 + ||A||_{L^1}). \tag{A.11}$$

THEOREM A.1. *Let $\{W_j\}$, $\{\bar{Y}_j\}$, $\{A_j\}$ be sequences in $BM_0([a,b],\mathbb{R}^n)$, $\mathbb{R}^n$, $L^1([a,b],\mathbb{R}^{n\times n})$, respectively. Assume that $\{A_j\}$ is bounded in $L^1([a,b],\mathbb{R}^{n\times n})$. Let $Y_j = Y_{A_j,dW_j,\bar{Y}_j}$. Then we have the following:*

(A.1.1) *$\{Y_j\}$ is bounded in $BM([a,b],\mathbb{R}^n)$ if and only if $\{W_j\}$ is bounded in $BM([a,b],\mathbb{R}^n)$ and $\{\bar{Y}_j\}$ is bounded in $\mathbb{R}^n$;*

(A.1.2) *If $\{A_j\}$ converges in $L^1([a,b],\mathbb{R}^n)$ to a limit A, $\bar{Y} \in \mathbb{R}^n$, $\bar{Y}_j \to \bar{Y}$, $\{W_j\}$ is bounded in $BM([a,b],\mathbb{R}^n)$, $W \in BM_0([a,b],\mathbb{R}^n)$, and $Y = Y_{A,dW,\bar{Y}}$, then $Y_j \to Y$ a.e. if and only if $W_j \to W$ a.e. In that case,* (a) *$Y_j - W_j$ converges to $Y - W$ uniformly on $[a,b]$ and, in particular,* (b) *for every $t \in [a,b]$, $Y_j(t) \to Y(t)$ if and only if $W_j(t) \to W(t)$.*

Proof. Statement (A.1.1) follows from the bounds (A.10) and (A.11). To prove (A.1.2), we first observe that under the hypotheses of (A.1.2) we can apply (A.1.1) and conclude that $\{Y_j\}$ is bounded in $BM([a,b],\mathbb{R}^n)$.

Let $\tilde{Y}_j = Y_j - W_j = Y_{A_j,A_jW_j\,dt,\bar{Y}_j}$, $\tilde{Y} = Y - W = Y_{A,AW\,dt,\bar{Y}_j}$. Then

$$\tilde{Y}_j(t) = \bar{Y}_j + \int_a^t A_j(s)Y_j(s)\,ds = \bar{Y}_j + \int_a^t A_j(s)\tilde{Y}_j(s)\,ds + \int_a^t A_j(s)W_j(s)\,ds, \tag{A.12}$$

so

$$\tilde{Y}_j(t) - \tilde{Y}(t) = \int_a^t A(s)(\tilde{Y}_j(s) - \tilde{Y}(s))\,ds + R_j(t), \tag{A.13}$$

where

$$R_j(t) = \bar{Y}_j - \bar{Y} + \int_a^t (A_j(s) - A(s))Y_j(s)\,ds + \int_a^t A(s)(W_j(s) - W(s))\,ds \tag{A.14}$$

and also

$$\text{(A.15)}\quad \tilde{Y}_j(t)-\tilde{Y}(t)=\bar{Y}_j-\bar{Y}+\int_a^t (A_j(s)-A(s))Y_j(s)\,ds+\int_a^t A(s)(Y_j(s)-Y(s))\,ds.$$

The assumption that the sequence $\{W_j\}$ is uniformly bounded and $\{\bar{Y}_j\}$ converges to $\bar{Y}$ implies that the sequences $\{\tilde{Y}_j\}$ and $\{Y_j\}$ are uniformly bounded as well. Since $A_j \to A$ in L^1, it is clear that the first two terms of the sum defining $R_j(t)$ go to zero uniformly as $j \to \infty$. Also,

$$\text{(A.16)}\qquad \left\|\int_a^t A(s)(W_j(s)-W(s))\,ds\right\| \le \int_a^b ||A(s)||\cdot||W_j(s)-W(s)||\,ds,$$

which goes to zero by the dominated convergence theorem if $s \to ||A(s)||$ is integrable and the sequence $\{W_j - W\}$ is uniformly bounded and goes to zero a.e. So, if we let $K_j = \sup\{||R_j(t)|| : a \le t \le b\}$, we see that $K_j \to 0$ as $j \to \infty$ if the assumptions of (A.1.2) hold and $W_j(t) \to W(t)$ for a.e. t. It then follows from Gronwall's inequality, applied to (A.13), that $||\tilde{Y}_j(t)-\tilde{Y}(t)|| \le K_j e^{||A||_{L^1}}$, so $\tilde{Y}_j - \tilde{Y} \to 0$ uniformly.

If $Y_j - Y \to 0$ a.e., then the first two terms of the right-hand side of (A.15) go to zero uniformly, and the third one is bounded by $\int_a^b ||A(s)||\cdot||Y_j(s)-Y(s)||\,ds$, which also goes to zero. So $\tilde{Y}_j - \tilde{Y} \to 0$ uniformly.

We have thus shown that $\tilde{Y}_j - \tilde{Y} \to 0$ uniformly if either $W_j \to W$ a.e. or $Y_j \to Y$ a.e. This clearly implies all the conclusions of (A.1.2). □

A very important subspace of $BM([a,b],\mathbb{R}^n)$ is $BV([a,b],\mathbb{R}^n)$, the space of all functions $W : [a,b] \to \mathbb{R}^n$ such that (a) W is of bounded variation, (b) W is right-continuous at every point of $]a,b]$, and (c) $W(a) = 0$. There is a canonical correspondence between $BV([a,b],\mathbb{R}^n)$ and the dual space $C^0([a,b],\mathbb{R}_n)^*$ of the Banach space $C^0([a,b],\mathbb{R}_n)$ of continuous functions $[a,b] \to \mathbb{R}_n$. (We write R_n rather than R^n because we want to think of the members of $C^0([a,b],\mathbb{R}_n)$ as *row-vector-valued functions*; cf. below.)

The members of $C^0([a,b],\mathbb{R}_n)^*$ are the $\mathbb{R}^n$*-valued Borel measures on* $[a,b]$. When $n = 1$, they are the *finite signed Borel measures on* $[a,b]$. The identification map from $BV([a,b],\mathbb{R}^n)$ to $C^0([a,b],\mathbb{R}_n)^*$ is the one that assigns to a function $W \in BV([a,b],\mathbb{R}^n)$ the unique Borel measure μ_W such that $\mu_W([a,t]) = W(t)$ for $a < t \le b$.

From now on we identify a function $W \in BV([a,b],\mathbb{R}^n)$ with its corresponding measure μ_W, and write $\int \alpha \cdot dW$ for $\int \alpha \cdot d\mu_W$. Clearly, $\int \alpha \cdot dW$ is defined, more generally, for an arbitrary bounded Borel measurable $\mathbb{R}_n$-valued map α on $[a,b]$.

Naturally, the integral $\int \alpha\, dW$ of a scalar bounded Borel measurable function $\alpha : [a,b] \to \mathbb{R}$ is also well defined, and is a vector in $\mathbb{R}^n$.

On $BV([a,b],R^n)$ there are at least three important topologies: (a) the one induced by the norm $||\cdots||_{BM}$, (b) the strong topology of $BV([a,b],\mathbb{R}^n)$ as the dual of the Banach space $C^0([a,b],\mathbb{R}_n)$, (c) the weak* topology, also corresponding to the duality with $C^0([a,b],\mathbb{R}_n)$. The second one is induced by the *total variation norm*:

$$\text{(A.17)}\qquad ||W||_{TV} = \sup\left\{\left|\int \alpha\cdot dW\right| : \alpha \in C^0([a,b],\mathbb{R}_n)\,,\ ||\alpha||_{BM} \le 1\right\}.$$

It is clear that $||W||_{BM} \le ||W||_{TV}$ for all W in $BV([a,b],\mathbb{R}^n)$. Therefore, *if a subset* S *of* $BV([a,b],\mathbb{R}^n)$ *is bounded in* TV *norm, then* S *is also bounded in* BM *norm.*

We will need the following.

LEMMA A.2. *If a sequence $\{W_j\}$ in $BV([a,b],\mathbb{R}^n)$ weak*-converges to a $W \in BV([a,b],\mathbb{R}^n)$, then* (a) *$\{W_j\}$ is bounded in TV norm,* (b) *$W_j(a) \to W(a)$ and $W_j(b) \to W(b)$,* (c) *every subsequence of $\{W_j\}$ has a subsequence that converges to W at all but at most countably many points of $[a,b]$,* (d) *$\int_a^b \beta(s)||W_j(t) - W(t)||\,dt \to 0$ for every function $\beta \in L^1([a,b],\mathbb{R})$.*

Proof. Statement (a) follows from the uniform boundedness theorem, and statement (b) is trivial, since $W_j(a) = 0 = W(a)$, $W_j(b) = \int 1\,dW_j$, and $W(b) = \int 1\,dW$. Statement (c) for $n = 1$ is a consequence of Helly's theorem, and the statement for arbitrary n then follows easily from the case $n = 1$. Statement (d) then clearly follows from (c), using the dominated convergence theorem together with the fact that boundedness in TV norm implies boundedness in BM norm. □

As a corollary of Lemma A.2, we get the integration by parts formula below.

LEMMA A.3. *If $\lambda : [a,b] \to \mathbb{R}_n$ is absolutely continuous, and $W \in BV([a,b],\mathbb{R}^n)$, then*

$$\lambda(b) \cdot W(b) = \int \lambda \cdot dW + \int_a^b \dot{\lambda}(t) \cdot W(t)\,dt. \tag{A.18}$$

Proof. The result is obviously true, because of the standard product rule for derivatives, if W is also absolutely continuous. In the general case, we use the fact that an arbitrary $W \in BV([a,b],\mathbb{R}^n)$ is the weak*-limit of a sequence $\{W_j\}$ of absolutely continuous functions. It then follows that (i) $\lambda(b) \cdot W_j(b) \to \lambda(b) \cdot W(b)$, because $W_j(b) \to W(b)$, (ii) $\int \lambda \cdot dW_j \to \int \lambda \cdot dW$, because $\{W_j\}$ weak*-converges to W, and λ is continuous, and (iii) $\int_a^b \dot{\lambda}(t) \cdot W_j(t)\,dt \to \int_a^b \dot{\lambda}(t) \cdot W(t)\,dt$, because part (d) of Lemma A.2 implies that $\int_a^b ||\dot{\lambda}(t)||.||W_j(t) - W(t)||\,dt \to 0$, since $\dot{\lambda} \in L^1$. □

An important consequence of Lemma A.3 is the following lemma.

LEMMA A.4. *Let $W \in BV([a,b],\mathbb{R}^n)$, $A \in L^1([a,b],\mathbb{R}^{n\times n})$, $\bar{Y} \in \mathbb{R}^n$, and write $Y = Y_{A,dW,\bar{Y}}$. Suppose $\lambda : [a,b] \to \mathbb{R}_n$ is absolutely continuous and satisfies $\dot{\lambda}(t) = -\lambda(t)A(t)$ for almost all t. Then*

$$\lambda(b) \cdot Y(b) - \lambda(a) \cdot Y(a) = \int \lambda \cdot dW. \tag{A.19}$$

Proof. Write $Y(t) = V(t) + W(t)$, where $V(t) = \bar{Y} + \int_a^t A(s).Y(s)\,ds$. Then

$$\lambda(b) \cdot Y(b) - \lambda(a) \cdot Y(a) = \lambda(b) \cdot V(b) - \lambda(a) \cdot V(a) + \lambda(b) \cdot W(b). \tag{A.20}$$

The function $t \to \lambda(t) \cdot V(t)$ is absolutely continuous and its derivative is $\dot{\lambda}(t) \cdot V(t) + \lambda(t) \cdot \dot{V}(t)$, which equals $-\lambda(t)A(t)Y(t) + \lambda(t)A(t)Y(t) - \dot{\lambda}(t) \cdot W(t)$, i.e., $-\dot{\lambda}(t) \cdot W(t)$. So

$$\lambda(b) \cdot Y(b) - \lambda(a) \cdot Y(a) = \lambda(b) \cdot W(b) - \int_a^b \dot{\lambda}(t) \cdot W(t)\,dt. \tag{A.21}$$

Equation (A.19) follows from this and the integration by parts formula (A.18). □

It is clear that, if $A \in L^1([a,b],\mathbb{R}^{n\times n})$, $\bar{Y} \in \mathbb{R}^n$, and $W \in BM_0([a,b],\mathbb{R}^n)$, then $Y_{A,dW,\bar{Y}} - W$ is absolutely continuous, so $W \in BV([a,b],\mathbb{R}^n)$ if and only if $Y_{A,dW,\bar{Y}}$ is of bounded variation, which happens if and only if $Y_{A,dW,\bar{Y}} - \bar{Y} \in BV([a,b],\mathbb{R}^n)$.

THEOREM A.5. *Let $\{W_j\}$, $\{\bar{Y}_j\}$, $\{A_j\}$ be sequences in $BV([a,b],\mathbb{R}^n)$, $\mathbb{R}^n$, $L^1([a,b],\mathbb{R}^{n\times n})$, respectively. Assume that $\{A_j\}$ is bounded in $L^1([a,b],\mathbb{R}^{n\times n})$ and $\{\bar{Y}_j\}$ is bounded in $\mathbb{R}^n$. Let $Y_j = Y_{A_j,dW_j,\bar{Y}_j}$. Then*

(A.5.1) *$\{Y_j - \bar{Y}_j\}$ is bounded in $BV([a,b],\mathbb{R}^n)$ if and only if $\{W_j\}$ is bounded in $BV([a,b],\mathbb{R}^n)$.*

(A.5.2) *If $\{A_j\}$ converges in $L^1([a,b],\mathbb{R}^n)$ to a limit A, $\bar{Y}_j \to \bar{Y}$, $W \in BV([a,b],\mathbb{R}^n)$, and we write $Y = Y_{A,dW,\bar{Y}}$, then $\{W_j\}$ weak*-converges to W if and only if $\{Y_j - \bar{Y}_j\}$ weak*-converges to $Y - \bar{Y}$. In that case, $\{Y_j - W_j - \bar{Y}_j\}$ converges to $Y - W - \bar{Y}$ strongly in $BV([a,b],\mathbb{R}^n)$. In particular,* (a) *$W_j(b) \to W(b)$,* (b) *$Y_j(b) \to Y(b)$, and* (c) *for every $t \in [a,b]$, $Y_j(t) \to Y(t)$ if and only if $W_j(t) \to W(t)$.*

Proof. We know from Theorem A.1 that $\{Y_j - \bar{Y}_j\}$ is bounded in BM norm if and only if $\{W_j\}$ is. Since boundedness in TV norm implies boundedness in BM norm, we see that boundedness of one of the two sequences in TV norm implies that there is a constant C such that $||Y_j(t)|| \le C$ for all j, t, and then the integral $I_j = \int_a^b ||A_j(t)Y_j(t)||\,dt$ is bounded by a fixed constant K. But $I_j = ||Y_j - \bar{Y}_j - W_j||_{TV}$. So boundedness in TV norm of one of the sequences $\{W_j\}$, $\{Y_j\}$ implies boundedness in TV norm of the other one.

If either $\{W_j\}$ weak*-converges to W, or $\{Y_j - \bar{Y}_j\}$ weak*-converges to $Y - \bar{Y}$, then the uniform boundedness theorem implies that either $\{W_j\}$ or $\{Y_j - \bar{Y}_j\}$ is bounded in TV norm, so both sequences are bounded in TV norm—and hence also in BM norm—by a constant C. Writing $Z_j = Y_j - \bar{Y}_j - W_j$, $Z = Y - \bar{Y} - W$, we then have $Z_j - Z = R_j^1 + R_j^2$, where

$$R_j^1(t) = \int_a^t (A_j(s) - A(s))Y_j(s)\,ds, \qquad R_j^2(t) = \int_a^t A(s)(Y_j(s) - Y(s))\,ds. \tag{A.22}$$

Then $||R_j^1||_{TV} \le C||A_j - A||_{L^1}$, so $||R_j^1||_{TV} \to 0$. Also,

$$||R_j^2||_{TV} \le ||A||_{L^1}.||\bar{Y}_j - \bar{Y}|| + \int_a^b ||A(t)||.||Y_j(t) - \bar{Y}_j - (Y(t) - \bar{Y})||\,dt. \tag{A.23}$$

If $\{Y_j - \bar{Y}_j\}$ weak*-converges to $Y - \bar{Y}$, then Lemma A.2 implies that $||R_j^2||_{TV} \to 0$, so $Z_j \to Z$ in $BV([a,b],\mathbb{R}^n)$.

If $\{W_j\}$ weak*-converges to W, then we rewrite R_j^2 as

$$R_j^2(t) = \int_a^t A(s)(Z_j(s) - Z(s))\,ds + R_j^3(t) + R_j^4(t), \tag{A.24}$$

where $R_j^3(t) = \int_a^t A(s)(\bar{Y}_j - \bar{Y})\,ds$ and $R_j^4(t) = \int_a^t A(s)(W_j(s) - W(s))\,ds$. We then have $||R_j^3||_{TV} \le ||A||_{L^1}.||\bar{Y}_j - \bar{Y}||$ and $||R_j^4||_{TV} \le \int_a^b ||A(s)||.||W_j(s) - W(s)||\,ds$. Lemma A.2 then implies that $||R_j^4||_{TV} \to 0$. So $Z_j(t) - Z(t) = \int_a^t A(s).(Z_j(s) - Z(s))\,ds + R_j(t)$, where $R_j = R_j^1 + R_j^3 + R_j^4$, so $||R_j||_{TV} \to 0$. Then Gronwall's inequality implies that $||Z_j - Z||_{BM} \le e^{||A||_{L^1}}||R_j||_{TV}$. If $Q_j(t) = \int_a^t A(s).(Z_j(s) - Z(s))\,ds$, then $||Q_j||_{TV} \le ||A||_{L^1}||Z_j - Z||_{BM}$, so $||Q_j||_{TV} \to 0$. Since $Z_j - Z = Q_j + R_j$, we see that $||Z_j - Z||_{TV} \to 0$.

So we have shown that if one of the two sequences $\{W_j - W\}$, $\{Y_j - \bar{Y}_j - (Y - \bar{Y})\}$ weak*-converges to 0, then $Y_j - \bar{Y}_j - W_j \to Y - \bar{Y} - W$ strongly in $BV([a,b],\mathbb{R}^n)$. This implies all the conclusions of (A.5.2). □

Recall that, if P, Q are normed spaces, and $F : \Omega \to Q$ is a map defined on a neighborhood Ω of a point p of P, then F is said to be *Fréchet differentiable* at p if there exists a bounded linear map L—called the *Fréchet differential* of F at p—such that $F(p') = F(p) + L(p' - p) + o(||p' - p||)$ as $p' \to p$. If X is a Banach space and $Q = X^*$, then we call F *weak* Fréchet differentiable* at p if for every $x \in X$ the map $\Omega \ni p' \to \langle F(p'), x\rangle \in \mathbb{R}$ is Fréchet differentiable at p. (We use $\langle x^*, x\rangle$, for $x \in X$, $x^* \in X^*$, as an alternative notation for $x^*(x)$.)

Suppose P is a normed space, X is a Banach space, $p \in P$, $F : \Omega \to X^*$ is a map defined on a neighborhood Ω of p, and F is weak* Fréchet differentiable at p. Then for each $x \in X$ there is a bounded linear functional θ_x on P such that $\langle F(p+v), x\rangle = \langle F(p), x\rangle + \theta_x(v) + o(||v||)$ as $v \to 0$ in P. Clearly, θ_x is unique and is given by the formula

$$\theta_x(v) = \lim_{h\to 0} h^{-1}\langle F(p+hv) - F(p), x\rangle. \tag{A.25}$$

For any given $v \in P$, use θ^v to denote the map $x \to \theta_x(v)$. Then θ^v is a limit of linear maps, so it is linear. If $\{h_j\}$ is any sequence going to 0, then the sequence of linear functionals ν_j on X given by $\nu_j(x) = h_j^{-1}\langle F(p + h_j v) - F(p), x\rangle$ converges pointwise to θ^v. Since each ν_j is a bounded linear functional on X, it follows for the uniform boundedness theorem that θ^v is bounded. Since $\theta^v(x) = \theta_x(v)$, which is linear with respect to v, we can conclude that $v \to \theta^v$ is a linear map from P to X^*. We use $D^wF(p)$ to denote this map, and $D^wF(p).v$ to denote its value for a particular v, so $D^wF(p).v = \theta^v$, and

$$\langle D^wF(p).v, x\rangle = \lim_{h\to 0} h^{-1}\langle F(p+hv) - F(p), x\rangle. \tag{A.26}$$

Lemma A.6. *Assume that P is a normed space, X is a Banach space, $p \in P$, $F : \Omega \to X^*$ is a map defined on a neighborhood Ω of p, and F is weak* Fréchet differentiable at p. Then*

(A.6.i) *there is a $C > 0$ such that*

$$||F(p+v) - F(p)|| \leq C||v|| \qquad \textit{for all sufficiently small} \qquad v \in P, \tag{A.27}$$

(A.6.ii) *the linear map $D^wF(p) : P \to X^*$ defined by* (A.26) *is bounded.*

Proof. Suppose (A.6.i) is not true. Then there exists a sequence $\{v_j\}$ in P, converging to 0 and such that $||F(p + v_j) - F(p)|| > j||v_j||$. Define, for each j, a member σ_j of X^* by

$$\sigma_j = \frac{F(p+v_j) - F(p) - \theta^{v_j}}{||v_j||}. \tag{A.28}$$

Then $\sigma_j(x) \to 0$ for each x, so the sequence $\{\sigma_j\}$ is pointwise bounded and hence uniformly bounded.

On the other hand, for each $x \in X$, the sequence $\{\frac{\theta^{v_j}(x)}{||v_j||}\}$ is bounded, because $\theta^{v_j}(x) = \theta_x(v_j)$, and θ_x is a bounded linear functional on P, so $|\theta^{v_j}(x)| \leq ||\theta_x||.||v_j||$. So, using the uniform boundedness theorem once again, we conclude that the sequence $\{\frac{\theta^{v_j}}{||v_j||}\}$ is bounded in X^*. Since $\{\sigma_j\}$ is bounded, we conclude that the sequence $\{\frac{F(p+v_j)-F(p)}{||v_j||}\}$ is bounded, contradicting the fact that $||F(p + v_j) - F(p)|| > j||v_j||$. So (A.27) holds for some $C > 0$, and we have proved (A.6.i).

It follows from (A.6.i) that $|\langle F(p + v) - F(p), x\rangle| \leq C||x||.||v||$ for all $x \in X$ if $v \in P$ is small enough. Applying this with hv in the role of v and letting $h \downarrow 0$, we find that

$$|\langle D^wF(p).v, x\rangle| \leq C||x||.||v|| \qquad \text{for all} \qquad x \in X\,,\ \ v \in P. \tag{A.29}$$

Therefore $||D^wF(p).v|| \leq C||v||$, so $D^wF(p)$ is a bounded linear map from P to X^*. $\square$

We are now ready to study the differentiability properties with respect to a parameter p, belonging to a normed space P, of a family of trajectories $x_p : [a,b] \to \mathbb{R}^n$ of time-varying vector fields f_p. We will prove, under minimal technical hypotheses, that if the initial condition map $p \to x_p(a)$ is differentiable at $p=0$, and the map $p \to f_p$ satisfies a "weak differentiability condition" at the trajectory x_0, then the map $p \to x_p$ is weak*-differentiable at $p=0$, as a map into $BV([a,b],\mathbb{R}^n)$, and in particular, the endpoint map $p \to x_p(b)$ is differentiable at $p=0$.

A *curve in* $\mathbb{R}^n$ is a continuous map $\xi : I \to \mathbb{R}^n$, whose domain $I = \mathrm{Dom}(\xi)$ is a nonempty interval. A curve ξ is an *arc* if $\mathrm{Dom}(\xi)$ is compact. The *graph* $G(\xi)$ of a curve ξ is the set $G(\xi) = \{(\xi(t),t) : t \in \mathrm{Dom}(\xi)\}$. If $\xi : [a,b] \to \mathbb{R}^n$ is an arc in $\mathbb{R}^n$, and $\varepsilon > 0$, then the *ε-tube about* ξ is the set $\mathcal{T}(\xi,\varepsilon) = \{(x,t) : x \in \mathbb{R}^n, a \leq t \leq b, ||x-\xi(t)|| \leq \varepsilon\}$.

A *time-varying vector field* on $\mathbb{R}^n$ is an $\mathbb{R}^n$-valued map f whose domain $\mathrm{Dom}(f)$ is a—possibly empty—subset S of the product $\mathbb{R}^n \times \mathbb{R}$. We use $TVVF(\mathbb{R}^n)$ to denote the set of all time-varying vector fields on $\mathbb{R}^n$. A *trajectory*, or *integral curve*, of an $f \in TVVF(\mathbb{R}^n)$ is a locally absolutely continuous curve ξ in $\mathbb{R}^n$ such that $G(\xi) \subseteq \mathrm{Dom}(f)$ and $\dot{\xi}(t) = f(\xi(t),t)$ for almost all $t \in \mathrm{Dom}(\xi)$. We use $\mathrm{Traj}(f)$ to denote the set of all trajectories of f, and $\mathrm{Traj}_c(f)$ to denote the set of all $\xi \in \mathrm{Traj}(f)$ such that $\mathrm{Dom}(\xi)$ is compact.

If $f \in TVVF(\mathbb{R}^n)$ and $\xi \in \mathrm{Traj}(f)$, we say that ξ is a *maximal* trajectory of f if it cannot be extended to a trajectory $\tilde{\xi} : \tilde{I} \to \mathbb{R}^n$ of f defined on an interval $\tilde{I}$ such that $I \subseteq \tilde{I}$ but $I \neq \tilde{I}$. We use $\mathrm{MTraj}(f)$ to denote the set of all maximal trajectories of f. Given $\bar{x} \in \mathbb{R}^n$, $\bar{t} \in \mathbb{R}$, we use $\mathrm{Traj}(f,\bar{x},\bar{t})$, $\mathrm{MTraj}(f,\bar{x},\bar{t})$ to denote, respectively, the set of all $\xi \in \mathrm{Traj}(f)$ such that $\xi(\bar{t}) = \bar{x}$, and the set $\mathrm{Traj}(f,\bar{x},\bar{t}) \cap \mathrm{MTraj}(f)$. Zorn's lemma implies that every $\xi \in \mathrm{Traj}(f,\bar{x},\bar{t})$ can be extended to a $\tilde{\xi} \in \mathrm{MTraj}(f,\bar{x},\bar{t})$, so $\mathrm{Traj}(f,\bar{x},\bar{t}) \neq \emptyset$ if and only if $\mathrm{MTraj}(f,\bar{x},\bar{t}) \neq \emptyset$. Clearly, $\mathrm{Traj}(f,\bar{x},\bar{t}) \neq \emptyset$ if and only if $(\bar{x},\bar{t}) \in \mathrm{Dom}(f)$. (Indeed, if $(\bar{x},\bar{t}) \in \mathrm{Dom}(f)$, then the map ξ with domain $\{\bar{t}\}$ such that $\xi(\bar{t}) = \bar{x}$ is in $\mathrm{Traj}(f)$.)

We say that an $f \in TVVF(\mathbb{R}^n)$ has the *forward existence property* at a point $(\bar{x},\bar{t})$ if for some $\varepsilon > 0$ there exists a $\xi \in \mathrm{Traj}(f,\bar{x},\bar{t})$ with domain $[\bar{t},\bar{t}+\varepsilon]$. We say that f has the *forward limit property* on a subset S of $\mathrm{Dom}(f)$ if, whenever $\xi : [a,b[\to \mathbb{R}^n$ is a trajectory of f which is contained in a compact subset of S, it follows that $\lim_{t\uparrow b}\xi(t)$ exists.

Suppose we are given the following data:

(D.1) a normed space P,

(D.2) a family $\mathbf{f} = \{f_p\}_{p\in P}$ of time-varying vector fields on $\mathbb{R}^n$, depending on a parameter $p \in P$,

(D.3) a $p_0 \in P$,

(D.4) a trajectory $\xi : [a,b] \to \mathbb{R}^n$ of f_{p_0}.

We then define functions $w_p : [a,b] \to \mathbb{R}^n$ by

$$w_p(t) = f_p(\xi(t),t) - f_{p_0}(\xi(t),t) \quad \text{for} \quad t \in [a,b], \tag{A.30}$$

so w_p is defined on $[a,b]$ whenever $G(\xi) \subseteq \mathrm{Dom}(f_p)$. If w_p is integrable, we let

$$W_p(t) = \int_a^t w_p(s)\,ds \qquad \text{for} \qquad t \in [a,b]. \tag{A.31}$$

We then let $\mathcal{W}$ be the map that sends p to $\mathcal{W}(p) = W_p$, with domain

$$\mathrm{Dom}(\mathcal{W}) = \{p \in P : G(\xi) \subseteq \mathrm{Dom}(f_p),\ w_p \in L^1([a,b],\mathbb{R}^n)\}. \tag{A.32}$$

We regard $\mathcal{W}$ as a map from $\mathrm{Dom}(\mathcal{W})$ to $BV([a,b],\mathbb{R}^n)$.

We then say that $\mathbf{f}$ is *weakly differentiable at* p_0 *along* ξ if there exists $\bar{\varepsilon}>0$ such that

(WD1) $\mathcal{T}(\xi,\bar{\varepsilon})\subseteq \mathrm{Dom}(f_p)$ for every p such that $||p-p_0||\le\bar{\varepsilon}$;

(WD2) for every p such that $||p-p_0||\le\bar{\varepsilon}$, f_p has the forward limit property on $\mathcal{T}(\xi,\bar{\varepsilon})$ and the forward existence property at every (x,t) such that $a\le t<b$ and $||x-\xi(t)||<\bar{\varepsilon}$;

(WD3) there exist $A\in L^1([a,b],\mathbb{R}^{n\times n})$ and functions $\psi_\varepsilon\in L^1([a,b],\mathbb{R})$, for $\varepsilon\in]0,\bar{\varepsilon}]$, such that

$$||f_p(x,t)-f_p(\xi(t),t)-A(t).(x-\xi(t))||\le\psi_\varepsilon(t)(||x-\xi(t)||+||p-p_0||) \tag{A.33}$$

whenever $a\le t\le b$, $||x-\xi(t)||\le\varepsilon$, $||p-p_0||\le\varepsilon$, and

$$\lim_{\varepsilon\downarrow 0}\int_a^b\psi_\varepsilon(t)\,dt=0; \tag{A.34}$$

(WD4) the function w_p defined by (A.30) is integrable for every p such that $||p-p_0||\le\bar{\varepsilon}$ and, if $\mathcal{W}$ is defined as above, then $\mathcal{W}$ is weak* Fréchet differentiable at p_0.

The second part of condition (WD4) says that for each $\alpha\in C^0([a,b],\mathbb{R}_n)$ the map $p\to\int\alpha\cdot dW_p$ is Fréchet differentiable at p_0. Lemma A.6 then implies—since $W_{p_0}=0$—that there exists a bounded linear map $D^w\mathcal{W}(p_0):P\to BV([a,b],\mathbb{R}^n)$ having the property that

(A.35)
$$\begin{aligned}\int\alpha\cdot dW_p&=\langle\mathcal{W}(p)-\mathcal{W}(p_0),\alpha\rangle\\&=\int\alpha\cdot d\Big(D^w\mathcal{W}(p_0).(p-p_0)\Big)+o(||p-p_0||)\qquad\text{as}\qquad p\to p_0\end{aligned}$$

for each $\alpha\in C^0([a,b],\mathbb{R}_n)$, and also that there is a constant C such that

(A.36) $||W_p||_{TV}\le C||p-p_0||$ for all $p\in P$ such that $\|p-p_0\|$ is sufficiently small.

The following result is the main theorem on differentiation of trajectories with respect to a parameter.

THEOREM A.7. *Let* P, $\mathbf{f}$, p_0, ξ *be data as in* (D.1,2,3,4). *Assume that* $\mathbf{f}$ *is weakly differentiable at* p_0 *along* ξ. *Let* $\bar{\varepsilon}>0$ *be such that* (WD1) *to* (WD4) *hold. Then the following hold:*

(A.7.a) *For every* $\varepsilon\in]\,0,\bar{\varepsilon}]$ *there exists a* $\delta\in]\,0,\varepsilon[$ *such that, whenever* $||p-p_0||\le\delta$, $||\bar{x}-\xi(a)||\le\delta$, *and* $\zeta\in\mathrm{MTraj}(f_p,\bar{x},a)$, *it follows that* $[a,b]\subseteq\mathrm{Dom}(\zeta)$ *and* $||\zeta(t)-\xi(t)||\le\varepsilon$ *for all* $t\in[a,b]$.

(A.7.b) *If* $\{\bar{x}_p\}_{||p-p_0||\le\bar{\varepsilon}}$ *is a family of points of* $\mathbb{R}^n$ *such that the map* $p\to\bar{x}_p$ *is Fréchet differentiable at* p_0 *and* $\bar{x}_{p_0}=\xi(a)$, *then, if* $\{\xi_p\}_{||p-p_0||\le\bar{\varepsilon}}$ *is a family of curves such that* $\xi_p\in\mathrm{MTraj}(f_p,\bar{x}_p,a)$, *and* x_p *denotes the restriction of* ξ_p *to* $[a,b]$—*which is well defined for small enough* $||p-p_0||$ *by* (A.7.a)—*then the map* $\mathcal{X}$ *that sends each* p *to the function* x_p *is weak* Fréchet differentiable at* p_0 *as a map from* P *to* $BV([a,b],\mathbb{R}^n)$. *Moreover,* $D^w\mathcal{X}(p_0)$ *can be computed as follows. For* $v\in P$,

$$D^w\mathcal{X}(p_0).v=Y_{A,\,d(D^w\mathcal{W}(0).v),\,L.v}, \tag{A.37}$$

where $L : P \to \mathbb{R}^n$ is the differential at p_0 of the map $p \to \bar{x}_p$.

Proof. We will assume, as we clearly may without loss of generality, that $p_0 = 0$. Also, we know that there is a $C > 0$ such that the bound (A.36) holds. So we can assume, by making $\bar{\varepsilon}$ smaller, if necessary, that $||W_p||_{TV} \leq C||p||$ whenever $||p|| \leq \bar{\varepsilon}$. By making C larger and $\bar{\varepsilon}$ smaller, if necessary, we may also assume that $C \geq 1$, and

$$8e^{||A||_{L^1}}||\psi_\varepsilon||_{L^1} < 1 \qquad \text{for} \qquad \varepsilon \in]0, \bar{\varepsilon}]. \tag{A.38}$$

Pick $\varepsilon \in]0, \bar{\varepsilon}[$, and choose $\delta > 0$ such that

$$C\delta < \varepsilon. \tag{A.39}$$

Notice that $\delta < \varepsilon$, since $C \geq 1$.

Let p, $\bar{x}$ be such that $||p|| \leq \delta$ and $||\bar{x} - \xi(a)|| \leq \delta$. Let $\zeta \in \mathrm{MTraj}(f_p, \bar{x}, a)$, and let $I = \mathrm{Dom}(\zeta)$. Then $a \in I$. Let c be the supremum of those t such that $a \leq t \leq b$ and $||\zeta(s) - \xi(s)|| \leq \varepsilon$ for $a \leq s \leq t$. Then $a < c \leq b$, $c \in I$, and either $c = b$ or $||\zeta(c) - \xi(c)|| = \varepsilon$. (The fact that $a < c$ follows from the local existence property, since $||\bar{x} - \xi(a)|| \leq \delta < \varepsilon < \bar{\varepsilon}$, and ζ is continuous. The fact that $c \in I$ follows from the forward limit property, since the graph of the restriction of ζ to $[a, c[$ is contained in $\mathcal{T}(\xi, \varepsilon)$. If $c < b$ and $||\zeta(c) - \xi(c)|| < \varepsilon$, then the forward existence property implies that ζ is not maximal.)

We will estimate $||\zeta(t) - \xi(t)||$ for $t \in [a, c]$, and show that—if ε is small enough, and δ is suitably chosen—$||\zeta(t) - \xi(t)|| < \varepsilon$ for all such t, including $t = c$. This will imply that $c = b$.

From now on, we work on the interval $[a, c]$. We have

$$\zeta(t) - \xi(t) = \bar{x} - \xi(a) + \int_a^t (f_p(\zeta(s), s) - f_0(\xi(s), s))\, ds, \tag{A.40}$$

so

$$\zeta(t) - \xi(t) = \bar{x} - \xi(a) + W_p(t) + \int_a^t (f_p(\zeta(s), s) - f_p(\xi(s), s))\, ds. \tag{A.41}$$

Let $r_p(t) = ||f_p(\zeta(t), t) - f_p(\xi(t), t) - A(t).(\zeta(t) - \xi(t))||$. Then

$$\zeta(t) - \xi(t) = \bar{x} - \xi(a) + W_p(t) + \int_a^t A(s).(\zeta(s) - \xi(s))\, ds + \int_a^t r_p(s)\, ds, \tag{A.42}$$

so

$$||\zeta(t) - \xi(t)|| \leq ||\bar{x} - \xi(a)|| + ||W_p||_{sup} + \int_a^t ||A(s)||.||\zeta(s) - \xi(s)||\, ds + \int_a^t r_p(s)\, ds, \tag{A.43}$$

and then

$$||\zeta(t) - \xi(t)|| \leq ||\bar{x} - \xi(a)|| + ||W_p||_{sup} + 2\varepsilon||\psi_\varepsilon||_{L^1} + \int_a^t ||A(s)||.||\zeta(s) - \xi(s)||\, ds, \tag{A.44}$$

because $r_p(s) \leq 2\varepsilon\psi_\varepsilon(s)$ for $s \in [a, c]$, since $||\zeta(s) - \xi(s)|| \leq \varepsilon$ and $||p|| \leq \varepsilon$.

Since $||\bar{x} - \xi(a)|| \leq \delta$, and $||W_p||_{sup} \leq ||W||_{TV} \leq C||p|| \leq C\delta$, Gronwall's inequality implies

$$||\zeta(t) - \xi(t)|| \leq e^{||A||_{L^1}}((C + 1)\delta + 2\varepsilon||\psi_\varepsilon||_{L^1}). \tag{A.45}$$

In view of (A.38), we have

$$2\varepsilon e^{||A||_{L^1}}||\psi_\varepsilon||_{L^1} < \frac{\varepsilon}{4}. \tag{A.46}$$

Now suppose that

$$\delta \le \kappa\varepsilon, \qquad \text{where} \qquad \kappa = (4C+4)^{-1}e^{-||A||_{L^1}}. \tag{A.47}$$

Then

$$||\zeta(t) - \xi(t)|| \le \frac{\varepsilon}{2} \qquad \text{for all} \qquad t \in [a,c]. \tag{A.48}$$

This is our desired estimate, from which it follows, as explained above, that $c = b$, and then $||\zeta(t) - \xi(t)|| \le \varepsilon$ for all $t \in [a,b]$. This completes the proof of (A.7.a).

Besides proving (A.7.a), we have also shown that, if κ is the constant given by (A.47), then the conclusion of (A.7.a) holds for every δ such that $0 < \delta \le \kappa\varepsilon$. (Notice that if $\delta \le \kappa\varepsilon$, then (A.39) holds.)

Now suppose we are given a map $P \ni p \to \bar{x}_p \in \mathbb{R}^n$ which is defined for $||p|| \le \bar{\delta}$ for some $\bar{\delta} > 0$. Assume that this map is Fréchet differentiable at 0, with differential L, and is such that $\bar{x}_0 = \xi(a)$. Write $z_p = Y_{A,dV_p,L.p}$, where $V_p = D^w\mathcal{W}(0).p$.

It then follows that there exist a constant $C' \ge 1$ and a $\bar{\delta}' \in]\,0, \bar{\delta}]$ such that (i) $C'\bar{\delta}' \le \kappa\bar{\varepsilon}$ and (ii) $||\bar{x}_p - \xi(a)|| \le C'||p||$ for $||p|| \le \bar{\delta}'$.

Pick in an arbitrary fashion $\xi_p \in \mathrm{MTraj}(f_p, \bar{x}_p, a)$ for $||p|| \le \bar{\delta}'$. Then, if $||p|| \le \bar{\delta}'$, it follows that (i) ξ_p is defined, (ii) $\xi_p \in \mathrm{MTraj}(f_p, \bar{x}, a)$ for an $\bar{x}$—namely, $\bar{x}_p$—such that $||\bar{x} - \xi(a)|| \le C'\bar{\delta}' \le \kappa\bar{\varepsilon}$. Since $C' \ge 1$, we also have $||p|| \le \kappa\bar{\varepsilon}$. Therefore we can conclude—using $\varepsilon = \bar{\varepsilon}$ and $\delta = \kappa\bar{\varepsilon}$—that $[a,b] \subseteq \mathrm{Dom}(\xi_p)$ and $(\xi_p(t), t) \in \mathcal{T}(\xi, \bar{\varepsilon})$ for $t \in [a,b]$.

Moreover, using (A.33) with $p = p_0$ it is easy to see that $\xi_0(t) = \xi(t)$ for every $t \in [a,b]$, since both trajectories satisfy the same initial condition.

Let x_p be the restriction of ξ_p to $[a,b]$. For $||p|| \le \bar{\delta}'$, we have

(A.49)

$$\begin{aligned} x_p(t) - \xi(t) - z_p(t) &= \bar{x}_p - \xi(a) - L.p \\ &\quad + \int_a^t A(s).(x_p(s) - \xi(s) - z_p(s))\,ds + W_p(t) - V_p(t) + R_p(t), \end{aligned}$$

where $R_p(t) = \int_a^t \tilde{r}_p(s)\,ds$ and $\tilde{r}_p(s) = f_p(x_p(s), s) - f_p(\xi(s), s) - A(s).(x_p(s) - \xi(s))$. If $p \ne 0$, let

$$\begin{aligned} Y_p &= ||p||^{-1}(x_p - x_0 - z_p), \\ U_p &= ||p||^{-1}(W_p - V_p + R_p), \\ \bar{Y}_p &= ||p||^{-1}(\bar{x}_p - x_0(a) - L.p). \end{aligned} \tag{A.50}$$

Then $Y_p = Y_{A,dU_p,\bar{Y}_p}$.

We have to prove that Y_p weak*-converges to 0 as $p \to 0$. Since $\bar{Y}_p \to 0$, using Theorem A.5 our conclusion will follow if we show that U_p weak*-converges to 0. Since $V_p = D^w\mathcal{W}(0).p$, it is clear that $||p||^{-1}(W_p - V_p)$ weak*-converges to 0. So all we need is to prove that $||p||^{-1}R_p$ weak*-converges to 0. If $||p||$ is sufficiently small, then we can apply the previous estimates using $\delta = ||p||$, $\varepsilon = \frac{\delta}{\kappa}$. We conclude that $||r_p(s)|| \leq 2\varepsilon\psi_\varepsilon(s)$ for all $s \in [a, b]$. This implies that

$$||\tilde{r}_p||_{L^1} \leq \frac{2||p||}{\kappa}||\psi_{\kappa^{-1}||p||}||_{L^1}. \tag{A.51}$$

This shows that $||R_p||_{TV} = ||r_p||_{L^1} = o(||p||)$ as $p \to 0$. So $||p||^{-1}R_p$ converges to 0 strongly—and a fortiori weakly*—in $BV([a, b], \mathbb{R}^n)$. This completes our proof. □

If $f \in TVVF(\mathbb{R}^n)$ and $\xi : [a, b] \to \mathbb{R}^n$ is a trajectory of f, we say that f *is differentiable along* ξ if the map $x \to f(x, t)$ is differentiable at $\xi(t)$ for almost every t. In that case, we write $D_\xi f(t) \overset{\text{def}}{=} \frac{\partial f}{\partial x}(\xi(t), t)$, and refer to the map $[a, b] \ni t \to D_\xi f(t) \in \mathbb{R}^{n\times n}$ as the *differential of* f *along* ξ.

Lemma A.8. *If* $\mathbf{f}$ *is weakly differentiable at* p_0 *along a trajectory* $\xi : [a, b] \to \mathbb{R}^n$ *of* f_{p_0}, *then* f_{p_0} *is differentiable along* ξ *and the map* A *of* (WD3) *is such that* $A(t) = D_\xi f_{p_0}(t)$ *for almost all* $t \in [a, b]$.

Proof. We begin by showing that the functions ψ_ε for which (WD3) holds can be chosen so that $\psi_\varepsilon(t)$ is a monotonically nonincreasing function of ε for each t. To see this, start with some family $\{\psi_\varepsilon\}_{\varepsilon\in]0,\bar{\varepsilon}]}$ of functions for which (WD3) holds, and define $\hat{\psi}_\varepsilon = \psi_{2^{-k}\bar{\varepsilon}}$ whenever $2^{-k-1}\bar{\varepsilon} < \varepsilon \leq 2^{-k}\bar{\varepsilon}$ and k is a nonnegative integer. Then (WD3) also holds if the $\hat{\psi}_\varepsilon$ are substituted for the ψ_ε. Next define $\bar{\psi}_\varepsilon(t) = \min\{\hat{\psi}_{\varepsilon'}(t) : \varepsilon \leq \varepsilon' \leq \bar{\varepsilon}\}$. Then the $\bar{\psi}_\varepsilon$ are well-defined integrable functions of t, because for each ε the set of functions $\{\hat{\psi}_{\varepsilon'} : \varepsilon \leq \varepsilon' \leq \bar{\varepsilon}\}$ is finite. It is clear that (WD3) also holds with the $\bar{\psi}_\varepsilon$ instead of the $\hat{\psi}_\varepsilon$, and in addition $\hat{\psi}_\varepsilon(t)$ is nonincreasing as a function of ε.

Now, if we pick the ψ_ε to be monotonically nonincreasing with respect to ε, it follows that the limit $\rho(t) = \lim_{\varepsilon\downarrow 0}\psi_\varepsilon(t)$ exists for every t. By the dominated convergence theorem, $\int_a^b \rho(t)dt = 0$. Since $\rho(t) \geq 0$ for all t, it follows that $\rho(t) = 0$ a.e.

Let t be such that $\rho(t) = 0$. Let $\theta(x) = \psi_{||x-\xi(t)||}(t)$. Then $\lim_{x\to\xi(t)}\theta(x) = 0$. On the other hand, (WD3) implies that

$$||f_{p_0}(x, t) - f_{p_0}(\xi(t), t) - A(t).(x - \xi(t))|| \leq \theta(x)||x - \xi(t)||. \tag{A.52}$$

Since $\theta(x) \to 0$ as $x \to \xi(t)$, (A.52) says precisely that $A(t)$ is the differential at $\xi(t)$ of the map $x \to f_{p_0}(x, t)$, completing our proof. □

If $f \in TVVF(\mathbb{R}^n)$ is differentiable along a trajectory $\xi : [a, b] \to \mathbb{R}^n$ of f, and the matrix-valued function $D_\xi f$ is integrable, then the ordinary differential equation

$$\dot{\lambda}(t) = -\lambda(t).D_\xi f(t), \tag{A.53}$$

for an $\mathbb{R}_n$-valued function λ on $[a, b]$, is the *adjoint variational equation*—or, simply, the *adjoint equation*—for f along ξ. A solution of the adjoint equation (i.e., an absolutely continuous function $\lambda : [a, b] \to \mathbb{R}_n$ such that (A.53) holds for almost all t) is known as an *adjoint vector for* f *along* ξ.

Theorem A.9. *Let* P, $\mathbf{f} = \{f_p\}_{p\in P}$, $p_0 \in P$, $\xi : [a, b] \to \mathbb{R}^n$ *be as in Theorem* A.7. *Let* $P \ni p \to \bar{x}_p\imath\mathbb{R}^n$ *be a map which is Fréchet differentiable at* p_0 *and such that* $\bar{x}_{p_0} = \xi(a)$, *and let* L_a *be its differential at* p_0. *Let* ξ_p *be, for* p *close enough to* p_0, *maximal trajectories of* f_p *such that* $\xi_p(a) = \bar{x}_p$, *and let* x_p *be the restriction of* ξ_p *to* $[a, b]$. *Then the map* $p \to x_p(b)$ *is differentiable at* p_0. *Let* L_b *be its differential at* p_0. *Let* $\lambda : [a, b] \to \mathbb{R}^n$ *be an adjoint vector for* f_{p_0} *along* ξ. *Then*

$$\lambda(b) \cdot L_b(v) - \lambda(a) \cdot L_a(v) = \lim_{\varepsilon\downarrow 0}\varepsilon^{-1}\int_a^b \langle\lambda(t), f_{p_0+\varepsilon v}(\xi(t), t) - f_{p_0}(\xi(t), t)\rangle dt \tag{A.54}$$

for every $v \in \mathbb{R}^n$.

Proof. Assume, without loss of generality, that $p_0 = 0$. According to Theorem A.7 and Lemma A.8, the map $p \to x_p$ is weak* Fréchet differentiable at 0—as a map from P to $BV([a,b],\mathbb{R}^n)$—and its differential at 0 is the map that sends $v \in P$ to $Y_{A,d(U.v),L_a(v)}$, where $U = D^w\mathcal{W}(0)$, and $\mathcal{W}$ is the map $p \to W_p$, where $W_p(t) = \int_a^t (f_p(\xi(s),s) - f_0(\xi(s),s))dt$, and $A = D_\xi f$. Then $p \to x_p(b)$ is Fréchet differentiable at 0, and the differential L_b is given by $L_b(v) = Y_{A,d(U.v),L_a(v)}(b)$. Clearly, $L_a(v) = Y_{A,d(U.v),L_a(v)}(a)$. Since λ satisfies $\dot\lambda = -\lambda.A$, Lemma A.4 tells us that

$$\lambda(b)\cdot L_b(v) - \lambda(a)\cdot L_a(v) = \int_a^b \lambda.d(U.v).$$

Since $U.v$ is the weak*-limit of $\varepsilon^{-1}W_{\varepsilon v}$ as $\varepsilon \downarrow 0$, our conclusion follows. □

We now turn to the differentiation of a Lagrangian cost functional.

If $f \in TVVF(\mathbb{R}^n)$, a *Lagrangian* for f in $\mathbb{R}^n$ is a real-valued function L such that

(L.1) $\mathrm{Dom}(f) \subseteq \mathrm{Dom}(L)$;

(L.2) the function $t \to L(\xi(t),t)$ is locally integrable for every $\xi \in \mathrm{Traj}(f)$.

If $f \in TVVF(\mathbb{R}^n)$ has components $f^1,\ldots,f^n$, and L is a Lagrangian for f, then we can consider the time-varying vector field $[L;f]$ in $\mathbb{R}^{n+1}$ whose components are $(f^0,f^1,\ldots,f^n)$, where $f^0 \equiv L$. A trajectory of $[L;f]$ can then be regarded as a pair (ξ^0,ξ), where $\xi \in \mathrm{Traj}(f)$ and ξ^0 is a real-valued locally absolutely continuous function on $\mathrm{Dom}(\xi)$ such that $\dot\xi^0(t) = L(\xi(t),t)$ for a.e. t. Any such function ξ^0 will be called a *running L-cost along* ξ. Our definition of a Lagrangian implies that a running L-cost along ξ exists for all $\xi \in \mathrm{Traj}(f)$, and it is clear that any two running L-costs along ξ must differ by a constant. If ξ is such that $\mathrm{Dom}(\xi)$ is of the form $[a,b]$ or $[a,b[$ or $[a,\infty[$, then there is a canonical choice of running L-cost along ξ, obtained by letting $\xi^0(t) = \int_a^t L(\xi(s),s)\,ds$. In that case, we will write ξ^L to denote the curve $t \to (\xi^0(t),\xi(t))$, where ξ^0 is the canonical running L-cost.

Now suppose we are given data P, $\mathbf{f}$, p_0, ξ as in (D.1,2,3,4), as well as

(D.5) a family $\mathbf{L} = \{L_p\}_{p\in P}$ such that each L_p is a Lagrangian for f_p.

Given a trajectory $\xi : [a,b] \to \mathbb{R}^n$ of f_{p_0}, we say that the pair $(\mathbf{f},\mathbf{L})$ is *weakly differentiable at* p_0 *along* ξ if the family $\{[L_p;f_p]\}_{p\in P}$ is weakly differentiable along $\xi^{L_{p_0}}$.

It is easy to see that $(\mathbf{f},\mathbf{L})$ is weakly differentiable at p_0 along ξ if and only if there exists $\bar\varepsilon > 0$ such that (WD1,2,3,4) hold, and in addition the following hold:

(WD5) *There exists* $h \in L^1([a,b],\mathbb{R}_n)$ *such that*

$$||L_p(x,t) - L_p(\xi(t),t) - h(t).(x-\xi(t))|| \le \psi_\varepsilon(t)(||x-\xi(t)|| + ||p-p_0||) \tag{A.55}$$

whenever $a \le t \le b$, $||x-\xi(t)|| \le \varepsilon$, *and* $||p-p_0|| \le \varepsilon$.

(WD6) *If we let* $w_p^0(t) = L_p(\xi(t),t) - L_{p_0}(\xi(t),t)$, $W_p^0(t) = \int_a^t w_p^0(s)\,ds$, *then the map* $\mathcal{W}^0$ *that sends* p *to* $\mathcal{W}^0(p) = W_p^0 \in BV([a,b],\mathbb{R})$ *is weak* Fréchet differentiable at* p_0.

If $f \in TVVF(\mathbb{R}^n)$, $\xi : [a,b] \to \mathbb{R}^n$ is a trajectory of f, and L is a Lagrangian for f, we say that L *is differentiable along* ξ if the map $x \to L(x,t)$ is differentiable at $\xi(t)$ for almost every t. In that case, we write $\nabla_\xi L(t) \overset{\mathrm{def}}{=} \nabla_x L(\xi(t),t)$, and refer to the map $[a,b] \ni t \to \nabla_\xi L(t) \in \mathbb{R}^n$ as the *differential of* L *along* ξ.

Lemma A.8, applied to the curve $\xi^{L_{p_0}}$, yields the following.

LEMMA A.10. *If* $(\mathbf{f},\mathbf{L})$ *is weakly differentiable at* p_0 *along a trajectory* $\xi : [a,b] \to \mathbb{R}^n$ *of* f_{p_0}, *then* f_{p_0} *and* L_{p_0} *are differentiable along* ξ *and the maps* A, h *of* (WD3,5) *are such that* $A(t) = D_\xi f_{p_0}(t)$ *and* $h(t) = \nabla_\xi L_{p_0}(\xi(t),t)$ *for almost all* $t \in [a,b]$.

Under the hypothesis of Lemma A.10, we can apply Theorem A.9 to the family $\{[L_p; f_p]\}_{p\in P}$ and the curve $\xi^{L_{p_0}}$. Define

$$J_p(\zeta) = \int_a^b L_p(\zeta(t), t)\, dt, \tag{A.56}$$

if $\zeta : [a, b] \to \mathbb{R}^n$ is a curve such that the integral exists. (We will only be using this when $\zeta \in \mathrm{Traj}(f_p)$, in which case $J_p(\zeta)$ is well defined.) An adjoint vector for $[L_{p_0}; f_{p_0}]$ along $\xi^{L_{p_0}}$ is then a pair (λ^0, λ) such that $\lambda^0 \in \mathbb{R}$, $\lambda : [a, b] \to \mathbb{R}_n$ is absolutely continuous, and

$$\dot\lambda(t) = -\lambda(t).A(t) - \lambda^0 h(t) \qquad \text{for a.e.} \qquad t \in [a, b]. \tag{A.57}$$

Theorem A.9 then implies the following.

Theorem A.11. *Assume that we are given data P, $\mathbf{f}$, p_0, ξ, $\mathbf{L}$ as in* (D.1,2,3,4,5). *Assume that $(\mathbf{f}, \mathbf{L})$ is weakly differentiable at p_0 along ξ. Let $P \ni p \to \bar x_p \in \mathbb{R}^n$ be a map which is Fréchet differentiable at p_0 and such that $\bar x_{p_0} = \xi(a)$ and let L_a be its differential at p_0. Let ξ_p be, for p close enough to p_0, maximal trajectories of f_p such that $\xi_p(a) = \bar x_p$, and let x_p be the restriction of ξ_p to $[a, b]$. Then the maps $p \to x_p(b)$ and $p \to J_p(x_p)$ are differentiable at p_0. Let L_b, L_b^0 be their differentials at p_0. Then, if (λ^0, λ) is an adjoint vector for $[L_{p_0}, f_{p_0}]$ along ξ, it follows that*

$$\begin{aligned}&\lambda(b) \cdot L_b(v) + \lambda^0 L_b^0(v) - \lambda(a) \cdot L_a(v)\\ &\quad= \lim_{\varepsilon\downarrow 0} \varepsilon^{-1} \int_a^b \Big(H_{p_0+\varepsilon v}(\lambda(t), \lambda_0, \xi(t), t) - H_{p_0}(\lambda(t), \lambda_0, \xi(t), t)\Big)\, dt,\end{aligned} \tag{A.58}$$

where

$$H_p(z, z^0, x, t) = \langle z, f_p(x, t)\rangle + z^0 L_p(x, t) \tag{A.59}$$

for $x \in \mathbb{R}^n$, $z \in \mathbb{R}_n$, $z^0 \in \mathbb{R}$, $t \in [a, b]$.

Appendix B. A lemma.

Lemma B.1. *Let f be a real-valued function on a compact interval $[a, b]$. Assume that there exists a finite or countable subset E of $[a, b]$ with the following properties:*

(B.1.a) $\liminf_{h\downarrow 0} \frac{f(x+h)-f(x)}{h} \geq 0$ *for all $x \in [a, b[\backslash E$,*

(B.1.b) $\liminf_{h\downarrow 0} f(x + h) \geq f(x)$ *for all $x \in [a, b[$,*

(B.1.c) $\liminf_{h\downarrow 0} f(x - h) \leq f(x)$ *for all $x \in]\,a, b]$.*

Then $f(b) \geq f(a)$.

Proof. Let $\varepsilon > 0$ be arbitrary. We will prove that $f(b) \geq f(a) - \varepsilon$. Let $\varepsilon' = \frac{\varepsilon}{1+b-a}$ and $E = \{x_k : k \in \mathbb{N}\}$.

For each $x \in [a, b]$, let $I(x) = \{k : x_k < x\}$. Define $g : [a, b] \to \mathbb{R}$ by

$$g(x) = \sum_{k\in I(x)} 2^{-k}\varepsilon'.$$

Then g is left-continuous and monotonically nondecreasing.

Let S be the set of all $x \in [a, b]$ such that $f(y) \geq f(a) - (y - a)\varepsilon' - g(y)$ for all $y \in [a, x]$. It is clear that $a \in S$, and also that S is an interval. So S is of the form $[a, c[$ or $[a, c]$, with $c \in [a, b]$.

Suppose $S = [a, c\,[$. Then $c > a$. By (B.1.c), given $\delta > 0$ there is a sequence $\{z_j\}$ of points of $[a, b]$ such that $z_j < c$, $z_j \to c$, and $f(c) \geq f(z_j) - \delta$. Then the z_j belong to S, so the inequalities $f(z_j) \geq f(a) - (z_j - a)\varepsilon' - g(z_j) \geq f(a) - (c-a)\varepsilon' - g(c)$ hold. It then follows that $f(c) \geq f(a) - (c-a)\varepsilon' - g(c) - \delta$. Since δ is arbitrary, we have $f(c) \geq f(a) - (c-a)\varepsilon' - g(c)$, so $c \in S$, which is a contradiction.

Therefore $S = [a, c]$ for some $c \in [a, b]$. We now show that $c = b$. Assume that $c < b$, and consider separately the cases $c \notin E$, $c \in E$.

If $c \notin E$, then (B.1.a) implies that there exists a $c' > c$ such that $f(x) - f(c) \geq -(x-c)\varepsilon'$ for every $x \in [c, c']$. But then, since $f(c) \geq f(a) - (c-a)\varepsilon' - g(c)$, we can conclude that $f(x) \geq f(a) - (x-a)\varepsilon' - g(c) \geq f(a) - (x-a)\varepsilon' - g(x)$ for $x \in [c, c']$. Therefore $c' \in S$, contradicting the fact that $S = [a, c]$.

Now suppose $c \in E$. Then $c = x_k$ for some k, and $x_k \neq b$. By (B.1.b), there exists $h > 0$ such that $c + h \leq b$ and $f(y) - f(c) \geq -2^{-k}\varepsilon'$ for all $y \in [c, c+h]$. If $y \in]\,c, c+h]$, then $k \notin I(c)$, and $I(c) \cup \{k\} \subseteq I(y)$. Therefore $g(y) \geq g(c) + 2^{-k}\varepsilon'$. Since $f(c) \geq f(a) - (c-a)\varepsilon' - g(c)$, we have $f(y) \geq f(c) - 2^{-k}\varepsilon' \geq f(a) - (c-a)\varepsilon' - 2^{-k}\varepsilon' - g(c) \geq f(a) - (y-a)\varepsilon' - g(y)$. So $c + h \in S$, which is a contradiction.

We have therefore proved that $c = b$. Then

$$f(b) \geq f(a) - (b-a)\varepsilon' - g(b) \geq f(a) - (b-a)\varepsilon' - \varepsilon' = f(a) - \varepsilon.$$

Since ε is arbitrary, we have shown that $f(b) \geq f(a)$, as desired. □

Appendix C. A simpler version of Brunovský's stratification condition. We prove that (Br.3.A) *holds if and only if $\mathcal{P}$ is a stratification of S.* It is clear that if $\mathcal{P}$ is a stratification of S, then (Br.3.A) holds. Let us prove the converse. Suppose that (Br.3.A) holds. Then $\mathcal{P}$ is a locally finite partition of S into nonempty connected embedded $\mathcal{C}^1$ submanifolds of S. Let us prove by contradiction that the frontier axiom (FA) holds. Suppose the axiom was violated for some pair (P_1, P_2) of members of $\mathcal{P}$. Of all these "bad" pairs, choose one for which P_1 has the largest possible dimension. Then P_1, P_2 are in $\mathcal{P}$, $P_1 \neq P_2$ (so $P_1 \cap P_2 = \emptyset$), and $P_1 \cap clos(P_2) \neq \emptyset$. Let $\tilde{P}_1 = P_1 \cap clos(P_2)$, so $\tilde{P}_1$ is a nonempty relatively closed subset of P_1. We will show that $\tilde{P}_1 = P_1$, contradicting the fact that (P_1, P_2) is a bad pair. To begin with, we cannot have $P_2 = \{0\}$, for if $P_2 = \{0\}$, then P_2 would be closed, so $P_1 \cap P_2 = P_1 \cap clos(P_2) = \tilde{P}_1 \neq \emptyset$, which is a contradiction. Also, $\dim(P_1) < n$, for if $\dim(P_1) = n$, then P_1 would be open, so the fact that $P_1 \cap clos(P_2) \neq \emptyset$ would imply that $P_1 \cap P_2 \neq \emptyset$, which is a contradiction. Finally, we may assume that $P_1 \neq \{0\}$, because if $P_1 = \{0\}$, then $\tilde{P}_1 = \{0\} = P_1$, because $\tilde{P}_1 \subseteq P_1$ and $\tilde{P}_1 \neq \emptyset$. So from now on we take it for granted that $P_1 \neq \{0\}$, $P_2 \neq \{0\}$, and $\dim(P_1) < n$, so, in particular, $P_1 \in \mathcal{P}'$. Let us first assume that $\dim(P_2) < n$. Then $P_2 \in \mathcal{P}'$. So both P_2 and P_1 belong to $\mathcal{P}'$, and in addition $P_1 \cap clos(P_2) \neq \emptyset$. Since $\mathcal{P}'$ is a stratification of S' by (Br.3.A), it follows that $P_1 \subseteq clos(P_2)$, contradicting the fact that the pair (P_1, P_2) was bad. This excludes the possibility that $\dim(P_2) < n$, and we are left with the case when $\dim(P_2) = n$, so P_2 is open in S. Let $\hat{P}_1 \stackrel{\text{def}}{=} P_1 \backslash \tilde{P}_1$. The fact that (P_1, P_2) is a bad pair implies that $\hat{P}_1 \neq \emptyset$. Since $\tilde{P}_1 \neq \emptyset$, $\tilde{P}_1$ is relatively closed in P_1, and P_1 is connected, it follows that $\hat{P}_1$ is not relatively closed in P_1. So $\tilde{P}_1 \cap clos_{P_1}(\hat{P}_1) \neq \emptyset$. Pick a point $\bar{x} \in \tilde{P}_1 \cap clos_{P_1}(\hat{P}_1)$. Let $m = \dim(P_1)$. We already know that $m < n$. Choose a coordinate chart $\mathbf{x} = (x^1, \ldots, x^n) : N \to \mathbb{R}^n$ that maps an open neighborhood N of $\bar{x}$ diffeomorphically onto the open cube $]-1, 1\,[^{\,n}$ in $\mathbb{R}^n$, in such a way that $\mathbf{x}(\bar{x}) = 0$ and $\mathbf{x}(P_1 \cap N) =]-1, 1\,[^{\,m} \times \{0\}^{n-m}$. For $\delta > 0$, let $N(\delta) = \mathbf{x}^{-1}(]-\delta, \delta\,[^{\,n})$. Since $\mathcal{P}$ is locally finite, if we let $Q(\delta)$ be the set of all $P \in \mathcal{P}$ that intersect $N(\delta)$, then there must exist a $\bar{\delta}$ such that the set $Q = Q(\delta)$ is finite and independent of δ for $\delta \in]\,0, \bar{\delta}]$. Clearly, $P_1 \in Q$ and $P_2 \in Q$. Let $Z(\delta) = clos_{N(\delta)}(P_2 \cap N(\delta))$. Then $(P_2 \cup \tilde{P}_1) \cap N(\delta) \subseteq Z(\delta)$. Let us exclude the possibility that $(P_2 \cup \tilde{P}_1) \cap N(\delta) = Z(\delta)$.

If $(P_2 \cup \tilde{P}_1) \cap N(\delta) = Z(\delta)$, then $P_2 \cap (N(\delta) \backslash P_1)$ is relatively closed in $N(\delta)\backslash P_1$. Since $P_2 \cap (N(\delta)\backslash P_1)$ is also relatively open in $N(\delta)\backslash P_1$ and nonempty, it follows that $P_2 \cap (N(\delta)\backslash P_1)$ must be a union of connected components of $N(\delta)\backslash P_1$. If $m < n-1$, then $N(\delta)\backslash P_1$ is connected, so $P_2 \cap (N(\delta)\backslash P_1) = N(\delta)\backslash P_1$. If $m = n-1$, then $N(\delta)\backslash P_1$ has two connected components, and $P_1 \cap N(\delta)$ is entirely contained in the closure of each one of them. So in both cases it follows that $P_1 \cap N(\delta) \subseteq clos(P_2)$. But then $P_1 \cap N(\delta) \subseteq \tilde{P}_1$, showing that $\tilde{P}_1$ is a relative neighborhood of $\bar{x}$ in P_1, which contradicts our choice of $\bar{x}$ as a limit point in P_1 of $P_1 \backslash \tilde{P}_1$. We now know that $(P_2 \cup \tilde{P}_1) \cap N(\delta)$ is a proper subset of $Z(\delta)$, for each $\delta \in]\,0, \bar{\delta}]$. Fix a sequence $\{\delta_j\}$ such that $0 < \delta_j \leq \bar{\delta}$ for all j and $\delta_j \downarrow 0$. Pick $y_j \in Z(\delta_j)\backslash((P_2 \cup \tilde{P}_1) \cap N(\delta_j))$. Then every y_j belongs to a $P^j \in Q$, but $y_j \notin P_1 \cup P_2$. (It is obvious that $y_j \notin P_2$ and $y_j \notin \tilde{P}_1$, since $y_j \in N(\delta_j)$. On the other hand, $y_j \in clos(P_2)$, so y_j cannot be in P_1, because if it was in P_1 it would belong to $\tilde{P}_1$.) So $P_1 \neq P^j \neq P_2$ for every j. Since Q is finite, we may assume, after passing to a subsequence, that all the P_j's are equal to one and the same $P \in Q$. By construction, $P \cap clos(P_2) \neq \emptyset$, because all the y_j are in $clos(P_2)$. Since $P \neq P_2$, P cannot be open, because $P \in \mathcal{P}$, so $P \cap P_2 = \emptyset$, thanks to the fact that $\mathcal{P}$ is a partition, so if P was open, then $P \cap clos(P_2)$ would be empty. So $\dim(P) < n$. Moreover, $P \neq \{0\}$ because $y_j \neq \bar{x}$ but $y_j \to \bar{x}$ and $y_j \in P$. So $P \in \mathcal{P}'$. Since $\bar{x} \in clos(P) \cap P_1$, and both P and P_1 are in $\mathcal{P}'$, it follows that $P_1 \subseteq clos(P)$ and $\dim(P_1) < \dim(P)$. Since we have chosen (P_1, P_2) to be a bad pair for which $\dim(P_1)$ is maximized, we can infer that (P, P_2) is not a bad pair. Since P and P_2 are in $\mathcal{P}$, and $P \cap clos(P_2) \neq \emptyset$, the inclusion $P \subseteq clos(P_2)$ must hold. But then $clos(P) \subseteq clos(P_2)$ as well, and this implies that $P_1 \subseteq clos(P_2)$, since we know that $P_1 \subseteq clos(P)$. So (P_1, P_2) is not a bad pair after all, and our proof is complete.

Acknowledgments. We want to thank Michael Malisoff for carefully reading the paper and for many useful discussions and suggestions. The final part of the work was carried out between January and June 1998 while the second author author was a visitor at the Institut Henri Poincaré in Paris. He is grateful to the Institut Henri Poincaré for its kind hospitality during this period.

REFERENCES

[1] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equation*, Birkhäuser, Boston, MA, 1997.

[2] V. G. Boltyanskii, *Sufficient conditions for optimality and the justification of the dynamic programming principle*, SIAM J. Control Optim., 4 (1966), pp. 326–361.

[3] A. Bressan and B. Piccoli, *Structural stability for time-optimal planar syntheses*, Dynam. Contin. Discrete Impuls. Systems, 3 (1997), pp. 335–371.

[4] A. Bressan and B. Piccoli, *A generic classification of time optimal planar stabilizing feedbacks*, SIAM J. Control Optim., 36 (1998), pp. 12–32.

[5] P. Brunovský, *Every normal linear system has a regular time-optimal synthesis*, Math. Slovaca, 28 (1978), pp. 81–100.

[6] P. Brunovský, *Existence of regular syntheses for general problems*, J. Differential Equations, 38 (1980), pp. 317–343.

[7] L. Cesari, *Optimization—Theory and Applications*, Springer–Verlag, New York, 1983.

[8] F. H. Clarke, Yu. S. Ledyaev, A. I. Subbotin, and E. D. Sontag, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[9] H. Federer, *Geometric Measure Theory*, Springer–Verlag, Berlin, 1969.

[10] A. F. Filippov, *Differential Equations with Discontinuous Right-Hand Side*, D. Reidel, Boston, 1989.

[11] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer–Verlag, New York, 1975.

[12] W. H. Fleming and M. Soner, *Controlled Markov Processes and Viscosity Solutions*,

Springer–Verlag, New York, 1993.

[13] A. T. Fuller, *Relay control systems optimized for various performance criteria*, Automatic and Remote Control, in Proceedings of the First World Congress IFAC, Moscow, Vol. 1, Butterworth, London, 1961, pp. 510–519.

[14] S. Lojasiewicz, Jr. and H. J. Sussmann, *Some examples of reachable sets and optimal cost functions that fail to be subanalytic*, SIAM J. Control Optim., 23 (1985), pp. 584–598.

[15] C. Marchal, *Chattering arcs and chattering controls*, J. Optim Theory Appl., 11 (1973), pp. 441–468.

[16] B. Piccoli, *Regular time-optimal syntheses for smooth planar systems*, Rend. Sem. Mat. Univ. Padova, 95 (1996), pp. 59–79.

[17] B. Piccoli, *Classification of generic singularities for the planar time-optimal syntheses*, SIAM J. Control Optim., 34 (1996), pp. 1914–1946.

[18] H. J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.

[19] H. J. Sussmann, *Piecewise analyticity of optimal cost functions and optimal feedback*, in Proceedings of the Joint Automatic Control Conference, Denver, CO, August 1979, IEEE, New York, 1979.

[20] H. J. Sussmann, *Analytic stratifications and control theory*, in Proceedings of the 1978 International Congress of Mathematicians, Helsinki, 1980, pp. 865–871.

[21] H. J. Sussmann, *Time-optimal control in the plane*, in Feedback Control of Linear and Nonlinear Systems, D. Hinrichsen and A. Isidori, eds., Springer–Verlag, New York, 1982, pp. 244–260.

[22] H. J. Sussmann, *Subanalytic sets and optimal control in the plane*, in Proceedings of the 21st IEEE Conference on Decision and Control, Orlando, FL, December 1982, IEEE, New York, 1982, pp. 295–299.

[23] H. J. Sussmann, *Lie brackets, real analyticity and geometric control theory*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 1–115.

[24] H. J. Sussmann *Lie brackets and real analyticity in control theory*, in Mathematical Control Theory, C. Olech, ed., Banach Center Publ. 14, PWN-Polish Scientific Publishers, Warsaw, 1985, pp. 515–542.

[25] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.

[26] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The general real-analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.

[27] H. J. Sussmann, *Regular synthesis for time-optimal control of single-input real analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.

[28] H. J. Sussmann, *Recent developments in the regularity theory of optimal trajectories*, in Conference on Linear and Nonlinear Mathematical Control Theory, Rend. Sem. Mat. Univ. Politec., Torino, Fascicolo speciale, (1987), pp. 149–182.

[29] H. J. Sussmann, *Envelopes, high-order optimality conditions and Lie brackets*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, December 1989, IEEE, New York, pp. 1107–1112.

[30] H. J. Sussmann, *Envelopes, conjugate points and optimal bang-bang extremals*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel, eds., Math. Appl. 29, D. Reidel, Dordrecht, Boston, MA, 1986, pp. 325–346.

[31] H. J. Sussmann, *Synthesis, presynthesis, sufficient conditions for optimality and subanalytic sets*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 1–19.

[32] M. I. Zelikin and V. F. Borisov, *Theory of Chattering Control, with Applications to Astronautics, Robotics, Economics and Engineering*, Birkhäuser, Boston, 1994.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 411–432

NONPARAMETRIC ESTIMATION AND ADAPTIVE CONTROL OF FUNCTIONAL AUTOREGRESSIVE MODELS[*]

BRUNO PORTIER[†] AND ABDERRAHIM OULIDI[‡]

Abstract. This paper deals with nonparametric estimation and adaptive control of nonlinear systems of the form $X_{n+1} = f(X_n) + U_n + \xi_{n+1}$ ($n \in \mathbb{N}$) where the state X_n is observed, f is an unknown function, and the control U_n is chosen in order to track a given reference trajectory. We estimate the function f using a nonparametric estimator and study two adaptive control laws built from this nonparametric estimator and derived from the self-tuning control. The first one can be used for open-loop stable systems and requires an additional exciting noise. The second one needs some a priori knowledge on function f but allows us to control open-loop unstable systems. We establish some general results on the nonparametric estimator of f like the uniform almost sure convergence over dilating sets and then prove that both adaptive control laws are asymptotically optimal in quadratic mean. In addition, we give a strongly consistent estimator of the covariance matrix of the unobservable white noise ξ_n.

Key words. adaptive control, discrete-time stochastic nonlinear system, nonparametric estimation, optimal adaptive tracking

AMS subject classifications. 93C40, 62G07, 93E20, 93C55, 93C10

PII. S0363012997316676

1. Introduction.

Since Aström and Wittenmark [2] introduced the self-tuning regulator in 1973, the stochastic adaptive tracking problem has drawn much attention from the control community (see, for example, Chen and Guo [7], Caines [6], Chen [8], Duflo [13], and the references therein). For linear models, the ARX and ARMAX models, the difficult problem of identifying unknown parameters and simultaneously tracking a reference signal has been completely solved using both the weighted least squares algorithm (see Bercu [3], [4] and Guo [18]) and a slight modification of the extended least squares algorithm (see Guo and Chen [16], Guo [17]).

Nevertheless, these linear models are not well suited for modelling when the relation between the state and its past is nonlinear. Several authors have proposed interesting methods for adaptive control of nonlinear models: neural networks-based methods, for example, have been increasingly used (see Narendra and Parthasarathy [27], Chen and Khalil [9] and Jagannathan, Lewis, and Pastravanu [22]). But, for these methods, theoretical results are not always proven.

In this paper, we consider the problem of adaptive control of discrete-time nonlinear stochastic systems. We will focus on NARX models on $\mathbb{R}^d$ ($d \in \mathbb{N}^*$) of the form

$$X_{n+1} = f(X_n) + U_n + \xi_{n+1} \quad (n \in \mathbb{N}), \tag{1.1}$$

where X_n, U_n, and ξ_n are the system output, input, and driven noise of the system, respectively. State X_n is observed, driving function f is unknown, ξ_n is an unobservable white noise, and control U_n is chosen in order to track a given deterministic reference trajectory, denoted by $(X_n^*)_{n\geq 1}$.

[*] Received by the editors February 12, 1997; accepted for publication (in revised form) March 13, 2000; published electronically August 31, 2000.

http://www.siam.org/journals/sicon/39-2/31667.html

[†] Laboratoire de Mathématique – U.M.R. C 8628, "Probabilités, Statistique et Modélisation," Université Paris-Sud, Bât. 425, 91 405 Orsay Cedex, France (bruno.portier@math.u-psud.fr).

[‡] I.M.A., Université Catholique de l'Ouest, 44 rue Rabelais, BP 808, 49008 Angers Cedex 01, France.

Clearly, this is quite a small class of models. Indeed, state is observed, time delay is equal to one, both control and state are in the same space ($\mathbb{R}^d$), and the system is always stabilizable. Our work, though, is a first step toward the study of nonlinear stochastic control systems and deals with the problem of unknown function f estimation. It should be noticed, however, that despite its simplicity, model (1.1) has already been used (in a slightly different version) to regulate the output gas flow-rate of an aerobic digestion process by adapting the liquid flow-rate of an influent of industrial wine distillery wastewater (see Hilgert et al. [21]).

In order to estimate function f, we introduce a kernel method-based recursive estimator (see Härdle [19], Devroye and Györfi [11]). We use general results on nonparametric estimation of regression functions. Using nonparametric estimation may come with a price, however. Compared with parametric methods, nonparametric ones have slower convergence rates. This is the counterpart of the flexibility of such nonparametric design. Let us mention that kernel estimation methods are also used in identification problems for nonlinear dynamic systems such as Hammerstein systems (see Greblicki [14], Greblicki and Pawlak [15]).

Nonparametric estimation of regression function has often been studied in a noncontrolled framework, i.e., $U_n \equiv 0$ in model (1.1) (see Collomb [10], Doukhan and Ghindès [12], Ango Nze and Portier [1], Truong and Stone [32], Bosq [5]). Unfortunately, all these papers deal with stationary and mixing processes; therefore, these results and techniques are not suitable for controlled processes: U_n can depend on the current state X_n and on the previous ones. Thus, another approach is necessary.

Duflo [13] and Senoussi [31] have given the first convergence results for $\widehat{f}_n$ in a control framework. For nonadaptive control laws ensuring that model (1.1) is stable (in the sense defined in section 2), they prove that $\forall\, A < \infty$

$$\sup_{\|x\| \leq A} \left\| \widehat{f}_n(x) - f(x) \right\| \xrightarrow[n\to\infty]{a.s.} 0, \tag{1.2}$$

where $\widehat{f}_n$ denotes the nonparametric estimator of function f. Hilgert, Senoussi, and Vila [20] generalize this result for models where the unknown function f depends on time n. In this paper, we extend result (1.2): we establish the uniform almost sure convergence of $\widehat{f}_n$ over dilating sets, i.e.,

$$\sup_{\|x\| \leq v_n} \left\| \widehat{f}_n(x) - f(x) \right\| \xrightarrow[n\to\infty]{a.s.} 0, \tag{1.3}$$

where $(v_n)_{n\geq 0}$ is a sequence of positive numbers increasing to infinity. Such a result allows us to study adaptive control laws built from $\widehat{f}_n$ which was not possible with result (1.2). Indeed, in order to obtain the asymptotic optimality of the tracking, we have to prove that

$$\frac{1}{n} \sum_{k=1}^{n} \left\| \widehat{f}_k(X_k) - f(X_k) \right\|^2 \xrightarrow[n\to\infty]{a.s.} 0 \tag{1.4}$$

and clearly, the uniform almost sure convergence over fixed compact sets is not sufficient.

In order to solve the problem of tracking, we propose two adaptive control laws, based on the certainty-equivalence principle. The first one can be used only when model (1.1) is open-loop stable. It requires an additional exciting term:

$$U_n \;=\; -\widehat{f}_n(X_n) \;+\; X^*_{n+1} \;+\; \gamma_{n+1}\, \eta_{n+1}, \tag{1.5}$$

where $\eta = (\eta_n)_{n\geq 1}$ is a Gaussian white noise and $(\gamma_n)_{n\geq 1}$ a sequence of real numbers decreasing to 0. The second adaptive control law requires some a priori knowledge about function f, but allows us to control open-loop unstable systems:

$$U_n \;=\; -\widehat{f}_n(X_n)\mathbb{1}_{E_n}(X_n) \;-\; f^*(X_n)\mathbb{1}_{\overline{E}_n}(X_n) \;+\; X^*_{n+1}, \tag{1.6}$$

where $\overline{E}_n$ denotes the complementary set of E_n. Function f^* characterizes the a priori knowledge about function f and allows us to compensate the possible lack of observations which disrupts the local estimator $\widehat{f}_n$. Set E_n, which will be specified in section 4.2, is introduced to ensure the closed-loop stability of model (1.1).

For both control laws, we prove that the tracking is asymptotically optimal, i.e.,

$$\frac{1}{n}\sum_{k=1}^{n} \|X_k - X^*_k\|^2 \;\underset{n\to\infty}{\xrightarrow{a.s.}}\; \mathrm{trace}(\Gamma), \tag{1.7}$$

where Γ denotes the covariance matrix of the noise ξ_n.

This paper is organized as follows. Section 2 is devoted to the model assumptions. Section 3 is concerned with the nonparametric estimation of function f. In section 4, we explain the adaptive control problem and we study the properties of control laws (1.5) and (1.6). In addition, the strong consistency of an estimator of Γ is proven. Finally, Appendices A to D contain proofs of the main results.

2. Model assumptions.

Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$, where $\mathcal{F}_n$ is the σ-algebra generated by events occurring up to time n. We assume that control $U = (U_n)_{n\geq 0}$ and noise $\xi = (\xi_n)_{n\geq 1}$ are adapted to $\mathcal{F}$, and that X_0 and U_0 are $\mathcal{F}_0$-measurable and arbitrarily chosen. Typically, here we have $\mathcal{F}_n = \sigma(X_0, U_0, \xi_1, \ldots, \xi_n)$. Finally, let us assume the following properties for model (1.1).

ASSUMPTION [$A1$] (*about function f*). *Function f is Lipschitz:*

$$\exists\, r_f < \infty,\ \forall\, x \in \mathbb{R}^d,\ \forall\, y \in \mathbb{R}^d,\quad \|f(x) - f(y)\| \;\leq\; r_f\,\|x - y\|\,,$$

where $\|\,.\,\|$ denotes the usual norm on $\mathbb{R}^d$.

ASSUMPTIONS [$A2$] (*about noise ξ*).

- *$\xi = (\xi_n)_{n\geq 1}$ is a sequence of independent and identically distributed random vectors with mean 0 and invertible covariance matrix Γ which is supposed to be unknown.*
- *ξ_n has a finite moment of order $m > 2$ and its distribution is absolutely continuous with respect to the Lebesgue measure. Its probability density function denoted by p is supposed to be C^1-class, p and its gradient are bounded.*

In this paper, the following definition for stability is used.

DEFINITION 2.1. *The process $(X_n)_{n\geq 0}$, defined by (1.1), is said to be stable if there exist constants $\mu > 0$ and $M < \infty$ such that for any initial law (the law of X_0),*

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \|X_k\|^{\mu} \;\leq\; M\,,\ \ a.s.$$

We will see in section 4 that the system driven by control (1.5) or (1.6) is stable with $\mu = m$. Let us now present nonparametric estimator $\widehat{f}_n$ and the uniform almost sure convergence of $\widehat{f}_n$ over dilating sets.

3. Estimation of function f. In order to estimate unknown function f, we use a recursive version of the classical kernel estimator of the regression function (see, for example, Härdle [19]). Let K be a function from $\mathbb{R}^d$ to $\mathbb{R}_+$ satisfying $\int K(y)\,dy = 1$ and let α be a real number in $]0\,,\,1/d[$. Function K is called the kernel and α the bandwidth parameter. For $x \in \mathbb{R}^d$, we estimate $f(x)$ by

$$\widehat{f}_n(x) = \frac{\displaystyle\sum_{i=0}^{n-1} i^{\alpha d} K\Big(i^{\alpha}(X_i - x)\Big)\Big(X_{i+1} - U_i\Big)}{\displaystyle\sum_{i=0}^{n-1} i^{\alpha d} K\Big(i^{\alpha}(X_i - x)\Big)} \tag{3.1}$$

if the denominator of $\widehat{f}_n(x)$ is not equal to 0, and by 0 otherwise.

Let us remark that $\widehat{f}_n(x)$ can be rewritten as $\sum_{i=0}^{n-1} w_i(X_i, x)\,(X_{i+1} - U_i)$, where $w_i(X_i, x) \geq 0$ and $\sum_{i=0}^{n-1} w_i(X_i, x) = 1$. So, $\widehat{f}_n(x)$ appears to be a weighted sum of $(X_{i+1} - U_i)_{0 \leq i \leq n-1}$. The closer X_i is to x, the greater is the weight $w_i(X_i, x)$, thanks to the kernel K and the bandwidth parameter α, and therefore, the greater is the contribution of $(X_{i+1} - U_i)$ to $\widehat{f}_n(x)$.

Nonparametric estimator (3.1) is said to be recursive because for $x \in \mathbb{R}^d$ and $n \geq 0$ we have

$$\begin{aligned} H_n(x) &= H_{n-1}(x) \ + \ n^{\alpha d} K\Big(n^{\alpha}(X_n - x)\Big)\,,\ H_{-1}(x) \ = \ 0, \\ N_n(x) &= N_{n-1}(x) \ + \ n^{\alpha d} K\Big(n^{\alpha}(X_n - x)\Big)\,\Big(X_{n+1} - U_n\Big)\,,\ N_{-1}(x) \ = \ 0, \end{aligned}$$

where $N_{n-1}(x)$ and $H_{n-1}(x)$ denote the numerator and the denominator of $\widehat{f}_n(x)$, respectively. This recursive property is useful in the adaptive control framework. Indeed, as soon as a new observation becomes available, $\widehat{f}_n(x)$ can easily be updated.

The following assumptions are made on kernel K.

Assumptions [$A3$] (*about kernel* K). $K : \mathbb{R}^d \to \mathbb{R}_+^*$ *is a positive bounded function, with a compact support such that*

- $\int K(y)\,dy \ = \ 1$ *and* $\int \|y\|\,K(y)\,dy \ < \infty$.
- $\exists\ r_K < \infty,\ \forall\, x \in \mathbb{R}^d,\ \forall\, y \in \mathbb{R}^d,\ \ \|K(x) - K(y)\| \ \leq \ r_K\ \|x - y\|$.

For example, Epanechnikov's kernel, defined by

$$K(y) = K(y_1, \ldots, y_d) = \prod_{j=1}^{d} (3/4)(1 - y_j)^2 1\!\!1_{\{|y_j| \leq 1\}},$$

satisfies [$A3$].

Now we give convergence results for estimator $\widehat{f}_n$. Part 1 of Theorem 3.1 will be used to study the convergence of $\widehat{f}_n(x)$ with control law (1.5) and part 2 with control law (1.6).

Theorem 3.1. *Assume that* [$A1$], [$A2$], *and* [$A3$] *hold. Let* $\alpha \in\ ]0\,,\,1/2d[$ *and let* $(v_n)_{n\geq 1}$ *be a sequence of positive real numbers increasing to infinity such that* $v_n = O(n^{\nu})$ *with* $\nu > 0$.

1. *If there exists a sequence* $(w_n)_{n\geq 0}$ *of positive real numbers, increasing to infinity, such that for any initial law*

$$\liminf_{n\to\infty} \left(\inf_{\|x\| \leq v_n} \frac{1}{w_n}\ H_{n-1}(x) \right) \ > \ 0,\ \ a.s., \tag{3.2}$$

where $H_{n-1}(x) = \sum_{i=0}^{n-1} i^{\alpha d} K\left(i^{\alpha}(X_i - x)\right)$*, then, for any* $s \in \left]1/2 + \alpha d\,,\, 1\right[$,

$$\sup_{\|x\| \le v_n} \left\| \widehat{f}_n(x) - f(x) \right\| \overset{a.s.}{=} o\left(\frac{n^s}{w_n}\right) + O\left(\frac{n^{1-\alpha}}{w_n}\right). \tag{3.3}$$

2. *More particularly, if the probability density function p is strictly positive and if there exists a finite constant M such that for any initial law*

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \|X_k\|^2 \le M, \ \ a.s., \tag{3.4}$$

then (3.2) *and* (3.3) *hold with* $w_n = n\,m_n$ *and* $m_n = \inf\{p(z);\ \ \|z\| \le v_n + R\}$ *for some finite constant R.*

Proof. The proof is given in Appendix B. □

Comments. We will say that we have uniform almost sure dilated convergence of $\widehat{f}_n$ to f if and only if there exists a sequence $(v_n)_{n\ge 0}$ of positive real numbers, increasing to infinity, such that

$$\sup_{\|x\| \le v_n} \left\| \widehat{f}_n(x) - f(x) \right\| \xrightarrow[n\to\infty]{a.s.} 0. \tag{3.5}$$

From part 1 of Theorem 3.1, we can observe that the convergence of $\widehat{f}_n$ to f depends only on the behavior of the denominator H_{n-1} of $\widehat{f}_n$. We will see in section 4 (see Theorem 4.1) how to choose sequences (v_n) and (w_n) in order to satisfy condition (3.2) when control law (1.5) is used to drive model (1.1).

Let us make some comments about $H_{n-1}(x)$. The quantity $n^{-1}H_{n-1}(x)$ behaves as a kind of frequency of the past values of the process $(X_i)_{0\le i\le n-1}$ located in a neighborhood of x. Condition (3.2) constrains the asymptotical behavior of $H_{n-1}(x)$, in order to have enough observations in a neighborhood of x, whatever x may be. Let us also mention that, when process (X_n) is asymptotically stationary (see Devroye and Györfi [11]), $n^{-1}H_{n-1}(x)$ is an estimator of the probability density function of the stationary distribution at point x.

When the assumptions of part 2 are fulfilled, result (3.5) is true as soon as

$$m_n^{-1} = \left(\inf\left\{p(z)\ ;\ \|z\| \le v_n + R\right\}\right)^{-1} = \inf\left(o\left(n^{\alpha}\right)\ ,\ O\left(n^{1-s}\right)\right). \tag{3.6}$$

Thus, it is the decrease of the probability density function p and the choice of a well-suited sequence (v_n) which give the rate of the uniform almost sure dilated convergence of $\widehat{f}_n$. If p rapidly decreases to 0, then sequence (v_n) must slowly increase to infinity. There is, however, a well-known property of kernel-based estimation: any new observation improves the estimator only in a neighborhood of this observation. According to the uniform convergence requirement, any heavy tailed noise is better than any Gaussian-type noise. Indeed, for any given n, heavy tailed noises explorate a larger zone than Gaussian-type noises.

The following corollary emphasizes this point. Let us consider the widely used Gaussian noise. In that case, we are able to easily exhibit sequence (v_n) and specify sequence (m_n) and convergence rate of $\widehat{f}_n$.

Corollary 3.2. *Assume that* [A1], [A2], *and* (3.4) *hold, and that* ξ *is a Gaussian white noise with mean zero and invertible covariance matrix* Γ.

1. *For $\alpha = 1/2(d+1)$, any initial law and any $A < \infty$,*

$$\sup_{\|x\| \leq A(\text{loglog } n)^{1/2}} \left\| \widehat{f}_n(x) - f(x) \right\| \overset{a.s.}{=} O\left(n^{-\lambda}\right) , \tag{3.7}$$

with $\lambda \in \,]0\,,\, 1/2(d+1)[$.

2. *For $\alpha \in \,]0\,,\, 1/2d[$, any initial law and any positive constant A such that $A^2/\lambda_{\min}(\Gamma) < \inf(\alpha\,,\, 1-s)$, where $\lambda_{\min}(\Gamma)$ denotes the minimum eigenvalue of the matrix Γ,*

$$\sup_{\|x\| \, \leq \, A(\log \, n)^{1/2}} \left\| \widehat{f}_n(x) - f(x) \right\| \overset{a.s.}{=} O\left(n^{-\lambda}\right) , \tag{3.8}$$

with $\lambda = \inf(\alpha\,,\, 1-s) - A^2/\lambda_{\min}(\Gamma)$.

Proof. Since ξ_n is Gaussian, [*A*2] is fulfilled and the probability density function p is such that

$$\inf\Big\{ p(z) \ ; \ \|z\| \leq v_n + R \Big\} \ \geq \ \text{cte} \exp\left(- \frac{(v_n + R)^2}{2\,\lambda_{\min}(\Gamma)} \right) \ \geq \ \text{cte} \exp\left(- \frac{v_n^2}{\lambda_{\min}(\Gamma)} \right) .$$

Thus, if we take $v_n = A\,(\log\log\, n)^{1/2}$ with $A \in \,]0\,,\, \infty[$, condition (3.6) is verified since $m_n \geq \text{cte}\,(\log n)^{-\delta}$ with $\delta = A^2/\lambda_{\min}(\Gamma)$. The choice $\alpha = 1/2(d+1)$ leads to part 1.

Now, if we choose $v_n = A\,(\log\, n)^{1/2}$ with $A > 0$, then $m_n \geq \text{cte}\, n^{-A^2/\lambda_{\min}(\Gamma)}$ and part 2 is easily derived. □

Prediction errors. In forecasting and adaptive control problems, the behavior of the prediction errors is a natural question to address (the prediction error at time k is defined by $\widehat{f}_k(X_k) - f(X_k)$). The following convergence result can be useful:

$$\frac{1}{n} \sum_{k=1}^{n} \left\| \widehat{f}_k(X_k) - f(X_k) \right\|^2 \overset{a.s.}{=} o(1). \tag{3.9}$$

If we are able to prove that $\sup_{\|x\| \leq v_n} \|\widehat{f}_n(x) - f(x)\| \overset{a.s.}{=} o(1)$, for a sequence $(v_n)_{n\geq 1}$ increasing to infinity, then

$$\frac{1}{n} \sum_{k=1}^{n} \left\| \widehat{f}_k(X_k) - f(X_k) \right\|^2 \mathbb{1}_{\{\, \|X_k\| \, \leq \, v_k\}} \overset{a.s.}{=} o(1).$$

Result (3.9) is then easily derived by establishing a similar result for prediction errors of the form $\|\widehat{f}_k(X_k) - f(X_k)\|^2 \mathbb{1}_{\{\, \|X_k\| \, > \, v_k\}}$. The following theorem deals with this particular point.

Theorem 3.3. *Assume that* [*A*1], [*A*2], *and* [*A*3] *hold and that there exist two constants $\mu > 0$ and $M < \infty$ such that*

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \|X_k\|^{\mu} \leq M\,, \ \ a.s. \tag{3.10}$$

Let $(v_n)_{n\geq 1}$ be a sequence of positive real numbers increasing to infinity. Let us denote $\xi_n^{\#} = \sup_{k\leq n} \|\xi_k\|$. Then, for any $b \in [\,0\,,\, \mu\,[$,

$$\frac{1}{n} \sum_{k=1}^{n} \left\| \widehat{f}_k(X_k) - f(X_k) \right\|^b \mathbb{1}_{\{\|X_k\| > v_k\}} \overset{a.s.}{=} O\left(\frac{\left(\xi_n^{\#}\right)^b}{n} \sum_{k=1}^{n} v_k^{-\mu} + \frac{1}{n} \sum_{k=1}^{n} v_k^{b-\mu} \right).$$

Proof. First, Lemma B.1 gives

$$\sum_{k=1}^{n} \left\| \widehat{f}_k(X_k) - f(X_k) \right\|^b \, \mathbb{1}_{\{\|X_k\| > v_k\}} \overset{a.s.}{=} O\left(\sum_{k=1}^{n} \left(\|X_k\|^b + 1 + (\xi_k^{\#})^b \right) \mathbb{1}_{\{\|X_k\| > v_k\}} \right).$$

Then, part 2 of Lemma A.1 finishes the proof. □

Example. Let us once again consider the Gaussian white noise case and let us show that the prediction errors converge to 0 in quadratic mean. Assume that we can exhibit a control law ensuring (control law (1.6) is well suited) that there exist $\mu > 2$ and $M < \infty$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \|X_k\|^{\mu} \le M, \quad \text{a.s.}$$

Then, since ξ_n is Gaussian, $\xi_n^{\#} \overset{a.s.}{=} O((\log n)^{1/2})$ and taking $v_n = A\,(\log n)^{1/2}$, with $A > 0$ such that $A^2/\lambda_{\min}(\Gamma) < \inf(\alpha\,,\, 1 - s)$, we derive from result (3.8) and Theorem 3.3 that

$$\frac{1}{n} \sum_{k=1}^{n} \left\| \widehat{f}_k(X_k) - f(X_k) \right\|^2 \overset{a.s.}{=} O\Big((\log n)^{2-\mu} \Big).$$

Now we explain the adaptive control problem and give the properties of control laws (1.5) and (1.6). In addition, the strong consistency of an estimator of Γ is obtained.

4. Study of two optimal adaptive control laws. In this section we consider a simple problem of adaptive control: the problem of adaptive tracking. The goal of adaptive tracking is to find a control sequence $U = (U_n)_{n\ge 1}$ which forces the output $X = (X_n)_{n\ge 1}$ to track the given deterministic reference trajectory $X^* = (X_n^*)_{n\ge 1}$.

If function f were known, the control U_n defined by

$$U_n = -f(X_n) + X_{n+1}^* \tag{4.1}$$

would ensure that the tracking is asymptotically optimal in quadratic mean. We have, indeed, the following:

$$\frac{1}{n} \sum_{k=1}^{n} \|X_k - X_k^*\|^2 = \frac{1}{n} \sum_{k=1}^{n} \|\xi_k\|^2 \xrightarrow[n\to\infty]{a.s.} \operatorname{trace}(\Gamma). \tag{4.2}$$

When function f is unknown, we know how to estimate it using the kernel estimator (3.1) and then, we can replace f by $\widehat{f}_n$ in (4.1). Thus, the adaptive tracking control U_n is given by

$$U_n = -\widehat{f}_n(X_n) + X_{n+1}^*. \tag{4.3}$$

Using this control law, the closed-loop system is of the form

$$X_{n+1} - X_{n+1}^* = \Big(f(X_n) - \widehat{f}_n(X_n) \Big) + \xi_{n+1}. \tag{4.4}$$

Let us remember, as seen in section 3, that in order to obtain good properties of convergence for $\widehat{f}_n$, the closed-loop system must be stable. Since the closed-loop properties of system (1.1) driven by (4.3) are unknown, we propose two adaptive control laws which are derived from (4.3) and for which stability properties of the closed-loop system can be obtained by adding some complementary assumptions on model (1.1). Then, for both control laws, we prove that the tracking is asymptotically optimal, i.e.,

$$\frac{1}{n}\sum_{k=1}^{n}\|X_k - X_k^*\|^2 \underset{n\to\infty}{\xrightarrow{a.s.}} \operatorname{trace}(\Gamma) \tag{4.5}$$

and $\widehat{\Gamma}_n$, defined by

$$\widehat{\Gamma}_n = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - X_k^*\right)\left(X_k - X_k^*\right)^T, \tag{4.6}$$

is a strongly consistent estimator of Γ.

4.1. The excited tracking control. In this section, we study the excited tracking control law proposed by Oulidi in [28]. First, let us make the following assumptions for model (1.1).

Assumptions [*A*4] (*about function f and noise ξ*).

- *Function f satisfies* [A1]*with $r_f < 1$.*
- *White noise ξ is bounded and satisfies* [*A*2].

Let us remark that the crucial assumption is the open-loop stability of model (1.1), which is implied by the value of the Lipschitz coefficient. Moreover, from a practical point of view, the noise assumption is not too restrictive.

In order to control model (1.1), we use the excited tracking control defined as follows.

Construction of U_n. Let $(X_n^*)_{n\geq 1}$ and $(\eta_n)_{n\geq 1}$, respectively, be a bounded deterministic reference trajectory and a Gaussian white noise with mean zero and invertible covariance matrix Γ_η, and let $(\gamma_n)_{n\geq 1}$ be the sequence of real numbers defined by

$$\gamma_1 > 0\,, \quad \gamma_n = C_\gamma\left(\log n\right)^{-\gamma} \quad \text{with} \quad C_\gamma < \infty \quad \text{and} \quad \gamma \in \,]0\,,\,1/2[.$$

Besides, $\eta = (\eta_n)_{n\geq 1}$ is supposed to be independent of X_0 and ξ. The excited tracking control, at time n, is given by

$$U_n \;=\; -\widehat{f}_n(X_n) \;+\; X_{n+1}^* \;+\; \gamma_{n+1}\,\eta_{n+1}. \tag{4.7}$$

The addition of an exciting noise in the control law allows us to obtain the uniform strong consistency for $\widehat{f}_n$ over dilating sets, unreachable when ξ is bounded. Besides, we can specify the convergence rate of $\widehat{f}_n$. Similar persistently excited control is used in the ARMAX framework to obtain the consistency of the least squares estimator (see Caines [6]).

Theorem 4.1. *Assume that* [*A*3] *and* [*A*4] *hold.*

1. *Then there exists a finite constant M such that for all initial law and any integer m,*

$$\limsup_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n}\|X_k\|^m \;\leq\; M\,, \quad a.s.$$

2. *Let us take* $\alpha \in \,]0\,,\,1/2d[$. *Then, for some finite constant* A, *we have*

$$\liminf_{n\to\infty}\left(\inf_{\|x\|\le\, A(\log n)^{1/2-\gamma}}\ \frac{1}{n^{1-\delta}(\log n)^{\gamma d}}\ H_{n-1}(x)\right) >\ 0,\ \ a.s.,$$

and

$$\sup_{\|x\|\le A(\log n)^{1/2-\gamma}} \left\|\widehat{f}_n(x)-f(x)\right\| \overset{a.s.}{=} o\left(n^{-\delta}\right)$$

with $\delta \in \,]0\,,\,1/2-\alpha d[$. *Moreover, the tracking is asymptotically optimal in quadratic mean*

$$\frac{1}{n}\sum_{k=1}^{n}\|X_k-X_k^*\|^2 \xrightarrow[n\to\infty]{a.s.} \text{trace}(\Gamma),$$

$$\frac{1}{n}\sum_{k=1}^{n}\|X_k-X_k^*-\xi_k\|^2 \overset{a.s.}{=} O\left((\log n)^{-2\gamma}\right)$$

and $\widehat{\Gamma}_n \xrightarrow[n\to\infty]{a.s.} \Gamma\,$.

Proof. The proof is given in Appendix C. The result of almost sure uniform dilated convergence of $\widehat{f}_n$ is obtained using part 1 of Theorem 3.1 in which assumption (3.2) is verified with $v_n = O\left((\log n)^{1/2-\gamma}\right)$ and $w_n = n^{1-\delta}\,(\log n)^{\gamma d}$. □

Now let us study control law (1.6).

4.2. Control with a priori knowledge on the model. In this section, we build a control law similar to the one proposed by Portier [30]. In this reference, an adaptive control algorithm, built on $\widehat{f}_n$ and using a priori knowledge about function f, was studied by simulation. The author shows that its adaptive control law gives very satisfactory results: function f is well estimated in the domain where the observations are clustered and this estimation leads to a good tracking of the given reference trajectory. However, no theoretical results prove that the tracking is optimal, though it appeared through the simulation results (see also Najim, Oppenheim, and Portier [26], Portier and Oppenheim [29]).

Let us introduce the knowledge we need on model (1.1). Assume there is a function f^* such that we have the following.

Assumption [*A*5]. *Function* f^* *is continuous and*

$$\exists\ a_f \in \left[\,0\,,\,1/2\,\right[,\exists\ A_f \in \left]\,0\,,\,\infty\right[,\forall x\in\mathbb{R}^d,\quad \|f(x)-f^*(x)\|\ \le\ a_f\,\|x\| +\ A_f.$$

Then we can build an adaptive control law which first ensures the stability of the closed-loop model and finally possesses the optimality property (4.5).

Construction of U_n. Let $(X_n^*)_{n\ge1}$ be a bounded deterministic given reference trajectory. At time n, the adaptive tracking control with a priori knowledge is given by

$$U_n\ =\ -\widehat{f}_n(X_n)\mathbb{1}_{E_n}(X_n)\ -\ f^*(X_n)\mathbb{1}_{\overline{E}_n}(X_n)\ +\ X_{n+1}^*, \tag{4.8}$$

where E_n is the set $\{x\in\mathbb{R}^d\,;\,\|\,\widehat{f}_n(x)-f^*(x)\,\|\le\, b_f\,\|x\|+B_f\}$ with $b_f \in\,]a_f\,,\,1-a_f[$ and $B_f \in\,]A_f\,,\,\infty[\ ;\ \overline{E}_n$ denotes the complementary set of E_n.

Function f^* and constants a_f and A_f characterize the a priori knowledge we have on model (1.1). From a practical point of view, knowing function f^* is more important (see Portier and Oppenheim [29]) than the assumption on constant a_f (and then on b_f) since the value of B_f can be chosen arbitrarily large. Using control law (4.8) in model (1.1), we obtain the following results.

Theorem 4.2. *Assume that* [*A*1], [*A*2], [*A*3], *and* [*A*5] *hold and that the probability density function* p *is strictly positive.*

1. *Then, there exists a finite constant* M *such that for any initial law,*

$$\limsup_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} \|X_k\|^m \ \leq\ M\,, \quad a.s.$$

2. *Let us take* $\alpha \in \left]0\,,\,1/2d\right[$. *If there is a sequence of positive real numbers* $(v_n)_{n\geq 1}$ *increasing to infinity such that* $v_n = O\left(n^{\nu}\right)$ *with* $\nu > 0$ *and*

$$\left(\inf\left\{p(z)\ ;\ \|z\| \leq v_n + R\right\}\right)^{-1} \ = \ \inf\left(o\left(n^{\alpha}\right)\,,\,O\left(n^{1-s}\right)\right)$$

for some constant $R < \infty$ *and with* $s \in \left]1/2 + \alpha d\,,\,1\right[$, *then we have the optimality of the tracking, i.e.,*

$$\frac{1}{n}\sum_{k=1}^{n} \|X_k - X_k^*\|^2 \ \ \xrightarrow[n\to\infty]{a.s.} \ \ \mathrm{trace}(\Gamma),$$

and $\widehat{\Gamma}_n \xrightarrow[n\to\infty]{a.s.} \Gamma\,.$

Proof. The proof is given in Appendix D. □

When ξ is a Gaussian white noise, we can specify the convergence rate of the optimality.

Corollary 4.3. *Assume that* [*A*1], [*A*3], *and* [*A*5] *hold and that* ξ *is a Gaussian white noise with zero mean and invertible covariance matrix* Γ. *Then, for* $\alpha \in \left]0\,,\,1/2d\right[$, *any initial law and any* $m > 2$,

$$\frac{1}{n}\sum_{k=1}^{n} \|X_k - X_k^* - \xi_k\|^2 \ \overset{a.s.}{=}\ O\left((\log n)^{2-m}\right).$$

Remark. The convergence rate is smaller than the one we have in the ARX framework, which is of (log n)/n (for example, see Lai and Wei [23], [24], Guo [17], and also Bercu [4]).

Proof. Noise ξ_n being Gaussian, [*A*2] is fulfilled and ξ_n has any moment of order m. Thus, part 1 of Theorem 4.2 holds for any m. Besides, part 2 of Corollary 3.2 gives

$$\sup_{\|x\|\,\leq\, A(\log\ n)^{1/2}} \left\|\widehat{f}_n(x) - f(x)\right\| \ \ \overset{a.s.}{=}\ \ O\left(n^{-\lambda}\right), \tag{4.9}$$

where $\lambda = \inf(\alpha\,,\,1-s) - A^2/\lambda_{\min}(\Gamma)$ with $A > 0$ and $A^2/\lambda_{\min}(\Gamma) < \inf(\alpha\,,\,1-s)$. Finally, from (D.13) (see Appendix D)

$$\begin{aligned}\sum_{k=1}^{n} \|X_k - X_k^* - \xi_k\|^2 \overset{a.s.}{=} O\left(\sum_{k=1}^{n-1} v_k^{2-m}\right) \ + \ O\left(n\,v_n^{2-m}\right) \\ + \ O\left(\sum_{k=1}^{n-1}\left(\sup_{\|x\|\,\leq\, v_k}\left\|\widetilde{f}_k(x)\right\|\right)^2\right),\end{aligned} \tag{4.10}$$

and we easily derive the corollary, setting $v_k = A\,(\log\ k)^{1/2}$ in the last equation. □

Appendix A. In this first appendix, we give three useful lemmas and their proofs.

Lemma A.1. *Let $(\phi_n)_{n\geq 1}$ be a sequence of random vectors. Assume that there are $\mu > 0$ and a finite constant M such that $\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \|\phi_k\|^{\mu} \leq M$.*

1. *Then, for R large enough, $\liminf_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}_{\{\|\phi_k\|\leq R\}} > 0$.*
2. *Let $(v_n)_{n\geq 1}$ be a sequence of positive real numbers increasing to infinity. Then,*

$$\forall\, b \in [0\,,\, \mu[,\quad \frac{1}{n}\sum_{k=1}^{n} \|\phi_k\|^{b}\, \mathbb{1}_{\{\,\|\phi_k\| > v_k\}} \;=\; O\left(\frac{1}{n}\sum_{k=1}^{n} v_k^{b-\mu}\right) + O\Big(v_n^{b-\mu}\Big).$$

Proof. Let $R > 0$. Since

$$\text{(A.1)}\qquad \frac{1}{n}\sum_{k=1}^{n} \|\phi_k\|^{\mu} \geq \frac{1}{n}\sum_{k=1}^{n} \|\phi_k\|^{\mu}\, \mathbb{1}_{\{\|\phi_k\|>R\}} \;\geq\; \frac{R^{\mu}}{n}\sum_{k=1}^{n} \mathbb{1}_{\{\|\phi_k\|>R\}},$$

we easily deduce that

$$\text{(A.2)}\qquad \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{\|\phi_k\|\leq R\}} \geq 1 - \frac{M}{R^{\mu}},$$

and part 1 holds for R large enough. Let $b \in [0\,,\,\mu[$. For $k \geq 1$,

$$\|\phi_k\|^{b}\, \mathbb{1}_{\{\,\|\phi_k\| > v_k\}} = \|\phi_k\|^{\mu}\; \|\phi_k\|^{b-\mu}\, \mathbb{1}_{\{\,\|\phi_k\| > v_k\}} \;\leq\; \|\phi_k\|^{\mu}\, v_k^{b-\mu}.$$

Thus, if we set $S_n = \sum_{k=1}^{n} \|\phi_k\|^{\mu}$ and $S_0 = 0$, we obtain that

$$\text{(A.3)}\qquad \begin{aligned} \frac{1}{n}\sum_{k=1}^{n} \|\phi_k\|^{b}\, \mathbb{1}_{\{\|\phi_k\|>v_k\}} &\leq \frac{1}{n}\sum_{k=1}^{n} v_k^{b-\mu}\Big(S_k - S_{k-1}\Big) \\ &= \frac{1}{n}\sum_{k=1}^{n} \Big(v_k^{b-\mu} - v_{k+1}^{b-\mu}\Big)\, S_k \;+\; v_{n+1}^{b-\mu}\,\frac{S_n}{n}. \end{aligned}$$

Then using the lemma's assumption, we derive that

$$\text{(A.4)}\qquad \begin{aligned} \frac{1}{n}\sum_{k=1}^{n} \|\phi_k\|^{b}\, \mathbb{1}_{\{\|\phi_k\|>v_k\}} &= O\Big(\frac{1}{n}\sum_{k=1}^{n} k\,\Big(v_k^{b-\mu} - v_{k+1}^{b-\mu}\Big)\Big) \;+\; O\Big(v_{n+1}^{b-\mu}\Big) \\ &= O\Big(\frac{1}{n}\sum_{k=1}^{n} v_k^{b-\mu}\Big) \;+\; O\Big(v_n^{b-\mu}\Big), \end{aligned}$$

which establishes part 2. □

Now we give a lemma useful for establishing a result of almost sure uniform dilated convergence for the square integrable martingales index-linked by $x \in \mathbb{R}^d$ which appear in the rest of the paper.

Lemma A.2. *For $x \in \mathbb{R}^d$, let us consider*

$$\text{(A.5)}\qquad M_n(x) = \sum_{i=1}^{n} i^{\lambda}\,\Big(K\Big(i^{\alpha}(X_i - x)\Big) \,-\, \mathbb{E}\Big[K\Big(i^{\alpha}(X_i - x)\Big)\,/\,\mathcal{F}_{i-1}\Big]\Big),$$

where $\lambda \in\,]0\,,\,1/2[\,$, $\alpha \in\,]0\,,\,1/2d[\,$, $(X_n)_{n\geq 0}$ is given by model (1.1), and K is the kernel of the nonparametric estimator. Assume that [A2] and [A3] hold. Then, for $A < \infty$, $\nu > 0$, and some $t \in\,]1/2 \,+\, \lambda\,,\,1[$,

$$\text{(A.6)}\qquad \sup_{\|x\|\leq A\, n^{\nu}} \|M_n(x)\| \overset{a.s.}{=} o\left(n^{t}\right).$$

Proof. The proof uses a theorem due to Oulidi that we can find with its proof in Duflo [13, Theorem 6.4.34, p. 220]. For the sake of completeness, we recall this theorem below.

Theorem. *Let us consider a probability space with a filtration $\mathbb{F}$ and a sequence $(M_n(x))_{n\geq 0, x\in\mathbb{R}^d}$ such that $\forall x$, $M_n(x)$ is a square-integrable martingale adapted to $\mathbb{F}$ and $\forall n$, the trajectories $x \to M_n(x)$ are continuous. Suppose there exists a sequence (s_n) adapted to $\mathbb{F}$ which tends a.s. to infinity and is such that, denoting $\Delta_{n+1}(x) = M_{n+1}(x) - M_n(x)$ and $h(x) = (2x \log\log x)^{1/2}$, there exist two constants $0 < \delta < \gamma$ and constants a, b, c, and d with the following properties:*
P1. *a.s., $\langle M(0)\rangle_n \leq a\, s_{n-1}^2$ and $|\Delta_n(0)| \leq b\, s_{n-1}^2\,(h(s_{n-1}^2))^{-1}$;*
P2. *for any pair (x, y), a.s.,*

$$\langle M(x) - M(Y)\rangle_n \leq c\, s_{n-1}^2\, \|x - y\|^\gamma,$$
$$|\Delta_n(x) - \Delta_n(y)| \leq d\, \|x - y\|^\delta\, s_{n-1}^2\,(h(s_{n-1}^2))^{-1}.$$

Then we have the following result. Let v be any function from $\mathbb{R}_+$ to itself which is increasing to infinity and such that for $\theta > 1$, $v\left(\theta^{n+1}\right) = O\left(v(\theta^n)\right)$. Then, for any $\beta > \sup(\delta\,, \gamma - \delta)$, a.s.,

$$\sup_{\|x\|\leq v(s_{n-1}^2)} \|M_n(x)\| = o\Big(v^\beta(s_{n-1}^2)\, h(s_{n-1}^2)\Big)\,. \qquad \square$$

Let us show that $M_n(x)$ defined by (A.5) matches the different assumptions of this theorem. To this aim, let us denote for $n \geq 1$ and $x \in \mathbb{R}^d$, $\Delta_n(x) = M_n(x) - M_{n-1}(x)$, and $\langle M(x)\rangle_n = \sum_{k=1}^n \mathbb{E}\left[\Delta_k^2(x) \,/\, \mathcal{F}_{k-1}\right]$, where $M_0(\,.\,) \equiv 0$. For $x \in \mathbb{R}^d$ and $n \geq 1$, $M_n(x)$ is a square integrable martingale adapted to $\mathcal{F}$. Let us denote $s_n^2 = n^{1+2\lambda+\alpha\delta}$ for some $\delta > 0$. Since K is bounded (cf. Assumption [$A3$]),

$$|\Delta_n(0)| \leq \text{cte } n^\lambda \;\leq\; \text{cte } s_n^2\,(h(s_n^2))^{-1}, \tag{A.7}$$

where $h(x) = (2x\, \log\log\, x)^{1/2}$. In addition, since p is bounded (cf. Assumption [$A2$]),

$$\begin{aligned} \langle M(0)\rangle_n &\leq \sum_{i=1}^n \mathbb{E}\left[i^{2\lambda}\, K^2\left(i^\alpha\, X_i\right) / \mathcal{F}_{i-1}\right] \\ &\leq \sum_{i=1}^n i^{2\lambda-\alpha d} \int K^2(t)\, p\Big(i^{-\alpha}t - f(X_{i-1}) - U_{i-1}\Big)\, dt \\ &\leq \|\, p\,\|_\infty\, \|K\|_\infty \sum_{i=1}^n i^{2\lambda-\alpha d} \;\leq\; \text{cte } n^{1+2\lambda-\alpha d} \;\leq\; \text{cte } s_n^2. \end{aligned} \tag{A.8}$$

Let $x, y \in \mathbb{R}^d$. Since K is bounded and Lipschitz, we have for any $\delta \in\,]0\,,\,1[$,

$$\begin{aligned} |\Delta_n(x) - \Delta_n(y)| &\leq \text{cte } n^{\lambda(1-\delta/2)}\, |\Delta_n(x) - \Delta_n(y)|^{\delta/2} \\ &\leq \text{cte } \|x - y\|^{\delta/2}\, n^{\lambda+\alpha\delta/2} \;\leq\; \text{cte } \|x - y\|^{\delta/2}\, s_n^2\,(h(s_n^2))^{-1}, \end{aligned} \tag{A.9}$$

$$\begin{aligned} \big\langle M(x) - M(y)\big\rangle_n &\leq \sum_{i=1}^n i^{2\lambda}\, \mathbb{E}\left[\Big(K\left(i^\alpha(X_i - x)\right) - K\left(i^\alpha(X_i - y)\right)\Big)^2 / \mathcal{F}_{i-1}\right] \\ &\leq \text{cte } \|x - y\|^\delta\, n^{1+2\lambda+\alpha\delta} \;\leq\; \text{cte } \|x - y\|^\delta\, s_n^2. \end{aligned} \tag{A.10}$$

Then, with (A.7)–(A.10), assumptions of Oulidi's theorem are fulfilled and therefore, for any $\omega > \delta/2 > 0$,

$$\sup_{\|x\| \leq v(\sigma_n^2)} \|M_n(x)\| \overset{a.s.}{=} o\Big(v^{\omega}(s_n^2)\, h(s_n^2)\Big), \tag{A.11}$$

where v is a real valued function increasing to infinity, such that $v\left(\theta^{n+1}\right) = O\left(v(\theta^n)\right)$ for $\theta > 1$. In particular, if we take $v(n) = A\, n^{\nu/(1+2\lambda+\alpha\delta)}$ for $A < \infty$ and $\nu > 0$, we obtain (A.6). This closes the proof. □

Lemma A.3 below collects some classic and sparse results that we can find in Duflo [13] or in Meyn and Tweedie [25]. Part 1 of Lemma A.3 will be used to establish part 1 of Theorems 4.1 and 4.2, and part 2 will ensure that some of martingales we consider are square integrable.

LEMMA A.3. *Let $0 < a < 1$ and $b < \infty$, and let $Z = (Z_n)$ and $\varepsilon = (\varepsilon_n)$ be two positive sequences of random variables. Assume that, for $n \geq 1$,*

$$Z_n \leq a\, Z_{n-1} \;+\; b \;+\; \varepsilon_n. \tag{A.12}$$

1. *If ε is a sequence of independent, identically distributed positive random variables with a moment of order $m \geq 1$, then $\sup_{k \leq n} Z_k \;=\; o(n^{1/m})$, a.s. and there is a finite constant M such that*

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Z_k^m \;\leq\; M\,, a.s.$$

2. *In addition, if Z_0 has a moment of order m, then $\sup_n \mathbb{E}\left[Z_n^m\right] \;<\; \infty$.*

For the sake of completeness, let us recall a sketch of proof.

Proof. From (A.12), we easily deduce that $Z_n \;\leq\; a^n\, Z_0 \;+\; \text{cte} \;+\; \sum_{k=1}^{n} a^{n-k}\varepsilon_k$. Then, $Z_n \overset{a.s.}{=} O(\sup_{k\leq n} \varepsilon_k)$ and since ε has a moment of order m, then $Z_n \overset{a.s.}{=} o(n^{1/m})$. In addition, for $a < a_1 < 1$, we have $Z_n^m \;\leq\; a_1\, Z_{n-1}^m \;+\; \text{cte} \;+\; \text{cte}\; \varepsilon_n^m$ and we derive that $(1 - a_1) \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Z_k^m \;\leq\; \text{cte} \;+\; \text{cte} \limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k^m$. Then, with the assumptions on ε, $\limsup_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k^m \;\leq\;$ cte and part 1 is established. Part 2 follows from the fact that $Z_n^m \;\leq\; a_1^n\, Z_0^m \;+\; \text{cte} \;+\; \text{cte}\; \sum_{k=1}^{n} a_1^{n-k} \varepsilon_k^m$. □

Appendix B. This appendix is concerned with the proof of Theorem 3.1. Let us denote $\widetilde{f}_n(x) = \widehat{f}_n(x) - f(x)$. Starting from the definition of $\widehat{f}_n(x)$, we can write

$$\widetilde{f}_n(x) \;=\; \frac{M_n(x) + R_{n-1}(x)}{H_{n-1}(x)} \mathbb{1}_{\{H_{n-1}(x) \neq 0\}} \;-\; f(x) \mathbb{1}_{\{H_{n-1}(x) = 0\}} \tag{B.1}$$

$$\text{with} \quad M_n(x) \;=\; \sum_{i=0}^{n-1} i^{\alpha d}\, K\Big(i^{\alpha}(X_i - x)\Big) \xi_{i+1},$$

$$R_{n-1}(x) \;=\; \sum_{i=0}^{n-1} i^{\alpha d}\, K\Big(i^{\alpha}(X_i - x)\Big) \Big(f(X_i) - f(x)\Big),$$

and easily derive that for any $x \in \mathbb{R}^d$,

$$\Big\|\widehat{f}_n(x) - f(x)\Big\| \leq \|f(x)\| \;+\; \sup_{k\leq n} \|\xi_k\| \;+\; \frac{\|R_n(x)\|}{H_{n-1}(x)} \mathbb{1}_{\{H_{n-1}(x) \neq 0\}}. \tag{B.2}$$

The proof of Theorem 3.1 consists of studying separately $M_n(x)$, $R_{n-1}(x)$, and $H_{n-1}(x)$ and then combining the results.

Study of $M_n(x)$. Since process $(\xi_n)_{n\geq 1}$ has a finite moment of order $m > 2$ and since for any x, $y \in \mathbb{R}^d$ and any $\delta \in]0\,,\,1[$,

$$n^{\alpha d}\, \left|K\left(n^{\alpha}(X_n - x)\right)\right| \leq \text{cte } n^{\alpha d}, \tag{B.3}$$

$$n^{\alpha d}\, \left|K\left(n^{\alpha}(X_n - x)\right) - K\left(n^{\alpha}(X_n - y)\right)\right| \leq \text{cte } n^{\alpha d + \alpha\delta}\, \|x - y\|^{\delta}, \tag{B.4}$$

the assumptions of the following corollary, also due to Oulidi [28] and which we can find with its proof in Duflo [13, Corollary 6.4.35, p. 223], are fulfilled with, for instance, $T_n = n^{\alpha d + \alpha\delta}$.

Corollary. *Let (ε_n) be a noise of dimension 1 adapted to a filtration $\mathbb{F} = (\mathcal{F}_n)_{n\geq 0}$ such that, for some $\alpha > 0$, $\sup_n \mathbb{E}[|\varepsilon_{n+1}|^{2+2\alpha} \,/\, \mathcal{F}_n] < \infty$, a.s. We consider a sequence $(Y_n(x))_{n\geq 0, x\in\mathbb{R}^d}$ such that $\forall\, x$, $Y_n(x)$ is adapted to $\mathbb{F}$ and, $\forall\, n$, the process $(Y_n(.))$ is continuous. Finally, we suppose we have a positive sequence (T_n) adapted to $\mathbb{F}$, such that $|T_n(0)| \leq T_n$, a.s., and there exist two constants $\delta > 0$ and $a < \infty$ such that, for any pair (x, y), $|Y_n(x) - Y_n(y)| \leq a\, \|x - y\|^{\delta}\, T_n$, a.s. For $M_n(x) = \sum_{k=1}^n Y_{k-1}(x)\,\varepsilon_k$ and $\tau_n = \sum_{k=0}^n T_k^2$, we assume that, a.s., $\tau_\infty = \infty$ and $\sum T_n^{2+2\alpha}\, (\tau_n)^{-1-\alpha}\, (\log\log\, n)^{\alpha} < \infty$. Then, for $\beta > \delta$ and any function $v : \mathbb{R}_+ \to \mathbb{R}_+$ increasing to infinity and such that for $\theta > 1$, $v\left(\theta^{n+1}\right) = O\left(v(\theta^n)\right)$, we have*

$$\sup_{\|x\|\leq\, v(\tau_{n-1})} \|M_n(x)\| \overset{a.s.}{=} o\Big(v^{\beta}(\tau_{n-1})\, (2\tau_{n-1}\, \log\log\, \tau_{n-1})^{1/2}\Big). \qquad \square$$

Taking $v(n) = A\, n^{\nu/(1+2\alpha\, d + 2\alpha\delta)}$ (for any $A < \infty$ and $\nu > 0$) in the corollary above, we obtain, for $s > 1/2 + \alpha d$

$$\sup_{\|x\|\leq\, A\, n^{\nu}} \|M_n(x)\| \overset{a.s.}{=} o(n^s). \tag{B.5}$$

In addition, let us note that the restriction $\alpha \in\,]0\,,\, 1/2d[$ derives from the fact that in that which follows the real s must be chosen as $s < 1$.

Study of $R_n(x)$. Since function f is Lipschitz-continuous (cf. [$A1$]),

$$\|R_n(x)\| \leq r_f \sum_{i=1}^n i^{\alpha d} K\Big(i^{\alpha}(X_i - x)\Big)\, \|X_i - x\|\,. \tag{B.6}$$

In addition, since K has a compact support (cf. [$A3$]), there exists a finite constant c_K such that $K(y) = 0$ for $\|y\| \geq c_K$. Then

$$\begin{aligned} \|R_n(x)\| \;&\leq\; r_f \sum_{i=1}^n i^{\alpha d - \alpha} K\Big(i^{\alpha}(X_i - x)\Big) i^{\alpha}\, \|X_i - x\|\, \mathbb{1}_{\left\{i^{\alpha}\, \|X_i - x\| \,\leq\, c_K\right\}} \\ &\overset{a.s.}{=} O\Big(T_n(x)\Big) \end{aligned} \tag{B.7}$$

with $T_n(x) = \sum_{i=1}^n i^{\alpha d - \alpha} K\Big(i^{\alpha}(X_i - x)\Big)$. In addition, since $T_n(x) \leq H_n(x)$, there exists a finite constant c_f such that for any $x \in \mathbb{R}^d$,

$$\frac{\|R_n(x)\|}{H_{n-1}(x)} \mathbb{1}_{\{H_{n-1}(x) \neq 0\}} \leq c_f. \tag{B.8}$$

Then, by combining this result with (B.2), we prove the following lemma which gives a first result on $\|\widehat{f}_n(x) - f(x)\|$.

LEMMA B.1. *Assume that* [A1] *and* [A3] *hold. Let us denote* $\xi_n^{\#} = \sup_{k\leq n}\|\xi_k\|$. *Then, there exists* $c_f<\infty$ *such that* $\forall\, x\in\mathbb{R}^d$ *and* $\forall\, n\geq 1$,

$$\left\|\widehat{f}_n(x)-f(x)\right\| \;\leq\; c_f \;+\; \|f(x)\| \;+\; \xi_n^{\#}, \qquad a.s.$$

We are left with the study of $T_n(x)$. Let us write $T_n(x) = M_n^T(x) + T_n^c(x)$ with

$$\begin{aligned} M_n^T(x) &= \sum_{i=1}^n i^{\alpha d-\alpha}\Big(K\Big(i^\alpha(X_i-x)\Big) - \mathbb{E}\left[K\Big(i^\alpha(X_i-x)\Big)/\mathcal{F}_{i-1}\right]\Big),\\ T_n^c(x) &= \sum_{i=1}^n i^{\alpha d-\alpha}\;\mathbb{E}\left[K\Big(i^\alpha(X_i-x)\Big)/\mathcal{F}_{i-1}\right]\\ &= \sum_{i=1}^n i^{-\alpha}\int K(t)\,p\Big(i^{-\alpha}t+x-f(X_{i-1})-U_{i-1}\Big)dt. \end{aligned}$$

Since p is bounded (cf. [$A2$]) and $\int K(t)\,dt=1$ (cf. [$A3$]),

$$\sup_{x\in\mathbb{R}^d}\|T_n^c(x)\| \overset{a.s.}{=} O\left(n^{1-\alpha}\right). \tag{B.9}$$

For $x\in\mathbb{R}^d$ and $n\geq 1$, $M_n^T(x)$ is a square integrable martingale for which we can apply Lemma A.2 with $\lambda=\alpha d-\alpha$. Then, for $A<\infty$, $\nu>0$, and $s'>\frac{1}{2}+\alpha d-\alpha$,

$$\sup_{\|x\|\leq A\,n^\nu}\left\|M_n^T(x)\right\| \overset{a.s.}{=} o\left(n^{s'}\right). \tag{B.10}$$

Moreover, since $\alpha\in\,]0\,,\,1/2d[$, the real s' can be chosen such that $s'<1-\alpha$. Therefore, from (B.9) and (B.10), we obtain that for $A<\infty$ and $\nu>0$

$$\sup_{\|x\|\leq A\,n^\nu}\|T_n(x)\| \overset{a.s.}{=} O\left(n^{1-\alpha}\right), \tag{B.11}$$

and therefore

$$\sup_{\|x\|\leq A\,n^\nu}\|R_n(x)\| \overset{a.s.}{=} O\left(n^{1-\alpha}\right). \tag{B.12}$$

Finally, as soon as $v_n=O(n^\nu)$, by combining result (B.5) and result (B.12) with assumption (3.2), we derive part 1 since

$$\sup_{\|x\|\leq v_n}\frac{\|M_n(x)\|}{H_{n-1}(x)} \overset{a.s.}{=} o\left(\frac{n^s}{w_n}\right), \tag{B.13}$$

$$\sup_{\|x\|\leq v_n}\frac{\|R_{n-1}(x)\|}{H_{n-1}(x)} \overset{a.s.}{=} O\left(\frac{n^{1-\alpha}}{w_n}\right). \tag{B.14}$$

Proof of part 2. To establish part 2, we have only to prove that

$$\liminf_{n\to\infty}\left(\inf_{\|x\|\leq v_n}\frac{1}{n\,m_n}\,H_{n-1}(x)\right) \;>\; 0, \text{ a.s.} \tag{B.15}$$

We study $H_n(x)$ by proceeding as for $T_n(x)$. For $x\in\mathbb{R}^d$, let us set

$$H_n(x) \;=\; M_n^H(x) \;+\; \Big(H_n^c(x) \;-\; J_n(x)\Big) \;+\; J_n(x) \tag{B.16}$$

with

$$\begin{aligned}
M_n^H(x) &= \sum_{i=1}^{n} i^{\alpha d} \left(K\Big(i^\alpha (X_i - x)\Big) \; - \; \mathbb{E}\left[K\Big(i^\alpha (X_i - x)\Big) \,/\, \mathcal{F}_{i-1} \right] \right), \\
H_n^c(x) &= \sum_{i=1}^{n} i^{\alpha d} \; \mathbb{E}\Big[K\left(i^\alpha (X_i - x)\right) / \mathcal{F}_{i-1} \Big] \\
&= \sum_{i=1}^{n} \int K(t)\, p\Big(i^{-\alpha} t + x - f(X_{i-1}) - U_{i-1}\Big)\, dt, \\
J_n(x) &= \sum_{i=1}^{n} \; p\Big(x - f(X_{i-1}) - U_{i-1}\Big).
\end{aligned}$$

For $x \in \mathbb{R}^d$ and $n \geq 1$, $M_n^H(x)$ is a square integrable martingale. Then, by Lemma A.2 with $\lambda = \alpha d$, we derive that for $A < \infty$, $\nu > 0$, and $s'' > \frac{1}{2} + \alpha d$,

$$\text{(B.17)} \qquad \sup_{\|x\| \leq A\, n^\nu} \left\| M_n^H(x) \right\| \; \overset{a.s.}{=} \; o\left(n^{s''} \right).$$

Since $\|Dp\|_\infty < \infty$ (cf. [$A2$]) and $\int \|t\|\, K(t)\, dt < \infty$ (cf. [$A3$]),

$$\text{(B.18)} \qquad \sup_{x \in \mathbb{R}^d} \| H_n^c(x) - J_n(x) \| \; \overset{a.s.}{=} \; O\left(n^{1-\alpha}\right).$$

Let $R < \infty$. For $x \in \mathbb{R}^d$ such that $\|x\| \leq v_n$, we have

$$\text{(B.19)} \qquad \|p\|_\infty \; \geq \; \frac{1}{n}\, J_n(x) \; \geq \; \frac{m_n}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{\|f(X_k) + U_k\| \leq R\}},$$

where $m_n = \inf\{p(z) \; ; \; \|z\| \leq v_n + R\}$. Since $m > 2$, $\limsup_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \|\xi_k\|^2 < \infty$. Combining this last result with assumption (3.4), we derive that

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \|f(X_k) + U_k\|^2 < \; \infty,$$

and then by applying part 1 of Lemma A.1, we obtain

$$\text{(B.20)} \qquad \liminf_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{\|f(X_k) + U_k\| \leq R\}} \; > \; 0, \; \text{ a.s.}$$

Then, combining (B.17), (B.18), (B.19), and (B.20) gives (B.15).

Appendix C. This appendix is concerned with the proof of Theorem 4.1.

1. *Stability of* (X_n). Using (4.7) in model (1.1), the following equation holds:

$$\text{(C.1)} \qquad X_{n+1} = -\widetilde{f}_n(X_n) \; + \; X_{n+1}^* \; + \; \xi_{n+1} \; + \; \gamma_{n+1} \eta_{n+1}.$$

Since ξ and $(X_n^*)_{n \geq 1}$ are bounded and f is Lipschitz, we deduce from Lemma B.1 that

$$\text{(C.2)} \qquad \|X_{n+1}\| \leq \; r_f \, \|X_n\| \; + \; A_2 \; + \; \gamma_{n+1} \; \|\eta_{n+1}\|$$

with $A_2 < \infty$. Then part 1 is obtained using Lemma A.3.

2. *Consistency of* $\widehat{f}_n$. To prove the almost sure uniform dilated convergence of $\widehat{f}_n$, we use part 1 of Theorem 3.1. Therefore, we must verify assumption (3.2) and, more precisely, find two sequences $(v_n)_{n\geq 1}$ and $(w_n)_{n\geq 1}$ such that

$$\liminf_{n\to\infty}\left(\inf_{\|x\|\leq v_n}\frac{1}{w_n}\,H_{n-1}(x)\right) > \ 0, \text{ a.s.} \tag{C.3}$$

For $n \geq 0$, let us set $\Pi_n = -\widetilde{f}_n(X_n) + X^*_{n+1}$ and consider the following decomposition:

$$H_n(x) \ = \ M^H_n(x) \ + \ \Big(H^c_n(x) - J_n(x)\Big) \ + \ \Big(J_n(x) - L_n(x)\Big) \ + \ L_n(x) \tag{C.4}$$

with

$$\begin{aligned}
M^H_n(x) &= H_n(x) \ - \ H^c_n(x),\\
H^c_n(x) &= \sum_{i=1}^{n} \ i^{\alpha d}\ \mathbb{E}\left[K\Big(i^{\alpha}\,(X_i - x)\Big)/\mathcal{F}_{i-1}\right]\\
&= \sum_{i=1}^{n} \ i^{\alpha d}\ \mathbb{E}\left[K\Big(i^{\alpha}\,(\Pi_{i-1} + \gamma_i\,\eta_i + \xi_i - x)\Big)/\mathcal{F}_{i-1}\right]\\
&= \sum_{i=1}^{n} \ \gamma_i^{-d}\iint K(v)\,p(u)\,p_\eta\left(\gamma_i^{-1}\left(i^{-\alpha}v + x - \Pi_{i-1} - u\right)\right)du\,dv,\\
J_n(x) &= \sum_{i=1}^{n} \ \gamma_i^{-d}\int p(u)\,p_\eta\left(\gamma_i^{-1}\,(x - \Pi_{i-1} - u)\right)\,du,\\
L_n(x) &= \sum_{i=1}^{n} \ \gamma_i^{-d}p_\eta\left(\gamma_i^{-1}\,(x - \Pi_{i-1})\right).
\end{aligned}$$

In this part, $\mathcal{F}_n = \sigma\,(X_0, U_0, \xi_1, \eta_1, \ldots, \xi_n, \eta_n)$. We now proceed as for $H_n(x)$ in Appendix B. For $x \in \mathbb{R}^d$ and $n \geq 1$, $M^H_n(x)$ is a square integrable martingale adapted to $\mathcal{F}$. Then, by Lemma A.2 with $\lambda = \alpha d$, we obtain that for $A < \infty$, $\nu > 0$, and $s \in \,]1/2 + \alpha d\,,\,1[$,

$$\sup_{\|x\|\leq\,A\,n^{\nu}}\left\|M^H_n(x)\right\| \ \overset{a.s.}{=} \ o\,(n^s)\,. \tag{C.5}$$

Since $\|Dp_\eta\| < \infty$, $\int \|v\|\,K(v)\,dv < \infty$ (cf. [$A3$]), and $\int \|u\|\,p(u)\,du < \infty$,

$$\sup_{x\in\mathbb{R}^d}\|H^c_n(x) - J_n(x)\| \ + \ \sup_{x\in\mathbb{R}^d}\|J_n(x) - L_n(x)\| \overset{a.s.}{=} O\left(\sum_{i=1}^{n}\gamma_i^{-d-1}\right). \tag{C.6}$$

We are left with the study of $L_n(x)$. Let $R > 0$. Using Lemma B.1 and Assumption [$A4$], we deduce that $\|\Pi_n\| \ \leq \ \|X_n\| \ + \ N$, with $N < \infty$. Hence,

$$L_n(x) \ \geq \ \sum_{i=1}^{n} \ \gamma_i^{-d}\inf_{\|y\|\leq N+R}p_\eta\left(\gamma_i^{-1}\,(x - y)\right)\mathbb{1}_{\{\|X_{i-1}\|\leq R\}}. \tag{C.7}$$

Since η is a Gaussian white noise with invertible covariance matrix Γ_η, we have

$$p_\eta\left(\gamma_i^{-1}\,(x - y)\right) \ \geq \ \text{cte}\ \exp\left(-\frac{\gamma_i^{-2}\ \|\,x - y\,\|^2}{2\,\lambda_{\min}(\Gamma_\eta)}\right), \tag{C.8}$$

and since $\gamma_n = C_\gamma (\log n)^{-\gamma}$ with $\gamma \in \,]0\,,\, 1/2[$,

$$\text{(C.9)}\quad \inf\left\{ p_\eta\left(\gamma_n^{-1}\,(x-y)\right)\ ;\ \|x\|^2 \le A\,\gamma_n^2\,\log n\,,\, \|y\|\, \le N+R \right\}\ \ge\ \text{cte}\, n^{-\delta},$$

with $\delta = A/2\,\lambda_{\min}(\Gamma_\eta)$. Finally,

$$\text{(C.10)}\qquad \inf_{\|x\|^2 \le A\,\gamma_n^2\,\log n} L_n(x)\ \ge\ \text{cte}\ \sum_{i=2}^{n} i^{-\delta}\,(\log i)^{-\gamma d}\,\mathbb{1}_{\{\|X_{i-1}\|\,\le\,R\}}.$$

In addition, by part 1 of Theorem 4.1, there exists a finite constant M such that $\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\|X_{i-1}\|^2\ \le\ M$, a.s. Then, from part 1 of Lemma A.1, we deduce that

$$\text{(C.11)}\qquad \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\|X_{i-1}\|\,\le\,R\}} >\ 0,\ \text{a.s.},$$

and therefore

$$\text{(C.12)}\quad \liminf_{n\to\infty}\left(\inf_{\|x\| \le\, (2\,\delta\,\lambda_{\min}(\Gamma_\eta))^{1/2} C_\gamma (\log n)^{1/2-\gamma}} \frac{L_n(x)}{n^{1-\delta}(\log n)^{\gamma d}}\right) >\ 0,\ \text{a.s.}$$

Then, combining (C.5), (C.6), and (C.12), we obtain that for any $\delta \in \,]0\,,\, 1/2 - \alpha d[$

$$\text{(C.13)}\qquad \liminf_{n\to\infty}\left(\inf_{\|x\| \le\, A(\log n)^{1/2-\gamma}} \frac{1}{n^{1-\delta}(\log n)^{\gamma d}}\ H_{n-1}(x)\right) >\ 0,\ \text{a.s.},$$

where $A\ =\ (2\,\delta\,\lambda_{\min}(\Gamma_\eta))^{1/2}\,C_\gamma$. Thus (3.2) is fulfilled and part 2 of Theorem 4.1 is proved.

3. *Optimality of the tracking.* From (C.1), we have

$$\begin{aligned} \left\|X_{n+1} - X_{n+1}^*\right\|^2 &= \left\|\widetilde{f}_n\left(X_n\right)\right\|^2\ 2\left\langle \widetilde{f}_n\left(X_n\right)\,,\, \xi_{n+1} + \gamma_{n+1}\,\eta_{n+1}\right\rangle \\ &\qquad\quad +\ \left\|\xi_{n+1}\right\|^2\ +\ \gamma_{n+1}^2\ \left\|\eta_{n+1}\right\|^2\ +\ 2\,\gamma_{n+1}\left\langle \eta_{n+1}, \xi_{n+1}\right\rangle, \\ \left\|X_n - X_n^* - \xi_n\right\|^2 &\le 2\ \left\|\widetilde{f}_{n-1}\left(X_{n-1}\right)\right\|^2\ +\ 2\,\gamma_n^2\ \left\|\eta_n\right\|^2, \end{aligned}$$

where $\langle\ .\ ,\ .\ \rangle$ denotes the inner product on $\mathbb{R}^d$. Using part 2 of Theorem 4.1, Lemma B.1, and Theorem 3.3, we derive that for any integer m,

$$\text{(C.14)}\quad \frac{1}{n}\sum_{k=1}^{n}\left\|\widetilde{f}_k\left(X_k\right)\right\|^2\ \mathbb{1}_{\left\{\|X_k\|\,\le\, A_\delta(\log k)^{1/2-\gamma}\right\}} \overset{a.s.}{=} o\left(n^{-2\delta}\right),$$

$$\text{(C.15)}\quad \frac{1}{n}\sum_{k=1}^{n}\left\|\widetilde{f}_k\left(X_k\right)\right\|^2\ \mathbb{1}_{\left\{\|X_k\|\,>\, A_\delta(\log k)^{1/2-\gamma}\right\}} \overset{a.s.}{=} O\left((\log n)^{(1/2-\gamma)(2-m)}\right).$$

Hence, for any integer m,

$$\text{(C.16)}\qquad \frac{1}{n}\sum_{k=1}^{n}\left\|\widetilde{f}_k\left(X_k\right)\right\|^2 \overset{a.s.}{=} O\left((\log n)^{-(1/2-\gamma)(m-2)}\right).$$

For $n\ \ge\ 1$, $M_n\ =\ \sum_{k=1}^{n}\langle \widetilde{f}_{k-1}\left(X_{k-1}\right), \gamma_k\,\eta_k + \xi_k\rangle$ is a square integrable martingale. Since $\|\,\langle M\rangle_n\,\| = O(\sum_{k=1}^{n}\|\widetilde{f}_{k-1}\left(X_{k-1}\right)\|^2)$ and $\sum_{k=1}^{n}\|\widetilde{f}_{k-1}\left(X_{k-1}\right)\|^2 \overset{a.s.}{=} o(n)$

by (C.16), we deduce from a strong law of large numbers for the martingales (for example, Duflo [13, Theorem 1.3.15, p. 20]) that for any $\lambda > 0$,

$$\text{(C.17)} \qquad \frac{1}{n}\sum_{k=1}^{n} \langle \widetilde{f}_{k-1}\,(X_{k-1})\ ,\, \gamma_k\,\eta_k + \xi_k\rangle \overset{a.s.}{=} O\left(\frac{(\log n)^{1+\lambda}}{n}\right)^{1/2}.$$

Since η is a Gaussian white noise, $\sum_{k=1}^{n} \|\eta_k\|^2 \overset{a.s.}{=} O(n)$ and then part 2 of Lemma A.1 gives

$$\text{(C.18)} \qquad \frac{1}{n}\sum_{k=1}^{n} \gamma_k^2\,\|\eta_k\|^2 \overset{a.s.}{=} O\left((\log n)^{-2\gamma}\right).$$

To close the proof of Theorem 4.1, we have only to combine all these results and use classical strong law of large numbers for the other terms.

Appendix D. This appendix is concerned with the proof of Theorem 4.2. For $x \in \mathbb{R}^d$, let us denote $\widetilde{f}^*(x) = f(x) - f^*(x)$.

1. *Stability of* (X_n). Let us first rewrite model (1.1) with control (4.8) in the following way:

$$\text{(D.1)} \quad X_{n+1} \;=\; \Big(f^*(X_n) - \widehat{f}_n(X_n)\Big)\,\mathbb{1}_{E_n}(X_n) \;+\; \widetilde{f}^*(X_n) \;+\; X^*_{n+1} + \xi_{n+1}.$$

Then, using Assumption $[A5]$,

$$\text{(D.2)} \qquad \|X_{n+1}\| \;\leq\; \underbrace{(a_f + b_f)}_{<\,1}\|X_n\| \;+\; \underbrace{\left(A_f + B_f + \left\|X^*_{n+1}\right\|\right)}_{<\,\infty} \;+\; \|\xi_{n+1}\|,$$

and applying Lemma A.3 gives part 1 of Theorem 4.2.

2. *Optimality of the control law.* For $k \geq 0$, let us set $\Pi_k = f(X_k) + U_k - X^*_{k+1}$. Then

$$\text{(D.3)} \qquad \left\|X_{k+1} - X^*_{k+1}\right\|^2 = \|\Pi_k\|^2 \;+\; 2\,\Pi_k^T\,\xi_{k+1} \;+\; \|\xi_{k+1}\|^2,$$

where Π_k^T denotes the transpose of Π_k. To prove the optimality, we study the convergence of the sum of each term in the last equation right-hand side.

The study of $(1/n)\sum_{k=0}^{n-1}\|\xi_{k+1}\|^2$ is straightforward. To explore $(1/n)\sum_{k=0}^{n-1}\Pi_k^T \times \xi_{k+1}$, we must face the real difficulty which concerns $(1/n)\sum_{k=0}^{n-1}\|\Pi_k\|^2$.

Step 1: *study of* $(1/n)\sum_{k=0}^{n-1}\|\xi_{k+1}\|^2$. Since ξ has a finite moment of order $m > 2$, we have the regular strong law of large numbers

$$\text{(D.4)} \qquad \frac{1}{n}\sum_{k=0}^{n-1} \|\xi_{k+1}\|^2 \xrightarrow[n\to\infty]{a.s.} \operatorname{trace}(\Gamma).$$

Step 2: *study of* $(1/n)\sum_{k=0}^{n-1}\|\Pi_k\|^2$. To this aim, let us rewrite

$$\begin{aligned}\Pi_k = &\Big(-\widetilde{f}_k(X_k)\mathbb{1}_{E_k}(X_k) + \widetilde{f}^*(X_k)\mathbb{1}_{\overline{E}_k}(X_k)\Big)\,\mathbb{1}_{V_k}(X_k)\\ &+\;\Big(-\widetilde{f}_k(X_k)\mathbb{1}_{E_k}(X_k) + \widetilde{f}^*(X_k)\mathbb{1}_{\overline{E}_k}(X_k)\Big)\,\mathbb{1}_{\overline{V}_k}(X_k),\end{aligned}$$

where V_k denotes the set $\left\{x \in \mathbb{R}^d\,;\, \|x\| \leq v_k\right\}$ and $\overline{V}_k$ its complementary.

- First, using assumption on function f^* (cf. $[A5]$),

$$\text{(D.5)}\sum_{k=1}^{n} \left\|\widetilde{f}^*(X_k)\right\|^2 \mathbb{1}_{\overline{E}_k}(X_k)\mathbb{1}_{\overline{V}_k}(X_k) \overset{a.s.}{=} O\left(\sum_{k=1}^{n}\left(\|X_k\|^2+1\right)\mathbb{1}_{\{\,\|X_k\|\,>\,v_k\}}\right).$$

Moreover, there are two constants c_1, $c_2 > 0$ such that, a.s.,

$$\text{(D.6)}\quad \begin{aligned}\left\|\widetilde{f}_k(X_k)\right\|^2 \mathbb{1}_{E_k}(X_k) &\leq \left\|\widetilde{f}_k(X_k)\right\|^2 \mathbb{1}_{\left\{x\in\mathbb{R}^d\,;\,\|\widetilde{f}_k(x)\|-\|\widetilde{f}^*(x)\|\,\leq\, b_f\|x\|+B_f\right\}}(X_k)\\ &\leq \left\|\widetilde{f}_k(X_k)\right\|^2 \mathbb{1}_{\left\{x\in\mathbb{R}^d\,;\,\|\widetilde{f}_k(x)\|\,\leq\, c_1\|x\|+\ c_2\right\}}(X_k)\\ &\leq \left(c_1\|X_k\|+\ c_2\right)^2\end{aligned}$$

and

$$\text{(D.7)}\quad \sum_{k=1}^{n} \left\|\widetilde{f}_k(X_k)\right\|^2 \mathbb{1}_{E_k}(X_k)\mathbb{1}_{\overline{V}_k}(X_k) \overset{a.s.}{=} O\left(\sum_{k=1}^{n}\left(\|X_k\|^2+1\right)\mathbb{1}_{\{\,\|X_k\|\,>\,v_k\}}\right).$$

Since $\sum_{k=1}^{n}\|X_k\|^m \overset{a.s.}{=} O(n)$ with $m > 2$, part 2 of Lemma A.1 applied to (D.5) and (D.7) leads to

$$\text{(D.8)}\qquad \begin{aligned}&\frac{1}{n}\sum_{k=1}^{n}\left(\left\|\widetilde{f}_k(X_k)\right\|^2\mathbb{1}_{E_k}(X_k)+\left\|\widetilde{f}^*(X_k)\right\|^2\mathbb{1}_{\overline{E}_k}(X_k)\right)\mathbb{1}_{\overline{V}_k}(X_k)\\ &\qquad\overset{a.s.}{=}\ O\left(\frac{1}{n}\sum_{k=1}^{n}v_k^{2-m}\right)\ +\ O\left(v_n^{2-m}\right).\end{aligned}$$

- Using once again Assumption $[A5]$, we derive

$$\text{(D.9)}\sum_{k=1}^{n} \left\|\widetilde{f}^*(X_k)\right\|^2 \mathbb{1}_{\overline{E}_k}(X_k)\mathbb{1}_{V_k}(X_k)$$
$$\overset{a.s.}{=} O\left(\sum_{k=1}^{n}\left(\|X_k\|^2+1\right)\mathbb{1}_{\left\{x\in\mathbb{R}^d\,;\,(b_f-a_f)\|x\|+(B_f-A_f)\leq\|\widetilde{f}_k(x)\|\right\}}(X_k)\ \mathbb{1}_{V_k}(X_k)\right).$$

In addition, for all constants $c_3, c_4 > 0$, there is $c_5 > 0$ such that for $k \geq 1$

$$\text{(D.10)}\qquad \begin{aligned}&\left\|\widetilde{f}_k(X_k)\right\|^2\mathbb{1}_{V_k}(X_k)\\ &\quad\geq\ \left\|\widetilde{f}_k(X_k)\right\|^2\mathbb{1}_{\left\{x\in\mathbb{R}^d\,;\,\|\widetilde{f}_k(x)\|\geq c_3\,\|x\|\,+c_4\right\}}(X_k)\mathbb{1}_{V_k}(X_k)\\ &\quad\geq\ c_5\left(\|X_k\|^2+1\right)\mathbb{1}_{\left\{x\in\mathbb{R}^d\,;\,\|\widetilde{f}_k(x)\|\geq c_3\|x\|\,+c_4\right\}}(X_k)\mathbb{1}_{V_k}(X_k).\end{aligned}$$

Thus, using this result in (D.9), we derive that

$$\text{(D.11)}\quad \sum_{k=1}^{n} \left\|\widetilde{f}^*(X_k)\right\|^2 \mathbb{1}_{\overline{E}_k}(X_k)\mathbb{1}_{V_k}(X_k) \overset{a.s.}{=} O\left(\sum_{k=1}^{n}\left\|\widetilde{f}_k(X_k)\right\|^2\mathbb{1}_{\{\|X_k\|\leq v_k\}}\right)$$

and then

$$\sum_{k=1}^{n}\left(\left\|\widetilde{f}_k(X_k)\right\|^2 \mathbb{1}_{E_k}(X_k)+\left\|\widetilde{f}^*(X_k)\right\|^2 \mathbb{1}_{\overline{E}_k}(X_k)\right)\mathbb{1}_{V_k}(X_k) \overset{a.s.}{=} O\left(\sum_{k=1}^{n}\left(\sup_{\|x\|\le v_k}\left\|\widetilde{f}_k(x)\right\|\right)^2\right). \tag{D.12}$$

Finally, combining (D.8) and (D.12) leads to

$$\frac{1}{n}\sum_{k=1}^{n}\|\Pi_k\|^2 \overset{a.s.}{=} O\left(\frac{1}{n}\sum_{k=1}^{n} v_k^{2-m}\right) + O\left(v_{n+1}^{2-m}\right) + O\left(\frac{1}{n}\sum_{k=1}^{n}\left(\sup_{\|x\|\le v_k}\left\|\widetilde{f}_k(x)\right\|\right)^2\right). \tag{D.13}$$

Since assumptions of part 2 of Theorem 3.1 are satisfied, $\sup_{\|x\|\le v_n}\|\widetilde{f}_n(x)\| \overset{a.s.}{=} o(1)$, and since $m>2$, then $v_n^{2-m}=o(1)$. Therefore, from (D.13), we deduce that

$$\frac{1}{n}\sum_{k=1}^{n}\|\Pi_k\|^2 \overset{a.s.}{=} o(1). \tag{D.14}$$

Step 3: *study of* $(1/n)\sum_{k=0}^{n-1}\Pi_k^T\,\xi_{k+1}$. For $n\ge 1$, $M_n=\sum_{k=0}^{n-1}\Pi_k^T\,\xi_{k+1}$ is a square integrable martingale. Since $\|\langle M\rangle_n\| \le \text{trace}(\Gamma)\sum_{k=0}^{n-1}\|\Pi_k\|^2 \overset{a.s.}{=} o(n)$ (by (D.14)), we deduce from a strong law of large numbers for the martingales (for example, Duflo [13, Theorem 1.3.15, p. 20]), that for any $\delta>0$,

$$\frac{1}{n}\sum_{k=0}^{n-1}\Pi_k^T\,\xi_{k+1} \overset{a.s.}{=} O\left(\frac{(\log n)^{1+\delta}}{n}\right)^{1/2}. \tag{D.15}$$

Finally, results (D.4), (D.14), and (D.15) give the optimality of the tracking.

To finish the proof of Theorem 4.2, let us remark that

$$\left\|\widehat{\Gamma}_n-\Gamma\right\| \overset{a.s.}{=} O\left(\frac{1}{n}\sum_{k=0}^{n-1}\|\Pi_k\|^2\right) + O\left(\left\|\frac{1}{n}\sum_{k=1}^{n}\xi_k\,\xi_k^T-\Gamma\right\|\right). \tag{D.16}$$

Acknowledgments. The authors thank Prof. Marie Duflo, Prof. Georges Oppenheim, and Prof. Jean-Michel Poggi for helpful discussions, as well as the two anonymous referees for their useful comments which allowed us to significantly improve this paper.

REFERENCES

[1] P. Ango Nze and B. Portier, *Estimation of the density and the regression functions of an absolutely regular stationary process*, Publ. Inst. Statist. Univ. Paris, 38 (1994), pp. 59–87.

[2] K.J. Aström and B. Wittenmark, *On self-tuning regulators*, Automatica, 9 (1973), pp. 185–199.

[3] B. Bercu, *Weighted estimation and tracking for ARMAX models*, SIAM J. Control Optim., 33 (1995), pp. 89–106.

[4] B. Bercu, *Central limit theorem and law of iterated logarithm for least squares algorithms in adaptive tracking*, SIAM J. Control Optim., 36 (1998), pp. 910–928.

[5] D. Bosq, *Nonparametric Statistics for Stochastic Processes*, Lecture Notes in Statist. 110, Springer-Verlag, New York, 1996.

[6] P.E. Caines, *Linear Stochastic System*, John Wiley, New York, Boston, 1985.

[7] H.F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.

[8] H.F. Chen, *Recursive Estimation and Control for Stochastic Systems*, John Wiley, New York, Boston, 1985.

[9] F.-C. Chen and H.K. Khalil, *Adaptive control of a class of nonlinear discrete-time systems using neural networks*, IEEE Trans. Automat. Control, 40 (1995), pp. 791–801.

[10] G. Collomb, *Estimation non paramétrique de la régression: revue bibliographique*, Internat. Statist. Rev., 49 (1981), pp. 75–93.

[11] L. Devroye and L. Györfi, *Nonparametric Density Estimation in L_1 View*, John Wiley, New York, 1985.

[12] P. Doukhan and M. Ghindès, *Estimation de la transition de probabilité d'une chaîne de Markov Döeblin-récurrente*, Stochastic Process. Appl., 15 (1983), pp. 271–293.

[13] M. Duflo, *Random Iterative Models*, Springer-Verlag, New York, 1997.

[14] W. Greblicki, *Nonlinearity estimation in Hammerstein systems based on ordered observations*, IEEE Trans. Signal Process., 44 (1996), pp. 1224–1233.

[15] W. Greblicki and M. Pawlak, *Nonparametric identification of Hammerstein systems*, IEEE Trans. Informat. Theory, 35 (1989), pp. 409–418.

[16] L. Guo and H.F. Chen, *The Aström-Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers*, IEEE Trans. Automat. Control, 36 (1991), pp. 802–812.

[17] L. Guo, *Further results on least squares based adaptive minimum variance control*, SIAM J. Control Optim., 32 (1994), pp. 187–212.

[18] L. Guo, *Self convergence of weighted least squares with applications to stochastic adaptive control*, IEEE Trans. Automat. Control, 41 (1996), pp. 79–89.

[19] W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK, 1990.

[20] N. Hilgert, R. Senoussi, and J.P. Vila, *Estimation non-paramétrique de suites de fonctions de processus autorégressifs non linéaires*, C. R. Acad. Sci. Paris, Sér. I Math., 323 (1996), pp. 1085–1090.

[21] N. Hilgert, J. Harmand, J.-P. Steyer, and J.-P. Vila, *Nonparametric identification and adaptive control of an anaerobic fluidized bed digester*, Control Engineering Practice, 8 (2000), pp. 367–376.

[22] S. Jagannathan, F.L. Lewis, and O. Pastravanu, *Discrete-time model reference adaptive control of nonlinear dynamical systems using neural networks*, Internat. J. Control, 64 (1996), pp. 217–239.

[23] T. Lai and C. Wei, *Extended least squares and their applications to adaptive control of dynamic systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 898–906.

[24] T. Lai and C. Wei, *Asymptotically efficient self-tuning regulators*, SIAM J. Control Optim., 25 (1987), pp. 466–481.

[25] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, Berlin, 1993.

[26] K. Najim, G. Oppenheim, and B. Portier, *Self-tuning control of nonlinear systems using nonparametric estimation*, in Adaptive Systems in Control and Signal Processing, L. Dugard, M. M'Saad, and I.D. Landau, eds., Pergamon Press, Oxford, UK, 1992.

[27] K.S. Narendra and K. Parthasarathy, *Identification and control of dynamical systems using neural networks*, IEEE Trans. Neural Networks, 1 (1990), pp. 4–27.

[28] A. Oulidi, *Estimation fonctionnelle et poursuite non linéaire*, Thèse, Université Paris-Sud, Orsay, France, 1993.

[29] B. Portier and G. Oppenheim, *Adaptive control of nonlinear dynamic systems: Study of the nonparametric estimator*, J. Syst. Eng., 1 (1993), pp. 40–50.

[30] B. Portier, *Estimation non paramétrique et commande adaptative de processus Markoviens non linéaires*, Thèse, Université Paris-Sud, Orsay, France, 1992.

[31] R. Senoussi, *Lois du logarithme itéré et identification*, Thèse d'état, Université Paris-Sud, Orsay, France, 1991.

[32] Y.K. Truong and C.J. Stone, *Nonparametric function estimation involving time series*, Ann. Statist., 20 (1992), pp. 77–97.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 433–456

CONJUGATE POINTS FOR VARIATIONAL PROBLEMS WITH EQUALITY AND INEQUALITY STATE CONSTRAINTS*

HIDEFUMI KAWASAKI[†] AND VERA ZEIDAN[‡]

Abstract. In this paper, variational problems with equality and inequality state constraints are considered. The theory of conjugate points for these problems is developed, and necessary conditions for weak local optimality are derived in terms of this concept and the Legendre condition. For the case of inequality constraints, the envelope-like effect is taken into consideration in the accessory problem.

Key words. conjugate points, accessory problem, inequality and equality state constraints, Jacobi system, Legendre condition, variational problems, necessary optimality conditions

AMS subject classifications. 49K05, 49N05

PII. S0363012998345925

1. Introduction.

In this paper, we deal with two variational problems. One is the variational problem with inequality state constraints:

$$\text{(VP)} \qquad \text{Minimize} \int_0^T f(t, x(t), \dot{x}(t))dt$$
$$\text{subject to} \quad x(0) = x_0, \quad x(T) = x_T, \quad x \in W^n_{1,\infty}[0, T],$$
$$g(t, x(t)) \leq 0 \ \text{ for all } t \in [0, T];$$

the other is the variational problem with equality state constraints:

$$\text{(VP}_0\text{)} \qquad \text{Minimize} \int_0^T f(t, x(t), \dot{x}(t))dt$$
$$\text{subject to} \quad x(0) = x_0, \quad x(T) = x_T, \quad x \in W^n_{1,\infty}[0, T],$$
$$h(t, x(t)) = 0 \ \text{ for all } t \in [0, T],$$

where $T > 0$ is a fixed time, x_0 and x_T are given points in R^n, and $W^n_{1,\infty}[0,T] := \{x : [0,T] \to R^n \mid x_i \text{ ; absolutely continuous, } ||x|| < \infty\}$ equipped with the norm $||x|| = \max_{t\in[0,T]} ||x(t)|| + \text{esssup}_{t\in[0,T]} ||\dot{x}(t)||$. We assume that $f : R^{2n+1} \to R$, $h : [0,T] \times R^n \to R^l$ and $g : [0,T] \times R^n \to R^m$ are continuous and have continuous partial derivatives with respect to (w.r.t.) x and $\dot{x}$ up to order 2 inclusive.

Since the inception of optimal control theory in the 1960s and its increased applications to modern problems, the study of variational problems with state constraints has been the center of attention for several researchers; see, for instance, [2], [9], [10], [11], [12], [17], [28], [29], [31], [36], [38] and the references provided therein. At this stage, many questions pertaining to the first-order optimality conditions have been satisfactorily answered. Recently, the light has been strongly focusing on the questions concerning second-order optimality conditions. For the case of abstract problems with

*Received by the editors October 5, 1998; accepted for publication (in revised form) March 29, 2000; published electronically September 7, 2000. This research was done during the first author's visits to the Department of Mathematics at Michigan State University and the Department of Industrial and Operations Engineering at the University of Michigan in 1997–1998.

http://www.siam.org/journals/sicon/39-2/34592.html

[†]Graduate School of Mathematics, Kyushu University, Fukuoka 812-0053, Japan (kawasaki@math.kyushu-u.ac.jp). This author was supported by the Program for Overseas Research from the Ministry of Education, Science and Culture, Japan, 8-Koh-226.

[‡]Department of Mathematics, Michigan State University, East Lansing, MI 48824 (zeidan@math.msu.edu).

inequality constraints, it was observed by Kawasaki [18] that a certain envelope-like effect occurs and is manifested by the appearance of an extra term in the second-order necessary condition. Later this phenomenon was also noticed by several authors (Ioffe [15], [16], Cominetti [8], Páles and Zeidan [32], [33], and Penot [36]).

For the case of variational problems with state constraints, (e.g., the problem (VP)), it is shown [32] that an extra term appears in the corresponding accessory problem.

In the context of second-order necessary conditions for weak local minima, the concept of conjugate point plays a crucial role in the history of variational problems with fixed endpoints and no state constraints (e.g., [4], [13], and [30]). This concept is usually given in terms of envelopes and is traditionally used as an effective tool to verify the nonnegativity of the second variation, or equivalently, to verify that the accessory problem has a zero minimum value. Hence, it is an important question to obtain optimality conditions in terms of conjugate points theory. This theory was extended in Zeidan and Zezza [43], [44] to the case when general endpoint conditions are present; this naturally includes the periodic ones.

However, the problem of deriving the conjugate point theory for the case when state constraints are present is a long-standing open question. This is the case even for the calculus of variations setting.

The main purpose of this paper is to develop a theory of conjugate points for the problems (VP) and (VP_0). The second aim is to establish in terms of this notion necessary conditions for optimality in (VP) and (VP_0). Our approach is analytical.

In section 2 we review the resulting envelope-like effect for an abstract infinite dimensional optimization problem. Then, the second variation is derived in Theorem 2.3 for (VP) via a direct proof. This result requires the nonemptiness of the set of second-order admissible variations $K(y)$ (see (2.2)) and, as expected, evokes in the second variation an extra term E (see Definition 2.2).

Since the constraint $K(y) \neq \emptyset$ is neither an equality nor an inequality and since the extra term involving E is so complicated to analyze, then the form of the accessory problem given by Theorem 2.3 is not appropriate to derive the conjugate point theorem. In section 3, we obtain from the second-variation of section 2 a "well-behaved" accessory problem (AP) for (VP) that has quadratic objectives and inequalities. The major results of the paper are given in section 4. There, we consider a quadratic problem (GAP) generalizing the accessory problem (AP) for which we develop necessary conditions for optimality phrased in terms of a certain "Jacobi system." This later inspires the introduction of the "conjugate point" notion for (GAP). As a consequence we obtain that the nonexistence of points in $(0, T)$ conjugate to 0 is necessary for optimality in (VP). In section 5 we obtain the Legendre condition for (VP). Conjugate point theory for (VP_0) is derived in section 6. In the final section we provide two numerical examples that illustrate the utility of the results.

2. Preliminary results. Both (VP) and (VP_0) are formulated as an abstract optimization problem in Banach spaces:

$$\text{(P)} \qquad \begin{array}{ll} \text{Minimize} & F(x) \\ \text{subject to} & G(x) \in K, \quad H(x) = 0, \end{array}$$

where X, V, W are Banach spaces, K is a closed convex *cone* in V with nonempty interior, $F : X \to R$, $G : X \to V$, and $H : X \to W$ are of C^2-class. For instance, to include (VP_0) (equality constraints) one would take $V = \{0\}$.

We first give some notations and definitions. We denote by $\langle \cdot, \cdot \rangle$ the canonical pairing between a Banach space V and its topological dual space V^*. For any $A \subset X$, its *interior*, *closure*, and *conical hull* are denoted by $\mathrm{int}A$, $\bar{A}$, and $\mathrm{cone}A$, respectively. All vectors in R^n except gradient vectors are column vectors. For $a \in R^n$, a^T denotes the transpose of a, and a_i denotes the ith component of a. The polar cone of K is defined by $K^\circ = \{v^* \in V^* | \langle v^*, v \rangle \ \leq 0 \quad \text{for all } v \in K\}$. For any $\hat{K} \subset V$ and $v^* \in V^*$, the support function is defined by $\delta^*(v^*|\hat{K}) = \sup\{\langle v^*, v\rangle | v \in \hat{K}\}$. For any twice continuously differentiable mapping H, we denote by $H'(x)$ and $H''(x)$ the first and second Fréchet differentials, respectively. A feasible function $\bar{x}$ is *regular* if $H'(\bar{x})$ has a closed range. A vector $y \in X$ is called a *critical* direction at $\bar{x}$ if

$$F'(\bar{x})y = 0, \quad H'(\bar{x})y = 0, \quad G'(\bar{x})y \in \overline{\mathrm{cone}}(K - G(\bar{x})). \tag{2.1}$$

For a given $y \in X$, we define the second-order admissible variation set

$$K(y) := \{w \in V | d(\theta^2 G(\bar{x}) + \theta G'(\bar{x})y + w,\, K) \to 0 \text{ as } \theta \to \infty\}, \tag{2.2}$$

where $d(a, K)$ denotes the distance from a point a to the set K; see Kawasaki [18], [20]. The set $K(y)$ is the closure of the set of second-order admissible variations defined in [32].

THEOREM 2.1 (see [32, Theorem 6]). *Let $\bar{x}$ be a regular local minimum of (P). Then, for each critical direction y satisfying $K(y) \neq \emptyset$, there exist $\lambda_0 \geq 0$, $v^* \in K^\circ$, and $w^* \in W^*$ not all zero such that*

$$L'(\bar{x}) = 0, \tag{2.3}$$

$$L''(\bar{x})(y, y) - 2\delta^*(v^*|K(y)) \geq 0, \tag{2.4}$$

$$\langle v^*, G(\bar{x})\rangle = 0, \tag{2.5}$$

where

$$L(x) := \lambda_0 F(x) + \langle v^*, G(x)\rangle + \langle w^*, H(x)\rangle.$$

Remark 2.1. As a consequence of (2.1), (2.3), and $v^* \in K^\circ$, we obtain the second complementary condition

$$\langle v^*, G'(\bar{x})y\rangle = 0. \tag{2.6}$$

Remark 2.2. The additional term $\delta^*(v^*|K(y))$ was first introduced in [18]. It is closely related to the second derivative of an envelope formed by the generalized inequality constraint $G(x) \in K$; see [18], [19], and [20]. Readers may also refer to [3], [5], [8], [15], [16], [21], [22], [23], [24], [25], [26], [32], [33], [35], [36], [39], [41], and [42]. Theorem 2.1 was also derived in [15, Thm. 5.1] under the Mangasarian–Fromovitz condition. In that case $\lambda_0 = 1$.

Now, we derive a second-order necessary optimality condition for the variational problem (VP) with inequality state constraints by applying Theorem 2.1 to (VP). Let $\bar{x}$ be an arbitrary weak minimum for (VP). For the sake of simplicity we use the following abbreviations:

$$\bar{f}(t) = f(t, \bar{x}(t), \dot{\bar{x}}(t)),\ \bar{g}(t) = g(t, \bar{x}(t)), \dots .$$

We shall invoke the *full-rank condition*, that is,

$$\{\bar{g}_{jx}(t)\}_{\bar{g}_j(t)=0} \text{ are linearly independent for any } t \in [0,T].$$

Note that it can be shown that the Mangasarian–Fromovitz condition holds if the full-rank condition holds together with

$$g(0,x_0) < 0, \quad g(T,x_T) < 0. \tag{2.7}$$

In order to apply Theorem 2.1 to (VP), we take $X := W^n_{1,\infty}[0,T]$, $V := (C[0,T])^m$, and $W := R^{2n}$. Furthermore, we take $K := \{v \in (C[0,T])^m \mid v(t) \le 0 \text{ for all } t\}$, $F(x) := \int_0^T f(t,x,\dot{x})dt$, $G(x)(t) := g(t,x(t))$, and $H(x) := (x(0)-x_0, x(T)-x_T)^T$, respectively.

Under the assumptions in section 1, the mappings $F : X \to R$, $G : X \to V$, and $H : X \to R^{2n}$ are twice continuously Fréchet differentiable, and their first and second Fréchet differentials are given by

$$F'(\bar{x})y = \int_0^T \{\bar{f}_x y + \bar{f}_{\dot{x}}\dot{y}\}dt, \tag{2.8}$$

$$F''(\bar{x})(y,y) = \int_0^T \{y^T \bar{f}_{xx} y + 2y^T \bar{f}_{x\dot{x}}\dot{y} + \dot{y}^T \bar{f}_{\dot{x}\dot{x}}\dot{y}\}dt, \tag{2.9}$$

$$G'(\bar{x})y = \bar{g}_x y, \quad G''(\bar{x})(y,y) = y^T \bar{g}_{xx} y, \tag{2.10}$$

$$H'(\bar{x})y = (y(0), y(T))^T, \quad H''(\bar{x})(y,y) = (0,0)^T; \tag{2.11}$$

see, e.g., Girsanov [11] and Ioffe and Tihomirov [14].

The set $K(y)$ in this setting is

$$\begin{aligned} K(y) := \{w \in (C[0,T])^m \mid \exists\ \Delta(\theta) \in (C[0,T])^m \text{ such that } \Delta(\theta) \to 0 \text{ as } \theta \to \infty, \\ \theta^2 \bar{g}(t) + \theta \bar{g}_x(t) y(t) + w(t) + \Delta(\theta)(t) \le 0 \quad \text{for all } t, \quad \text{for all } \theta > 0\}. \end{aligned}$$

In order to deal with the extra term $\delta^*(v^*|K(y))$ in (2.4), we need the function $E(t)$ introduced in [19] and [20]. A function closely related to E was also introduced for this context in [32] and [33].

Definition 2.2. *For any u, $v \in C[0,T]$ that satisfy $u(t) \ge 0$ for all t and $v(t) \ge 0$ for any t such that $u(t) = 0$, we define a function $E : [0,T] \to [-\infty, \infty]$ by*

$$E(t) := \begin{cases} \max\left\{\limsup_{n\to\infty} \dfrac{v(t_n)^2}{4u(t_n)};\ \{t_n\} \text{ satisfies } (2.13)\right\} & \textit{if } t \in T_0, \\ 0 & \textit{if } t \in T_1 \backslash T_0, \\ -\infty & \text{otherwise}, \end{cases} \tag{2.12}$$

where

$$T_0 := \left\{ t \in T \mid {}^{\exists}t_n \to t \text{ s.t. } u(t_n) > 0,\ -\frac{v(t_n)}{u(t_n)} \to +\infty \right\}, \tag{2.13}$$

$$T_1 := \{t \in T \mid u(t) = v(t) = 0\}. \tag{2.14}$$

Furthermore, it is convenient to use the following notation: Let $y(t)$ be critical. For each $j = 1, \ldots, m$, take in Definition 2.2

$$u(t) = -\bar{g}_j(t), \quad v(t) = -\bar{g}_{jx}(t)y(t). \tag{2.15}$$

Then we denote the function $E(t)$ by $E_j(t; y)$, and $(E_1(t; y), \ldots, E_m(t; y))^T$ by $E(t; y)$, respectively. The following second-order necessary optimality condition was essentially obtained in Páles, Zeidan [32, Thm. 6] for a more general problem. Since the problem (VP) is simpler than that considered in [32], a direct and short proof is provided below in order to complete the presentation of this paper.

THEOREM 2.3. *If $\bar{x}$ is a weak minimum for* (VP)*, then for each critical direction $y \in W^n_{1,\infty}[0, T]$ satisfying $K(y) \neq \emptyset$, there exists a constant vector $a \in R^n$, a constant $\lambda_0 \geq 0$, and a nondecreasing function $\lambda : [0, T] \to R^m$ that is right-continuous except at $t = 0$ such that λ_0 and $d\lambda$ are not zero and*

$$\lambda_0 \left(\bar{f}_{\dot{x}}(t) - \int_0^t \bar{f}_x ds \right) - \int_{(0,t]} d\lambda^T \bar{g}_x = a^T \text{ a.e. } t \in [0, T], \tag{2.16}$$

(2.17)
$$\int_0^T \lambda_0 \{y^T \bar{f}_{xx} y + 2y^T \bar{f}_{x\dot{x}} \dot{y} + \dot{y}^T \bar{f}_{\dot{x}\dot{x}} \dot{y}\} dt + \int_{[0,T]} y^T (d\lambda^T \bar{g})_{xx} y + 2 \int_{[0,T]} d\lambda(t)^T E(t; y) \geq 0,$$

and

$$d\lambda_j(t) = 0 \quad \text{on} \quad \{t \mid \bar{g}_j(t) = \bar{g}_{jx}(t)y(t) = 0\}^c, \quad j = 1, \ldots, m. \tag{2.18}$$

If in addition the full-rank condition holds then $\lambda_0 \neq 0$, say $\lambda_0 = 1$, and in this case, λ does not depend on y.

Remark 2.3. Since $y = 0$ is a critical direction and since $0 \in K(0)$, then by Theorem 2.3, $y = 0$ has an associated $\lambda_0 \geq 0$ and λ satisfying (2.16) and (2.20). Under the full-rank condition and (2.7), Theorem 2.3 asserts that this particular λ_0 is nonzero and that when we take $\lambda_0 = 1$, λ is independent of y and is unique up to a constant.

Proof. By Riesz's representation theorem, the Lagrange function $L(x)$ is represented as

$$L(x) = \int_0^T \lambda_0 f dt + \int_{[0,T]} d\lambda^T g + \sum_{k=0,\ T} \nu_k^T (x(k) - x_k), \tag{2.19}$$

where $\nu_0,\ \nu_T \in R^n$ and $\lambda : [0, T] \to R^m$ is a componentwise nondecreasing function; see, e.g., Rudin [40]. For any critical direction y, we get from (2.3), (2.19), and the integration by parts that

$$\begin{aligned} 0 = L'(\bar{x})y &= \int_0^T \lambda_0 \{\bar{f}_x y + \bar{f}_{\dot{x}} \dot{y}\} dt + \int_{[0,T]} d\lambda^T \bar{g}_x y + \sum_{k=0}^{T} \nu_k^T y(k) \\ &= \int_0^T \lambda_0 \left\{ \bar{f}_{\dot{x}} - \int_0^t \bar{f}_x dt - \int_{(0,t]} d\lambda^T \bar{g}_x \right\} \dot{y} dt \end{aligned}$$

for all $y \in W^n_{1,\infty}[0,T]$ satisfying $y(0) = y(T) = 0$. This implies (2.16); see, e.g., [11], [14], and [27]. On the other hand, the complementarity condition (2.5) becomes

$$\int_{[0,T]} d\lambda^T \bar{g} = 0. \tag{2.20}$$

Since λ is nondecreasing and $\bar{g}$ is nonpositive, (2.20) implies that

$$d\lambda_j(t) = 0 \ \text{ if } \ \bar{g}_j(t) < 0. \tag{2.21}$$

Hence the second complementarity condition (2.6) becomes

$$\int_{\bar{g}(t)=0} d\lambda^T \bar{g}_x = 0. \tag{2.22}$$

Since $G'(\bar{x})y \in \overline{\text{cone}}(K - G(\bar{x}))$, it follows from Lemma 6.1 in [18] that

$$\bar{g}_x(t)y(t) \le 0 \ \text{ if } \ \bar{g}(t) = 0. \tag{2.23}$$

Combining (2.4), (2.21), (2.22), and (2.23), we get (2.18). On the other hand, it was shown in [20, p. 222] and [34, Cor. (4.2) (iv)] that

$$\delta^*(v^*|K(y)) = -\int_{[0,T]} d\lambda(t)^T E(t;y). \tag{2.24}$$

Combining (2.9), (2.10), and (2.24), we get (2.17).

Now, assume the full-rank condition. Then, if $\lambda_0 = 0$, (2.16) yields that

$$\int_{(t_1,t_2]} d\lambda^T \bar{g}_x = 0 \qquad \text{for all } \ t_1, t_2 \in [0,T].$$

The right continuity of λ, the full-rank condition, and (2.21) imply that $d\lambda = 0$ on $[0,T]$. Hence a contradiction is obtained. Thus, $\lambda_0 \neq 0$ and can be taken to be 1. Similar arguments show that if λ and μ are two nondecreasing functions that are right continuous except at $t = 0$ and satisfy (2.16) and (2.21), then they must have $d\lambda = d\mu$ on $[0,T]$. □

3. The accessory problem (AP). In this section, we will give an accessory problem (AP) for (VP). Assume in the rest of the paper that the full rank condition and (2.7) hold. By Theorem 2.3 and Remark 2.3, there exist unique (up to a constant) multipliers $\lambda_0 = 1$ and λ such that Theorem 2.3 leads to a prototype of our accessory problem:

$$\begin{aligned}(\text{AP}_0)\ \text{Minimize } & \frac{1}{2}\int_0^T \{y^T \bar{f}_{xx} y + 2y^T \bar{f}_{x\dot{x}} \dot{y} + \dot{y}^T \bar{f}_{\dot{x}\dot{x}} \dot{y}\} dt \\ & + \frac{1}{2}\sum_{j=1}^m \int_{[0,T]} \{y^T \bar{g}_{jxx} y + 2E_j\} d\lambda_j, \\ & \text{subject to } y \ \text{ being critical and } \ K(y) \neq \emptyset.\end{aligned}$$

From Theorem 2.3, it follows that $y(t) \equiv 0$ is a minimum of (AP_0).

Note that when we attempt to give a first-order necessary optimality condition for (AP_0) as well as the classical Jacobi equation, we encounter two difficulties:

(1) The constraint $K(y) \neq \emptyset$ is neither an inequality constraint nor an equality constraint.

(2) The definition of $E(t;y)$ is so complicated that the extra term $\int_{[0,T]} E_j(t;y)d\lambda_j$ is hard to analyze as a function of y.

Hence the key of this section is to overcome these two difficulties. Before solving them, we first deal with critical directions.

Lemma 3.1. *When $\bar{x}(t)$ is a weak minimum for* (VP)*, then a function $y \in W^n_{1,\infty}[0,T]$ is critical if and only if*

$$\int_{[0,T]} d\lambda(t)^T \bar{g}_x(t)y(t) = 0, \tag{3.1}$$

$$\bar{g}_{jx}(t)y(t) \leq 0 \quad \text{if} \quad \bar{g}_j(t) = 0 \tag{3.2}$$

for all j, and

$$y(0) = y(T) = 0. \tag{3.3}$$

Proof. It is evident that $H'(\bar{x})y = 0$ is equivalent to (3.3). As was seen in the proof of Theorem 2.3, the condition $G'(\bar{x})y \in \overline{\text{cone}}(K - G(\bar{x}))$ is equivalent to (3.2). Since the first-order optimality condition (2.3) is satisfied, the condition $F'(\bar{x})y = 0$ is equivalent to $\langle v^*, G'(\bar{x})y\rangle = 0$, which is furthermore equivalent to (3.1). This completes the proof. □

In this result, we characterize the nonemptiness of $K(y)$ in terms of inequalities. Thus, the first difficulty encountered above is circumvented.

Lemma 3.2. *For any critical direction $y(t)$, the necessary and sufficient condition for the nonemptiness of $K(y)$ is that there exists $\beta \in W^m_{1,\infty}[0,T]$ such that*

$$\bar{g}_{jx}(t)y(t) + \sqrt{-2\bar{g}_j(t)}\beta_j(t) \leq 0 \tag{3.4}$$

for all t and $j = 1, \ldots, m$. Furthermore, when $\beta(t)$ satisfies (3.4)*, it holds that*

$$\sum_{j=1}^m \int_{[0,T]} \beta(t)^2 d\lambda_j \geq 2\sum_{j=1}^m \int_{[0,T]} E_j(t;y)d\lambda_j, \tag{3.5}$$

where $\lambda_j(t)$ is the nondecreasing function guaranteed in Theorem 2.3.

Proof. First we note that $K(y)$ is nonempty if and only if $E_j(t;y) < \infty$ for all t and $j = 1, \ldots, m$; see Theorem 2.1 in [19]. Next, for each $j = 1, \ldots, m$, put

$$u_j(t) := -\bar{g}_j(t), \quad v_j(t) := -\bar{g}_{jx}(t)y(t). \tag{3.6}$$

Sufficiency: Let β satisfy (3.4). Then, for any converging sequence $t_n \to t$ satisfying

$$-\frac{v_j(t_n)}{u_j(t_n)} \to \infty, \tag{3.7}$$

we get from (3.4) that

$$0 < \frac{-v_j(t_n)}{\sqrt{2u_j(t_n)}} \leq -\beta_j(t_n). \tag{3.8}$$

Tending $n \to \infty$, we get

$$0 \leq \limsup_{n\to\infty} \frac{v_j(t_n)^2}{2u_j(t_n)} \leq \beta_j(t)^2. \tag{3.9}$$

It follows from the definition of $E_j(t; y)$ that $0 \leq 2E_j(t; y) \leq \beta_j(t)^2 < \infty$. When there is no sequence $t_n \to t$ satisfying (3.7), it is evident from the definition of $E_j(t; y)$ that $2E_j(t; y) \leq 0 \leq \beta_j(t)^2$. Since $d\lambda \geq 0$, we readily get (3.5) from these inequalities.

Necessity: Since $v_j(t) \geq 0$ if $u_j(t) = 0$, condition (3.4) implies that

$$\beta_j(t) \leq \frac{v_j(t)}{\sqrt{2u_j(t)}} \quad \text{if} \quad u_j(t) > 0 \quad \text{for all} \quad j = 1, \ldots, m. \tag{3.10}$$

Since β_j is continuous, this condition yields that

$$-\frac{v_j(t)}{\sqrt{2u_j(t)}} \text{ is bounded above on } \{t : u_j(t) > 0\} \text{ for all } \quad j = 1, \ldots, m. \tag{3.11}$$

Conversely, assume that (3.11) holds for some upper bounds $M_j > 0$. Set $\beta_j(t) = M_j$ on $[0, T]$. The result is that the function $\beta = (\beta_j)$ is in $W^m_{1,\infty}[0, T]$ and satisfies (3.4) on $[0, T]$. Therefore, (3.4) is equivalent to (3.11).

Now, we proceed by contradiction. If (3.11) does not hold, then there exist some j and a sequence t_n such that $u(t_n) > 0$ and

$$\lim_{n\to\infty} \frac{-v_j(t_n)}{\sqrt{2u_j(t_n)}} = \infty. \tag{3.12}$$

Here we may assume that t_n converges to some point $\bar{t}$. Then it is evident that $v_j(t_n) < 0$ for all sufficiently large n,

$$\lim_{n\to\infty} \frac{v_j(t_n)^2}{4u_j(t_n)} = \infty, \tag{3.13}$$

and

$$-\frac{v_j(t_n)}{u_j(t_n)} = \frac{v_j(t_n)^2}{u_j(t_n)} \frac{1}{-v_j(t_n)} \to \infty. \tag{3.14}$$

Therefore $E_j(\bar{t}; y) = \infty$, so that $K(y)$ is empty. This completes the proof. □

Combining Lemmas 3.1 and 3.2, we get another variational problem, which we call the accessory problem for the variational problem (VP):

$$\begin{aligned}
\text{(AP) Minimize } & \frac{1}{2}\int_0^T \{y^T \bar{f}_{xx} y + 2y^T \bar{f}_{x\dot{x}} \dot{y} + \dot{y}^T \bar{f}_{\dot{x}\dot{x}} \dot{y}\} dt \\
& + \frac{1}{2}\sum_{j=1}^m \int_{[0,T]} \{y^T \bar{g}_{jxx} y + \beta_j^2\} d\lambda_j \\
\text{subject to } & y \in W^n_{1,\infty}[0, T], \quad \beta \in W^m_{1,\infty}[0, T], \\
& \bar{g}_{jx}(t) y(t) + \sqrt{-2\bar{g}_j(t)}\beta_j(t) \leq 0 \quad \text{if} \quad d\lambda_j(t) = 0, \\
& \bar{g}_{jx}(t) y(t) + \sqrt{-2\bar{g}_j(t)}\beta_j(t) = 0 \quad \text{if} \quad d\lambda_j(t) > 0, \\
& y(0) = y(T) = 0.
\end{aligned}$$

Theorem 2.3 implies that the objective function of (AP) is nonnegative. Hence, it is evident that $(y(t), \beta(t)) \equiv (0, 0)$ is a minimum of the accessory problem (AP).

4. Conjugate points for the (VP) problem. The goal of this section is to eventually derive necessary conditions for the (VP) problem in terms of conjugate points. In order to achieve this goal, we first derive the conjugate point theory for the general quadratic problem (GAP) below, and then the conjugate points for (VP) will be defined exactly as the conjugate points for its accessory problem (AP), which is quadratic. In this section, we first give a first-order necessary optimality condition (Jacobi system) for the accessory problem (AP). Next, we will define conjugate points by using the Jacobi system, and finally we will prove that the open interval $(0, T)$ includes no points conjugate to 0 for any weak minimum for (VP).

$$\text{(GAP) Minimize } \frac{1}{2}\int_0^T \{y^T P y + 2y^T Q\dot{y} + \dot{y}^T R\dot{y}\}dt + \frac{1}{2}\sum_{j=1}^{m}\int_{[0,T]}\{y^T C_j y + \beta_j^2\}d\lambda_j$$

subject to $y \in W^n_{1,\infty}[0,T], \quad \beta \in W^m_{1,\infty}[0,T],$

$$a_j(t)^T y(t) + \alpha_j(t)\beta_j(t) \le 0 \text{ if } d\lambda_j(t) = 0, \tag{4.1}$$

$$a_j(t)^T y(t) + \alpha_j(t)\beta_j(t) = 0 \quad \text{if } d\lambda_j(t) > 0, \tag{4.2}$$

$$y(0) = y(T) = 0,$$

where R, Q, and P are $n\times n$-matrix-valued essentially bounded measurable functions for $j = 1, \dots, m$, C_j is a continuous $n\times n$-matrix-valued function, $a_j(t)$ in $W^n_{1,\infty}[0,T]$ and $\alpha_j(t)$ in $W_{1,\infty}[0,T]$, and $d\lambda_j$ is a given nonnegative regular Borel measure. We assume that $R(t)$, $P(t)$, and $C_j(t)$ are symmetric, $\alpha_j(0)$ and $\alpha_j(T)$ are positive for all $j = 1, \dots, m$, the measure satisfies

$$\alpha_j(t)d\lambda_j(t) = 0 \quad \text{on} \quad [0,T] \tag{4.3}$$

for all $j = 1, \dots, m$, and $\{a_j^T(t)\}_{\alpha_j(t)=0}$ are linearly independent for all $t \in [0,T]$.

In the following, we use these abbreviations:

$$\alpha(t) := \begin{pmatrix}\alpha_1(t)\\ \vdots \\ \alpha_m(t)\end{pmatrix}, \quad D(\alpha) := \begin{pmatrix}\alpha_1 & & 0\\ & \ddots & \\ 0 & & \alpha_m\end{pmatrix}, \quad A(t) := \begin{pmatrix}a_1(t)^T\\ \vdots \\ a_m(t)^T\end{pmatrix}. \tag{4.4}$$

The following theorem is a first-order necessary optimality condition for the generalized accessory problem (GAP).

THEOREM 4.1. *If* $(y, \beta) \in W^n_{1,\infty}[0,T]\times W^m_{1,\infty}[0,T]$ *is a minimum for the accessory problem* (GAP), *then there exist* $d \in R^n$ *and* $\mu : [0,T] \to R^m$ *with bounded variation that is right-continuous except at* $t = 0$ *such that* $y(0) = y(T) = 0$, *and the Jacobi system holds:*

(J1)

$$Q(t)^T y(t) + R(t)\dot{y}(t) - \int_0^t \{Py + Q\dot{y}\}dt - \sum_{j=1}^{m}\int_{(0,t]} C_j y d\lambda_j - \int_{(0,t]} A^T d\mu = d \quad \text{a.e. } t,$$

$$\beta_j(t)d\lambda_j(t) = \alpha_j(t)d\mu_j(t) = 0, \tag{J2}$$

$$a_j(t)^T y(t) \le 0 \quad \text{if} \quad \alpha_j(t) = 0, \tag{J3}$$

(J4) $$a_j(t)^T y(t) + \alpha_j(t)\beta_j(t) \le 0 \quad \textit{if} \ \ d\lambda_j(t) = 0,$$

(J5) $$a_j(t)^T y(t) d\mu_j(t) = a_j(t)^T y(t) d\lambda_j(t) = 0,$$

and

(J6) $$d\mu_j(t) \ge 0 \quad \textit{if} \ \ d\lambda_j(t) = 0$$

for all $t \in [0, T]$ *and* $j = 1, \ldots, m$.

Proof. For the sake of simplicity, we denote by $F(y, \beta)$ the quadratic objective function of (GAP). Now, let (y, β) be a minimum for (GAP). Then, since all the functions in the constraints are linear w.r.t. (y, β), it is easily seen that there exists no $(z, \gamma) \in W^n_{1,\infty}[0, T] \times W^m_{1,\infty}[0, T]$ such that $F'(y, \beta)(z, \gamma) < 0$, $Az + D(\alpha)\gamma \in K_\lambda$, and $z(0) = z(T) = 0$, where K_λ denotes the set of all $v \in W^m_{1,\infty}[0, T]$ that satisfies

(4.5) $$v_j(t) \le 0 \text{ if } d\lambda_j(t) = 0 \ \text{ and } \ a_j(t)^T y(t) + \alpha_j(t)\beta_j(t) = 0,$$

(4.6) $$v_j(t) \in R \text{ if } d\lambda_j(t) = 0 \ \text{ and } \ a_j(t)^T y(t) + \alpha_j(t)\beta_j(t) < 0,$$

(4.7) $$v_j(t) = 0 \text{ if } d\lambda_j(t) > 0$$

for all $j = 1, \ldots, m$. This implies that the zero-vector in $R \times (C[0, T])^m \times R^{2n}$ does not belong to the following convex cone:

(4.8) $$C := \left\{ \begin{pmatrix} F'(y,\, \beta)(z, \gamma) + p \\ Az + D(\alpha)\gamma - v \\ z(0) \\ z(T) \end{pmatrix} ; \begin{array}{l} p > 0 \\ v \in K_\lambda \\ z \in W^n_{1,\infty}[0, T] \\ \gamma \in W^m_{1,\infty}[0, T] \end{array} \right\}.$$

As we shall later see, C has nonempty interior. Hence, by the separation theorem, there exist $\lambda_0 \in R$, ν_0, $\nu_T \in R^n$, and $\mu : [0, T] \to R^m$ of bounded variations that are right-continuous except at $t = 0$ not all zero such that

(4.9) $$\lambda_0\{F'(y, \beta)(z, \gamma) + p\} + \int_{[0,T]} d\mu^T \{Az + D(\alpha)\gamma - v\} + \nu_0^T z(0) + \nu_T^T z(T) \ge 0$$

for all $p > 0$, $v \in K_\lambda$, $z \in W^n_{1,\infty}[0, T]$, and $\gamma \in W^m_{1,\infty}[0, T]$. We easily get $\lambda_0 \ge 0$. From (4.9) and the definition of K_λ, we obtain

(4.10) $$d\mu_j(t) \ge 0, \quad \text{if} \ \ d\lambda_j(t) = 0 \ \text{ and } \ a_j(t)^T y(t) + \alpha_j(t)\beta_j(t) = 0,$$

and

(4.11) $$d\mu_j(t) = 0, \quad \text{if} \ \ d\lambda_j(t) = 0 \ \text{ and } \ a_j(t)^T y(t) + \alpha_j(t)\beta_j(t) < 0,$$

which yields the last assertion (J6). Tending $p \to +0$ and taking $v = 0$ in (4.9), we have

(4.12) $$\lambda_0 F'(y, \beta)(z, \gamma) + \int_{[0,T]} d\mu^T \{Az + D(\alpha)\gamma\} + \nu_0^T z(0) + \nu_T^T z(T) \ge 0$$

for all $z \in W^n_{1,\infty}[0, T]$ and $\gamma \in W^m_{1,\infty}[0, T]$. Let us show $\lambda_0 > 0$. Suppose that $\lambda_0 = 0$ in (4.12); then we have

(4.13) $$\int_{[0,T]} d\mu^T \{Az + D(\alpha)\gamma\} + \nu_0^T z(0) + \nu_T^T z(T) = 0.$$

Hence, for any (z, γ) satisfying $z(0) = z(T) = 0$, we have, by integration by parts, that

$$\int_0^T \left(\int_{(0,t]} d\mu^T A \right) \dot{z}(t)dt + \int_0^T \left(\int_{[0,t]} d\mu^T D(\alpha) \right) \dot{\gamma}(t)dt = 0. \tag{4.14}$$

Hence both $\int_{(0,t]} d\mu^T A$ and $\int_{(0,t]} d\mu^T D(\alpha)$ are constant, so that $d\mu^T (A, D(\alpha)) = 0$. Since the matrix $(A(t), D(\alpha(t)))$ has full rank, we have $d\mu(t) = 0$. Hence from (4.13) we have $\nu_0 = \nu_T = 0$. This leads to a contradiction. Hence we may assume that $\lambda_0 = 1$, and we get from (4.12) that

$$F'(y, \beta)(z, \gamma) + \int_{[0,T]} d\mu^T \{Az + D(\alpha)\gamma\} = 0 \tag{4.15}$$

for any (z, γ) such that $z(0) = z(T) = 0$. That is,

$$\int_0^T \{y^T Pz + \dot{y}^T Q^T z + y^T Q\dot{z} + \dot{y}^T R\dot{z}\}dt + \sum_{j=1}^m \int_{[0,T]} \{y^T C_j z d\lambda_j + d\lambda^T D(\beta)\gamma\}$$
$$+ \int_{[0,T]} d\mu^T \{Az + D(\alpha)\gamma\} = 0.$$

By integration by parts, we have

$$\int_0^T \dot{z}(t)^T \left\{ Q(t)^T y(t) + R(t)\dot{y}(t) - \int_0^t \{Py + Q\dot{y}\}dt - \sum_{j=1}^m \int_{(0,t]} C_j y d\lambda_j - \int_{(0,t]} A^T d\mu \right\} dt$$

$$- \int_0^T \dot{\gamma}(t)^T \left\{ \int_{(0,t]} D(\beta)d\lambda + \int_{(0,t]} D(\alpha)d\mu \right\} dt = 0. \tag{4.16}$$

From (4.16), we get (J1) and

$$D(\beta(t))d\lambda(t) + D(\alpha(t))d\mu(t) = 0 \quad \text{on} \quad [0, T]. \tag{4.17}$$

Combining (4.3) and (4.17), we get (J2).

From (4.1), (4.2), (4.10), and (4.11) we get (J3), (J4),

$$\{a_j(t)^T y(t) + \alpha_j(t)\beta_j(t)\}d\lambda_j(t) = 0, \tag{4.18}$$

and

$$\{a_j(t)^T y(t) + \alpha_j(t)\beta_j(t)\}d\mu_j(t) = 0. \tag{4.19}$$

Combining (4.18), (4.19), and (J2), we get (J5). Finally, we prove that the convex cone C has nonempty interior. For any $u \in (W_{1,\infty}^+[0, T])^m$, that is, for any nonnegative-valued function in $W_{1,\infty}^m[0, T]$, the full-rank property of $(A, D(\alpha))$ implies the following linear equation has an unique solution $w(t) \in R^m$:

$$(A(t), D(\alpha(t))) \begin{pmatrix} A(t)^T \\ D(\alpha(t)) \end{pmatrix} w(t) = u(t). \tag{4.20}$$

Put

$$z(t) := A(t)^T w(t), \quad \gamma(t) := D(\alpha(t))w(t). \tag{4.21}$$

Then, by Cramer's formula, z in $W^n_{1,\infty}[0,T]$ and γ in $W^m_{1,\infty}[0,T]$. From (4.20), we have

$$A(t)z(t) + D(\alpha(t))\gamma(t) = u(t). \tag{4.22}$$

Since $\alpha(0)$ and $\alpha(T)$ are positive and since α is continuous, there exists $\delta > 0$ such that α is positive on $[0,\delta] \cup [T-\delta, T]$. Next, for any $\xi_0,\ \xi_T \in R^n$, put

$$\bar{z}(t) := \begin{cases} \frac{1}{\delta}\{tz(t) + (\delta - t)\xi_0(t)\} & \text{on } [0,\delta], \\ z(t) & \text{on } [\delta, T-\delta], \\ \frac{1}{\delta}\{(T-t)z(t) + (t-T+\delta)\xi_T(t)\} & \text{on } [T-\delta, T], \end{cases} \tag{4.23}$$

where

$$\xi_0(t) = \frac{1}{\delta}\left[z(\delta)t + (\delta - t)\xi_0\right]$$

and

$$\xi_T(t) = \frac{1}{\delta}\left[z(T-\delta)(T-t) + (t-T+\delta)\xi_T\right].$$

Similarly for a suitable $\eta \in W^m_{1,\infty}[0,T]$, satisfying $\eta(\delta) = \gamma(\delta)$ and $\eta(T-\delta) = \gamma(T-\delta)$, and which we will define later, put

$$\bar{\gamma}(t) := \begin{cases} \frac{1}{\delta}\{t\gamma(t) + (\delta - t)\eta(t)\} & \text{on } [0,\delta], \\ \gamma(t) & \text{on } [\delta, T-\delta], \\ \frac{1}{\delta}\{(T-t)\gamma(t) + (t-T+\delta)\eta(t)\} & \text{on } [T-\delta, T]. \end{cases} \tag{4.24}$$

Then it is evident from (4.22)–(4.24) that

$$\bar{z}(0) = \xi_0, \quad \bar{z}(T) = \xi_T, \ \dot{\bar{z}}(\delta) = \dot{z}(\delta), \quad \text{and} \quad \dot{\bar{z}}(T-\delta) = \dot{z}(T-\delta), \tag{4.25}$$

$\bar{z}$ and $\bar{\gamma}$ are in $W^n_{1,\infty}$ and $W^m_{1,\infty}$, and

$$A(t)\bar{z}(t) + D(\alpha(t))\bar{\gamma}(t) = u(t) \quad \text{on} \quad [\delta, T-\delta]. \tag{4.26}$$

Furthermore, for any $t \in [0,\delta]$, it follows from (4.22)–(4.24) that

(4.27)

$$\begin{aligned} &u(t) - \{A(t)\bar{z}(t) + D(\alpha(t))\bar{\gamma}(t)\} \\ &= A(t)z(t) + D(\alpha(t))\gamma(t) - A(t)\frac{tz(t) + (\delta - t)\xi_0(t)}{\delta} - D(\alpha(t))\frac{t\gamma(t) + (\delta - t)\eta(t)}{\delta} \\ &= \frac{\delta - t}{\delta}\{A(t)(z(t) - \xi_0(t)) + D(\alpha(t))\gamma(t)\} - \frac{\delta - t}{\delta}D(\alpha(t))\eta(t). \end{aligned}$$

We can choose $\eta(t)$ so that the right-hand side of (4.27) is nonnegative for all $t \in [0,\delta]$. Indeed, put $\rho(t) := A(t)(z(t) - \xi_0(t)) + D(\alpha(t))\gamma(t)$. Then the right-hand side of (4.27) is nonnegative if and only if

$$\rho_j(t) \geq \alpha_j(t)\eta_j(t) \ \text{ for all } \ j = 1, \dots, m. \tag{4.28}$$

Since $\alpha_j(t) > 0$ on $[0, \delta]$ and since both ρ and α are in $W^m_{1,\infty}[0, T]$, (4.28) is achieved by a certain function $\eta(t)$ in $W^m_{1,\infty}[0, T]$. This fact together with (4.27) implies that

$$u(t) \geq A(t)\bar{z}(t) + D(\alpha(t))\bar{\gamma}(t) \quad \text{on} \quad [0, \delta]. \tag{4.29}$$

Similarly, we can prove that the inequality in (4.29) also holds on $[T-\delta, T]$ by choosing a suitable $\eta(t)$. Finally, take

$$v(t) := \begin{cases} 0 & \text{on } [\delta, T - \delta], \\ A(t)\bar{z}(t) + D(\alpha(t))\bar{\gamma}(t) - u(t) \leq 0 & \text{on } [0, \delta] \cup [T - \delta, T]. \end{cases} \tag{4.30}$$

Then it is easily seen from assumption (4.3) that $v \in K_\lambda$ and $u = A\bar{z} + D(\alpha)\bar{\gamma} - v$. Therefore the convex cone C contains $[r, \infty) \times (W^+_{1,\infty}[0, T])^m \times R^{2n}$, where r is some real number. This completes the proof. □

Definition 4.2. *A pair $(y, \beta) \in W^n_{1,\infty}[0, T] \times W^m_{1,\infty}[0, T]$ is said to satisfy the Jacobi system for* (GAP) *if there exist a constant vector $d \in R^n$ and $\mu : [0, T] \to R^m$ with bounded variation that is right-continuous except at $t = 0$ such that* (J1)–(J6) *are satisfied.*

Theorem 4.3. *Let c be in $(0, T]$, and let $(y, \beta) \in W^n_{1,\infty}[0, T] \times W^m_{1,\infty}[0, T]$ satisfy the generalized Jacobi system on $[0, c]$ and the end points condition*

$$y(0) = y(c) = 0 \quad \text{and} \quad \beta(c)^T \int_{(c,T]} d\lambda = 0. \tag{4.31}$$

Define $\bar{y}(t)$ and $\bar{\beta}(t)$ by

$$\bar{y}(t) := \begin{cases} y(t) & \text{on } [0, c], \\ 0 & \text{on } [c, T], \end{cases} \qquad \bar{\beta}(t) := \begin{cases} \beta(t) & \text{on } [0, c], \\ \beta(c) & \text{on } [c, T], \end{cases} \tag{4.32}$$

respectively. Then $(\bar{y}, \bar{\beta})$ is also a feasible solution for the generalized accessory problem (GAP)*, and the value of the objective function at $(\bar{y}, \bar{\beta})$ is zero.*

Proof. It is clear that $(\bar{y}, \bar{\beta})$ is a feasible solution for (GAP). It suffices to show that

$$\int_0^c \{y^T P y + 2y^T Q\dot{y} + \dot{y}^T R\dot{y}\}dt + \sum_{j=1}^m \int_{[0,c]} \{y^T C_j y + \beta_j^2\} d\lambda_j = 0. \tag{4.33}$$

By (J2), (J1), and integration by parts, this integration is equal to

$$\int_0^c \left(y(t)^T Q(t) + \dot{y}(t)^T R(t) - \int_0^t y^T P dt - \int_0^c \dot{y}^T Q^T dt - \sum_{j=1}^m \int_{(0,t]} y^T C_j d\lambda_j \right) \dot{y}(t) dt$$

$$= \int_0^c \left(\int_{(0,t]} d\mu^T A \right) \dot{y} dt = \int_{[0,c]} d\mu^T A y = 0,$$

where the first equality follows from the complementarity condition (J5). This completes the proof. □

Definition 4.4. *Let y belong to $W^m_{1,\infty}[0, T]$ and let c be in $(0, T]$. Then we denote by $\dot{y}_{ess}(c - 0)$ the set of all essential cluster points of $\dot{y}(t)$ as $t \to c - 0$, that is, those that persist upon the removal of any set of measure zero; see Clarke* [6, Ex. 2.2.5].

DEFINITION 4.5. *A point $c \in (0,T]$ is said to be conjugate to $t=0$ for* (GAP) *if there exist a pair $(y,\beta) \in W^m_{1,\infty}[0,T] \times W^m_{1,\infty}[0,T]$ and an essential cluster point $\xi \in \dot{y}_{ess}(c-0)$ such that they satisfy, with some μ and d, the Jacobi system* (J1)–(J6) *on $[0,c]$, the end points condition* (4.31),

$$a_j(c)^T\xi \geq 0 \quad \textit{if} \quad \alpha_j(c) = 0 \quad \textit{and} \quad d\lambda_j(c) = 0, \tag{4.34}$$

$$a_j(c)^T\xi = 0 \quad \textit{if} \quad d\lambda_j(c) > 0, \tag{4.35}$$

for all j, and

$$\underset{t\to c-0,\ \dot{y}(t)\to\xi}{\textit{limesssup}}\ \dot{y}(t)^T R(t)\dot{y}(t) > 0. \tag{4.36}$$

In particular, when both $\dot{y}(t)$ and $R(t)$ have the left limits at $t=c$, conditions (4.34), (4.35)*, and* (4.36) *reduce, respectively, to*

$$a_j(c)^T\dot{y}(c-0) \geq 0 \quad \textit{if} \quad \alpha_j(c) = 0 \quad \textit{and} \quad d\lambda_j(c) = 0, \tag{4.37}$$

$$a_j(c)^T\dot{y}(c-0) = 0 \quad \textit{if} \quad d\lambda_j(c) > 0, \tag{4.38}$$

and

$$\dot{y}(c-0)^T R(c-0)\dot{y}(c-0) > 0. \tag{4.39}$$

The following result is a necessary condition in terms of the conjugate point theory for the optimal value of (GAP) to be zero.

THEOREM 4.6. *Assume that* (GAP) *has a minimum value equal to zero. Then the open interval $(0,T)$ contains no points conjugate to $t=0$.*

Remark 4.1. Note that we do not assume that the strengthened Legendre condition holds and thus, $R(c-0)$ need not be positive definite. Hence, Theorem 4.6 is applicable to the shortest path problem in Euclidean space; see Example 7.2 below.

Proof of Theorem 4.6. Let (y,β) satisfy the Jacobi system (J1)–(J6) on $[0,c]$ and the end points condition (4.31). Let $(\bar{y},\bar{\beta})$, defined by (4.32), be the extensions of (y,β). Then, by Theorem 4.3, $(\bar{y},\bar{\beta})$ is a global minimum for the generalized accessory problem (GAP). Hence, by Theorem 4.1, there exist a function $\bar{\mu}(t)$ and a constant $\bar{d}$ that satisfy the Jacobi system (J1)–(J6). In particular, (J1) reduces to

$$Q(t)^Ty(t) + R(t)\dot{y}(t) - \int_0^t \{Py + Q\dot{y}\}dt - \sum_{j=1}^m \int_{(0,t]} \{C_jyd\lambda_j + a_jd\bar{\mu}_j\} = \bar{d} \tag{4.40}$$

a.e. on $[0,c]$, and

$$-\int_0^c \{Py + Q\dot{y}\}dt - \sum_{j=1}^m \int_{(0,c]} \{C_jyd\lambda_j + a_jd\bar{\mu}_j\} = \bar{d}. \tag{4.41}$$

By subtracting (4.41) from (4.40) and multiplying it by $\dot{y}(t)^T$, we get

$$\begin{aligned} &\dot{y}(t)^TQ(t)^Ty(t) + \dot{y}(t)^TR(t)\dot{y}(t) + \dot{y}(t)^T\int_t^c \{Py + Q\dot{y}\}dt \\ &+ \sum_{j=1}^m \dot{y}(t)^T\int_{(t,c]} C_jyd\lambda_j + \sum_{j=1}^m \dot{y}(t)^T\int_{(t,c]} a_jd\bar{\mu}_j = 0 \end{aligned} \tag{4.42}$$

a.e. on $[0,c]$. Here, since $\dot{y}(t)^T Q(t)^T$ is essentially bounded and $y(c)=0$, the first term of (4.42) converges to zero as $t \to c-0$. Similarly, the third and the fourth terms tend to zero. Now, let ξ be the essential cluster point in the definition of conjugate points, and let $t_k \to c-0$ satisfy (4.42) and $\dot{y}(t_k) \to \xi$. Then the last term in (4.42) converges to $\xi^T \sum_{j=1}^m a_j(c) d\bar{\mu}_j(c)$. So we get from (4.42) that

$$\lim_{k\to\infty} \dot{y}(t_k)^T R(t_k)\dot{y}(t_k) + \sum_{j=1}^{m} \xi^T a_j(c) d\bar{\mu}_j(c) = 0. \tag{4.43}$$

Hence it follows from (J2) and (4.35) that

$$\lim_{k\to\infty} \dot{y}(t_k)^T R(t_k)\dot{y}(t_k) + \sum_{\alpha_j(c)=0,\ d\lambda_j(c)=0} \xi^T a_j(c) d\bar{\mu}_j(c) = 0. \tag{4.44}$$

However, for any j satisfying $\alpha_j(c)=0$ and $d\lambda_j(c)=0$, we get from (4.34) and (J6) that $\xi^T a_j(c) d\bar{\mu}_j(c) \geq 0$, so that $\lim_{k\to\infty} \dot{y}(t_k)^T R(t_k)\dot{y}(t_k) \leq 0$, which contradicts (4.36). This completes the proof. □

In the rest of this section, we apply the concept of conjugate points and Theorem 4.6 to the problem (VP).

DEFINITION 4.7. *A point $c \in (0,T]$ is said to be conjugate to $t=0$ for* (VP) *if c is a conjugate point to $t=0$ for* (AP).

The following theorem is a necessary condition for optimality in (VP) in terms of the conjugate point theory. It is an immediate consequence of Theorems 2.3 and 4.6. We assume that $\bar{g}(t) \in W^m_{1,\infty}[0,T]$ and $\bar{g}_{jx}(t) \in W^n_{1,\infty}[0,T]$ for all j and that the full rank condition and (2.7) hold.

THEOREM 4.8 (main theorem). *Let $\bar{x}(t)$ be a weak minimum for the variational problem* (VP). *Then the open interval $(0,T)$ contains no points conjugate to $t=0$.*

5. Legendre condition for (VP). In the definition of conjugate points (Definition 4.5), we used a kind of Legendre condition. In this section, we justify it, that is, we will prove that the usual Legendre condition holds for (VP).

First, let $\bar{x}(t)$ be a weak minimum for (VP). Then there exists $\varepsilon > 0$ such that $F(\bar{x}) \leq F(\bar{x}+\varepsilon y)$ if $\bar{x}+\varepsilon y$ is feasible and $\parallel y \parallel < 1$, where F is the objective function of (VP) and $\parallel y \parallel$ is the norm in $W^n_{1,\infty}[0,T]$.

Hence, by putting $u := \dot{y}$, $(\bar{y}(t), \bar{u}(t)) \equiv (0,0)$ is a global minimum for the following optimal control problem:

$$\begin{aligned}
\text{(OCP) Minimize } & \int_0^T f(t, \bar{x}(t)+\varepsilon y(t), \dot{\bar{x}}(t)+\varepsilon u(t))dt \\
\text{subject to } & y(0)=y(T)=0, \quad (y,u) \in W^n_{1,\infty}[0,T] \times L^n_\infty[0,T], \\
& \dot{y} = u, \\
& g(t, \bar{x}(t)+\varepsilon y(t)) \leq 0 \quad \text{for all } t \in [0,T], \\
& |y(t)| < 1, \quad |u(t)| < 1,
\end{aligned}$$

where $|\cdot|$ designates the Euclidean norm.

THEOREM 5.1. *If $\bar{x}$ is a weak minimum for* (VP), *then $f_{\dot{x}\dot{x}}(t, \bar{x}(t), \dot{\bar{x}}(t))$ is nonnegative definite for a.e. t.*

Proof. By the maximum principle, see, e.g., Clarke [6, Thm. 5.2.1], there exist $\lambda_0 \in \{0,1\}$, m-dimensional vector-valued absolutely continuous p, and a nondecreasing function $\lambda : [0,T] \longrightarrow R^m$ that is right-continuous except $t=0$ such that

$(\lambda_0, p, d\lambda)$ are not all zero,

$$\dot{p}(t) = -H_y\left(t, \bar{y}(t), p(t) + \int_{(0,t]} \bar{g}_x^T d\lambda, \bar{u}(t), \lambda_0\right), \tag{5.1}$$

$$H\left(t, \bar{y}(t), p(t) + \int_{(0,t]} \bar{g}_x^T d\lambda, w, \lambda_0\right) \geq H\left(t, \bar{y}(t), p(t) + \int_{(0,t]} \bar{g}_x^T d\lambda, \bar{u}(t), \lambda_0\right) \tag{5.2}$$

for all $|w| < 1$ and a.e. t, and

$$d\lambda_j(t) = 0 \;\text{ if }\; \bar{g}_j(t) < 0 \tag{5.3}$$

for all j, where

$$H(t, y, q, w, \lambda_0) := q^T u - \lambda_0 f(t, \bar{x}(t) + \varepsilon y, \dot{\bar{x}}(t) + \varepsilon w) \tag{5.4}$$

for $(t, y, q, w, \lambda_0) \in [0, T] \times R^{3n+1}$. Since

$$H\left(t, \bar{y}(t), p(t) + \int_{(0,t]} \bar{g}_x^T d\lambda, w, \lambda_0\right) = \left(p(t) + \int_{(0,t]} \bar{g}_x^T d\lambda\right)^T w - \lambda_0 f(t, \bar{x}(t), \dot{\bar{x}}(t) + \varepsilon w) \tag{5.5}$$

takes the maximum at $w = 0$, we get

$$p(t) + \int_{(0,t]} \bar{g}_x^T d\lambda - \lambda_0 \varepsilon f_{\dot{x}}(t, \bar{x}(t), \dot{\bar{x}}(t)) = 0 \tag{5.6}$$

for a.e. t. Furthermore, by differentiating (5.5) twice w.r.t. w, we see that $\lambda_0 \varepsilon^2 f_{\dot{x}\dot{x}}(t, \bar{x}(t), \dot{\bar{x}}(t))$ is nonnegative definite for a.e. t. Finally, we show that $\lambda_0 \neq 0$. Indeed, if $\lambda_0 = 0$, then from (5.6) we have

$$p(t) + \int_{(0,t]} \bar{g}_x^T d\lambda = 0 \tag{5.7}$$

for a.e. t. For any $t \in [0, T]$, by choosing a suitable sequence t_k converging to t from the left and taking the limit in (5.7), we get

$$p(t) + \int_{(0,t)} \bar{g}_x^T d\lambda = 0, \tag{5.8}$$

where the commutativity of the limit with the integration is guaranteed by Lebesgue's convergence theorem. Similarly, by taking a suitable right limit in (5.7), we see that (5.7) holds for all $t \in [0, T]$. So, from (5.7) and (5.8), we have $\bar{g}_x(t)^T(\lambda(t) - \lambda(t-0)) = 0$ for all t. Since $\{\bar{g}_{jx}(t)\}_{\bar{g}_j(t)=0}$ are linearly independent and $d\lambda_j(t) = 0$ for any j such that $\bar{g}_j(t) < 0$, we see that $\lambda(t)$ does not jump at any point t. On the other hand, it follows from the adjoint equation (5.1) that $\dot{p} = -\lambda_0 \varepsilon \bar{f}_x = 0$. Since $p(0) = 0$, we have $p(t) = 0$. Next, differentiate (5.7). Then we have $\dot{p} + \bar{g}_x^T \dot{\lambda} = 0$. Hence $\bar{g}_x^T \dot{\lambda} = 0$, which implies that $\dot{\lambda} = 0$. This contradicts that $(\lambda_0, p, d\lambda)$ are not all zero. Therefore λ_0 is positive. This completes the proof. □

Remark 5.1. Theorem 5.1 above asserts that the usual Legendre condition holds even if there exists an inequality state constraint. This fact is surprising, because the result is different in the case of an equality state constraint. Namely, when the inequality constraint $g(t, x(t)) \leq 0$ is replaced with $h(t, x(t)) = 0$, we get only

$$\xi^T f_{\dot{x}\dot{x}}(t, \bar{x}(t), \dot{\bar{x}}(t))\xi \geq 0 \;\text{ if }\; h_x(t, \bar{x}(t))\xi = 0; \tag{5.9}$$

see, e.g., Hestenes [13, Thm. 5.1].

6. Conjugate points for ($\mathbf{VP_0}$) problem. In this section, we deal with the variational problem (VP_0) with equality state constraints. Since all the proofs are similar to and far simpler than those of the inequality case, we omit them. Assume that $\bar{h}_x(t)$ has full rank for all t.

THEOREM 6.1. *If $\bar{x}(t)$ is a weak minimum for* (VP_0)*, then there exists a constant vector $a \in R^n$ and a function $\lambda : [0,T] \longrightarrow R^m$ with bounded variation that is right-continuous except at $t = 0$ such that for each critical direction $y \in W^n_{1,\infty}[0,T]$, we have*

$$\bar{f}_{\dot{x}}(t) - \int_0^t \bar{f}_x ds - \int_{(0,t]} d\lambda^T \bar{h}_x = a^T \quad a.e.\ t \in [0,T], \tag{6.1}$$

$$\int_0^T \{y^T \bar{f}_{xx} y + 2y^T \bar{f}_{x\dot{x}} \dot{y} + \dot{y}^T \bar{f}_{\dot{x}\dot{x}} \dot{y}\} dt + \int_{[0,T]} y^T (d\lambda^T \bar{h})_{xx} y \geq 0. \tag{6.2}$$

The accessory problem for (VP_0) is given as follows:

$$(\mathrm{AP}^0)\ \text{Minimize}\ \frac{1}{2} \int_0^T \{y^T \bar{f}_{xx} y + 2y^T \bar{f}_{x\dot{x}} \dot{y} + \dot{y}^T \bar{f}_{\dot{x}\dot{x}} \dot{y}\} dt + \frac{1}{2} \int_{[0,T]} y^T (d\lambda^T \bar{h})_{xx} y$$

$$\text{subject to } y \in W^n_{1,\infty}[0,T], \quad \bar{h}_{jx}(t) y(t) \equiv 0, \quad y(0) = y(T) = 0.$$

THEOREM 6.2. *If y is a weak minimum for* (AP^0)*, then there exist $d \in R^n$ and $\mu : [0,T] \longrightarrow R^m$ of bounded variation that is right-continuous except at $t = 0$ such that the Jacobi system is satisfied, that is, a.e. t,*

($\mathrm{J1}_0$)

$$\bar{f}_{\dot{x}x}(t) y(t) + \bar{f}_{\dot{x}\dot{x}}(t) \dot{y}(t) - \int_0^t \{\bar{f}_{xx} y + \bar{f}_{x\dot{x}} \dot{y}\} dt - \sum_{j=1}^m \int_{(0,t]} \{\bar{h}_{jxx} y d\lambda_j + \bar{h}^T_{jx} d\mu_j\} = d,$$

a.e. t, and

$$\bar{h}_x(t) y(t) \equiv 0. \tag{$\mathrm{J2}_0$}$$

THEOREM 6.3. *Let c be in $(0,T]$, y satisfy* ($\mathrm{J1}_0$) *and* ($\mathrm{J2}_0$) *on $[0,c]$, and c satisfy the end points condition $y(0) = y(c) = 0$. Next, define $\bar{y}(t)$ by $y(t)$ on $[0,c]$, and 0 on $[c,T]$. Then $\bar{y}$ is also a feasible solution for* (AP^0)*, and the value of the objective function at $\bar{y}$ is zero.*

DEFINITION 6.4. *A point $c \in (0,T]$ is said to be conjugate to $t = 0$ if there exist a feasible solution $y \in W^m_{1,\infty}[0,T]$ for* (AP^0) *and an essential cluster point $\xi \in \dot{y}_{ess}(c-0)$ such that they satisfy, for some μ and d, the Jacobi system* ($\mathrm{J1}_0$)*,* ($\mathrm{J2}_0$) *on $[0,c]$, the end points condition $y(0) = y(c) = 0$, $\bar{h}_x(c)\xi = 0$, and $limesssup_{t \to c-0,\ \dot{y}(t) \to \xi}\, \dot{y}(t)^T \bar{f}_{\dot{x}\dot{x}}(t) \dot{y}(t) > 0$.*

THEOREM 6.5. *Let $\bar{x}(t)$ be a weak minimum for the variational problem* (VP_0)*. Then the open interval $(0,T)$ contains no points conjugate to $t = 0$.*

Remark 6.1. Traditionally, it was thought that inequality constraints could be treated in the context of equality constraints by adding a slack variable of the form v^2 (see, e.g., [13, p. 261] or [30, p. 148]). However, this method produces necessary conditions for optimality that are considerably weaker than those obtained via Theorem 2.1. For instance, if we add the slack variable $\frac{v^2}{2}$ to $g(t,x(t))$, the constraint $g(t,x(t)) \leq 0$ reduces to

$$g(t,x(t)) + \frac{v^2}{2}(t) = 0 \qquad \text{for all} \quad t \in [0,T]. \tag{6.3}$$

Now apply the results of this section to (VP), where $g(t,x(t)) \leq 0$ is replaced by (6.3). We obtain that the accessory problem (AP) has a zero minimum over a set of directions $(y,\beta) \in W^n_{1,\infty}[0,T] \times W^m_{1,\infty}[0,T]$ which is a strict subset of the set of critical directions used in section 3. In fact, (y,β) must now satisfy $y(0)=y(T)=0$ and, for all t and for all j,

$$a_j^T(t)y(t) + \alpha_j(t)\beta_j(t) = 0, \tag{6.4}$$

where $a_j^T(t) = \bar{g}_{jx}(t)$ and $\alpha_j(t) = \sqrt{-2\bar{g}_j(t)}$. Furthermore, the Jacobi system associated with (GAP) subject to (6.4) would be (J1), (J2), (J5), (J6), and (6.4). Hence, with this Jacobi system the number of conjugate points is smaller than that obtained by using (J1)–(J6), as in Definition 4.5. Thus, the necessary condition that we would obtain in terms of conjugate points is weaker than that given in Theorems 4.6 and 4.8.

7. Examples. We begin this section by introducing a connection condition. Usually, optimal solutions for (VP) have a couple of phases. For instance, in Figure 7.1 below, $\bar{x}(t)$ has two phases: straight line x_0P and sin-curve Px_T. This fact makes calculation of the Jacobi system, especially (J1), complicated. For example, when we compute the integrations in (J1) for t that correspond to a point in Px_T, we have to divide the interval of integration into two subintervals like $[0,t]=[0,\tau)\cup[\tau,t]$. However, we can avoid this division by introducing a connection condition at the boundary point $t=\tau$.

THEOREM 7.1. *Let y and μ satisfy* (J1), *and let τ be in* $(0,T)$. *Then it holds that*

(7.1)
$$\{Q^T(t)y(t) + R(t)\dot{y}(t)\}|_{t=\tau-0}^{t=\tau+0} = \sum_{j=1}^{m}\{C_j(\tau)y(\tau)\lambda_j(t)|_{t=\tau-0}^{t=\tau} + a_j(\tau)\mu_j(t)|_{t=\tau-0}^{t=\tau}\}.$$

We call (7.1) *the connection condition.*

Proof. Take the left and the right limits at τ in (J1). Then, since $\lambda(t)$ and $\mu(t)$ are right-continuous, we get the desired result by comparing the limits. □

With the connection condition, we do not need to take care of the value of the constant vector $d \in R^n$ in (J1). It suffices to use the fact that the left-hand side of (J1) is constant in each phase; see (7.7) below. Furthermore, we can easily derive another connection condition for the Euler–Lagrange equation (2.16).

$$\bar{f}_{\dot{x}}(t)|_{t=\tau-0}^{t=\tau+0} = \lambda(t)^T|_{t=\tau-0}^{t=\tau}\bar{g}_x(\tau). \tag{7.2}$$

Example 7.1.

$$\text{Minimize} \int_0^T f(t,x,\dot{x})dt = \frac{1}{2}\int_0^T (\dot{x}^2 - x^2)dt$$
$$\text{subject to } x(0)=0, \quad x(T)=-1, \quad x \in W_{1,\infty}[0,T],$$
$$g(t,x(t)) = l(t) - x(t) \leq 0 \quad \text{for all } t \in [0,T],$$

where $\pi \leq \tau < T < 2\pi$ are fixed and

$$l(t) := \min\left\{\frac{t-\pi}{\sin(T-\tau)},\ 0,\ \frac{\tau - t}{\sin(T-\tau)}\right\}.$$

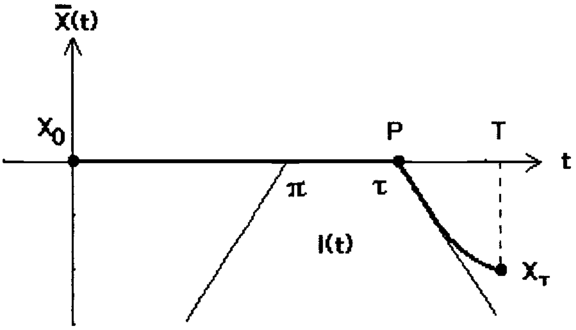


FIG. 7.1.

It is easily seen that the following pair satisfies the boundary condition and the Euler–Lagrange equation (2.16):

$$\bar{x}(t) := \begin{cases} 0 & \text{on } [0,\tau], \\ -\frac{\sin(t-\tau)}{\sin(T-\tau)} & \text{on } [\tau,T], \end{cases} \quad \text{and} \quad \lambda(t) := \begin{cases} 0 & \text{on } [0,\tau), \\ \frac{1}{\sin(T-\tau)} & \text{on } [\tau,T]. \end{cases}$$

Next, since $R(t) = \bar{f}_{\dot{x}\dot{x}}(t) = 1$, $Q(t) = \bar{f}_{x\dot{x}}(t) = 0$, $P(t) = \bar{f}_{xx}(t) = -1$, $C_1(t) = \bar{g}_{xx}(t) = 0$, and $a_1(t) = \bar{g}_x(t)^T = -1$, (J1) reduces to

$$\dot{y} + \int_0^t y\,dt + \mu(t) - \mu(0) = d. \tag{7.3}$$

On the other hand, it follows from (J2) that $d\mu = 0$ on $[0,\pi) \cup (\tau,T]$, so that $\ddot{y}+y = 0$ a.e. on $[0,\pi) \cup (\tau,T]$. Since $y(0) = 0$, we have $y(t) = A\sin t$ on $[0,\pi]$. Let us now see if $t = \pi$ is conjugate to 0 or not. By taking $\beta(t) \equiv 0$, we can easily see that the pair (y,β) satisfies the Jacobi system. Furthermore, $t = \pi$ satisfies (4.39) if $A \neq 0$. Indeed, if $A \neq 0$, then $R(\pi-0)\dot{y}(\pi-0)^2 = A^2 > 0$.

Case 1. In the case of $\tau > \pi$, it holds that $d\lambda(\pi) = 0$, $\bar{g}(\pi) = 0$, and $\bar{g}_x(\pi)\dot{y}(\pi - 0) = A$. Hence, by taking $A > 0$, we see that $t = \pi$ is conjugate to 0, and hence $\bar{x}$ is not a weak minimum for the problem.

Case 2. In the case of $\tau = \pi$, it holds that $d\lambda(\pi) > 0$ and $\bar{g}_x(\pi)\dot{y}(\pi - 0) = A$. Hence, by (4.38), A must be zero if $t = \pi$ is conjugate to 0. However, this contradicts (4.39). Hence $t = \pi$ is not conjugate to 0. Furthermore, since $y(t) = B\sin t$ for some $B \in R$ on $[\pi,T]$, $(\pi,T]$ contains no points conjugate to $t = 0$. Therefore $\bar{x}$ could be optimal.

Example 7.2. The problem is to find the shortest path which does not cross the inside of the unit ball $B \subset R^3$ and which joins two given points x^0, $x^1 \notin B$; see Figure 7.2. That is,

$$\text{Minimize} \int_{t_0}^{T} f(t,x(t),\dot{x}(t))dt = \int_{t_0}^{T} \sqrt{\dot{x}_1(t)^2 + \dot{x}_2(t)^2 + \dot{x}_3(t)^2}dt$$

$$\text{subject to } x(t_0) = x^0, \quad x(T) = x^1, \quad x \in W^n_{1,\infty}[t_0,T],$$

$$g(t,x(t)) = \frac{1}{2}(1 - x_1(t)^2 - x_2(t)^2 - x_3(t)^2) \leq 0 \quad \text{for all } t \in [t_0,T],$$

where $t_0 < 0$, $\frac{\pi}{2} < T$ are fixed.

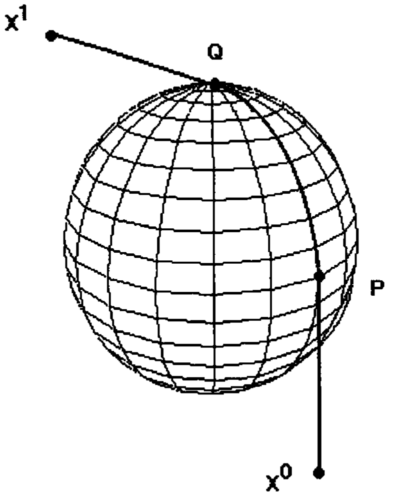


FIG. 7.2.

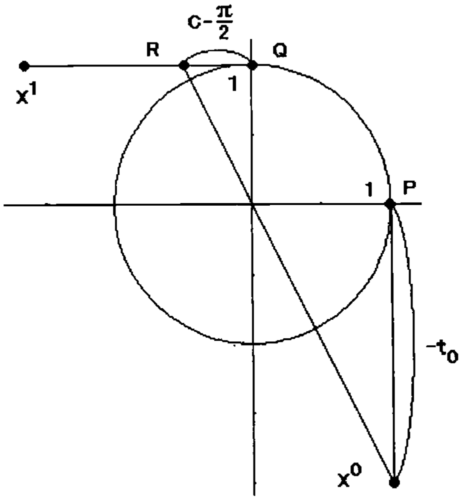


FIG. 7.3.

In this example, we consider the case where the endpoints are given by

$$x^0 = (1, 0, t_0)^T,\ x^1 = \Big(\frac{\pi}{2} - T, 0, 1\Big)^T. \tag{7.4}$$

As a candidate for an optimal solution, let us take

$$\bar{x}(t) := \begin{cases} (1, 0, t)^T & \text{on } t_0 \le t < 0, \\ (\cos t, 0, \sin t)^T & \text{on } 0 \le t < \frac{\pi}{2}, \\ (\frac{\pi}{2} - t, 0, 1)^T & \text{on } \frac{\pi}{2} \le t \le T. \end{cases} \tag{7.5}$$

Then it is easily seen that $\bar{x}(t)$ satisfies the Euler–Lagrange equation by taking

$$\lambda(t) := \begin{cases} 0 & \text{on } t_0 \le t < 0, \\ t & \text{on } 0 \le t < \frac{\pi}{2}, \\ \frac{\pi}{2} & \text{on } \frac{\pi}{2} \le t \le T. \end{cases} \tag{7.6}$$

Now, let us compute the Jacobi system.

Phase 1. $t_0 \le t < 0$. On this interval, it is easily seen that (J1) reduces to

$$\begin{pmatrix} \dot{y}_1(t) \\ \dot{y}_2(t) \\ 0 \end{pmatrix} + \int_{(t_0,t]} \begin{pmatrix} 1 \\ 0 \\ t \end{pmatrix} d\mu = d.$$

Since $y(t_0) = 0$, we see that $d\mu(t) \equiv 0$, $y_1(t) = d_1(t - t_0)$, $y_2(t) = d_2(t - t_0)$, and $y_3(t)$ can be arbitrary except $y_3(t_0) = 0$. Suppose that there exists a point c conjugate to t_0 in this interval. Then both $y_1(t)$ and $y_2(t)$ must vanish, so that $\dot{y}(c-0)^T \bar{f}_{\dot{x}\dot{x}}(c-0)\dot{y}(c-0) = \dot{y}_1(c-0)^2 + \dot{y}_2(c-0)^2 = 0$, which contradicts (4.39). Therefore there is no point conjugate to t_0 in this interval.

Phase 2. $0 \le t < \pi$. It is easily seen that (J1) reduces to

(7.7)
$$\begin{pmatrix} \cos^2 t & 0 & \sin t \cos t \\ 0 & 1 & 0 \\ \sin t \cos t & 0 & \sin^2 t \end{pmatrix} \begin{pmatrix} \dot{y}_1(t) \\ \dot{y}_2(t) \\ \dot{y}_3(t) \end{pmatrix} + \int_{(0,t]} \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix} dt + \int_{(0,t]} \begin{pmatrix} \cos t \\ 0 \\ \sin t \end{pmatrix} d\mu = c,$$

where $c \in R^3$ is a suitable constant vector. Here we note that the interval of integration is not $(t_0, t]$ but $(0, t]$. On the other hand, it follows from (J5) that $\bar{g}_x y dt = 0$. That is, $y_1(t)\cos t + y_3(t)\sin t = 0$. Hence there exists a real-valued function $\rho(t)$ such that

$$(y_1(t), y_3(t)) = \rho(t)(\sin t, -\cos t). \tag{7.8}$$

Hence $\rho(t)$ is absolutely continuous. Substituting (7.8) into (7.7), we get

$$\rho(t)\cos t + \int_{(0,t]} \rho(t)\sin t\, dt + \int_{(0,t]} \cos t d\mu = c_1 \tag{7.9}$$

and

$$\rho(t)\sin t - \int_{(0,t]} \rho(t)\cos t\, dt + \int_{(0,t]} \sin t d\mu = c_3. \tag{7.10}$$

By differentiating (7.9), we have $(\dot{\rho}(t) + \dot{\mu}(t))\cos t = 0$. Similarly, we get from (7.10) that $(\dot{\rho}(t) + \dot{\mu}(t))\sin t = 0$. Therefore

$$\dot{\rho} + \dot{\mu} = 0. \tag{7.11}$$

On the other hand, it follows from the second component of (7.7) that $\ddot{y}_2 + y_2 = 0$. Hence

$$y_2(t) = A\cos t + B\sin t, \tag{7.12}$$

where A and B are real constants.

Then it is easily seen that the connection condition (7.1) at $t = 0$ reduces to

$$d_1 = \dot{y}_1(0+0) = \rho(0), \quad d_2 = \dot{y}_2(0+0) = B. \tag{7.13}$$

Furthermore, since $y(t)$ is continuous at $t = 0$, we have

$$d_1 = 0, \ A = -d_2 t_0. \tag{7.14}$$

Combining (7.12), (7.13), and (7.14), we get

$$y_2(t) = B(\sin t - t_0 \cos t). \tag{7.15}$$

Now, suppose that there exists a point c conjugate to t_0 in this interval. Then we get from (7.15) that $\tan c = t_0 < 0$, which contradicts that $0 \leq c < \frac{\pi}{2}$. Therefore there is no point conjugate to t_0 in this interval.

Phase 3. $\frac{\pi}{2} \leq t \leq T$. As was the case in Phase 1, it follows from (J1) that $d\mu = 0$, $y_2(t)$ and $y_3(t)$ are straight lines, and $y_1(t)$ is arbitrary. If there exists a point c conjugate to t_0 in $[\frac{\pi}{2}, T]$, then y_2 and y_3 are given by $y_2(t) = b_2(t-c)$ and $y_3(t) = b_3(t-c)$, where b_k's are constants. Furthermore, it follows from the connection condition (7.1) and the continuity of $y(t)$ at $t = \frac{\pi}{2}$ that $b_2 = Bt_0$, $B = b_2(\frac{\pi}{2} - c)$, and $b_3 = 0$. Hence

$$B\left\{t_0\left(\frac{\pi}{2} - c\right) - 1\right\} = 0. \tag{7.16}$$

On the other hand, it follows from (4.39) that $0 < \dot{y}(c-0)^T \bar{f}_{\dot{x}\dot{x}}(c-0)\dot{y}(c-0) = \dot{y}_2^2(c-0)^2 + \dot{y}_3^2(c-0)^2 = b_2^2 = B^2 t_0^2$. Hence we get from (7.16) that

$$t_0\left(\frac{\pi}{2} - c\right) - 1 = 0. \tag{7.17}$$

In Figure 7.3, R indicates the point $\bar{x}(c)$. Since the inequality constraint is inactive at $t = c$, conditions (4.37) and (4.38) are trivially satisfied. Furthermore, it is easily seen that (J2)–(J6) are satisfied by taking $\beta(t) \equiv 0$. Therefore, when $T > \frac{\pi}{2} - \frac{1}{t_0}$, $\bar{x}(t)$ is not a weak minimal solution.

Acknowledgments. The authors would like to thank Michigan State University and the University of Michigan. They also thank Professor John R. Birge for the excellent research opportunity.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Plenum, New York, 1987.

[2] A. V. Arutynov, *On the theory of the maximum principle in optimal control problems with phase constraints*, Soviet Math. Dokl., 304 (1989), pp. 11–14.

[3] A. Ben-tal and J. Zowe, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Programming, 19 (1982), pp. 39–76.

[4] G. A. Bliss, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

[5] R. W. Chaney, *Second-order necessary conditions in semismooth optimization*, Math. Programming, 40 (1988), pp. 95–109.

[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.

[7] R. F. A. Clebsch, *Über die Reduction der zweiten Variation auf ihre einfachste Form*, J. Reine Angew. Math., 55 (1858), pp. 254–273.

[8] R. Cominetti, *Metric regularity, tangent sets, and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[9] A. L. Dontchev and I. V. Kolmanovsky, *State constraints in the linear regulator problem; Case study*, J. Optim. Theory. Appl., 87 (1995) pp. 323–347.

[10] M. M. A. Ferreira and R. B. Vinter, *When is the maximum principle for state constrained problems nondegenerate?*, J. Math. Anal. Appl., 187 (1994), pp. 438–467.

[11] I. V. Girsanov, *Lectures on Mathematical Theory of Extremum Problems*, Lecture Notes in Econom. and Math. Systems 67, Springer, New York, 1972.

[12] R. F. Hartl, S. P. Sethi, and R. G. Vickson, *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Rev., 37 (1995), pp. 181–218.

[13] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley and Sons, New York, 1966.

[14] A. Ioffe and V. M. Tikhomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.

[15] A. Ioffe, *On some recent developments in the theory of second order optimality conditions*, in Optimization, S. Dolezki, ed., Lecture Notes in Math. 1405, Springer, New York, 1989, pp. 55–68.

[16] A. Ioffe, *Variational analysis of a composite function: A formula for the lower second order epi-derivative*, J. Math. Anal. Appl., 160 (1991), pp. 379–405.

[17] D. H. Jacobson, M. M. Lele, and J. L. Speyer, *New necessary conditions of optimality for control problems with state-variable inequality constraints*, J. Math. Anal. Appl., 35 (1971), pp. 255–284.

[18] H. Kawasaki, *An envelope-like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Programming, 41 (1988), pp. 73–96.

[19] H. Kawasaki, *The upper and lower second order directional derivatives of a sup-type function*, Math. Programming, 41 (1988), pp. 327–339.

[20] H. Kawasaki, *Second order necessary optimality conditions for minimizing a sup-type function*, Math. Programming, 49 (1991), pp. 213–229.

[21] H. Kawasaki, *Second order necessary and sufficient optimality conditions for minimizing a sup-type function*, Appl. Math. Optim., 26 (1992), pp. 195–220.

[22] H. Kawasaki, *A second-order property of spline functions with one free knot*, J. Approx. Theory, 78 (1994), pp. 293–297.

[23] H. Kawasaki, *First- and second-order directional derivatives of a max-type function induced from an inequality state constraint*, Bull. Inform. Cybernet., 29 (1997), pp. 41–49.

[24] H. Kawasaki, *Optimization and best approximation*, Sugaku Expositions, 10 (1997), pp. 1–17.

[25] H. Kawasaki and S. Koga, *Legendre conditions for a variational problem with one-sided inequality phase constraints*, J. Oper. Res. Soc. Japan, 38 (1995), pp. 483–492.

[26] S. Koga and H. Kawasaki, *Legendre-type optimality conditions for a variational problem with inequality state constraints*, Math. Program., 84 (1999), pp. 421–434.

[27] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.

[28] K. Malanowski, *Sufficient optimality conditions for optimal control subject to state constraints*, SIAM J. Control Optim., 35 (1997), pp. 205–227.

[29] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.

[30] A. A. Milyutin and N. P. Osmolovskii, *Calculus of Variations and Optimal Control*, American Mathematical Society, Providence, RI, 1998.

[31] P. D. Loewen and R. T. Rockafellar, *The adjoint arc in nonsmooth optimization*, Trans. AMS, 325 (1991), pp. 39–72.

[32] Zs. Páles and V. M. Zeidan, *Nonsmooth optimum problems with constraints*, SIAM J. Control Optim., 32 (1994), pp. 1476–1502.

[33] Zs. Páles and V. M. Zeidan, *First and second order necessary conditions for control problems with constraints*, Trans. Amer. Math. Soc., 346 (1994), pp. 421–453.

[34] Zs. Páles and V. M. Zeidan, *Optimum problems with certain lower semicontinuous set-valued constraints*, SIAM J. Optim., 8 (1998), pp. 707–727.

[35] J.-P. Penot, *On regularity conditions in mathematical programming*, Math. Programming, 19 (1982), pp. 167–199.

[36] J. P. Penot, *Optimality conditions in mathematical programming and composite optimization*, Math. Programming, 67 (1994), pp. 225–245.

[37] L. S. Pontryagin, V. G. Boltjanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience Publishers John Wiley and Sons,

New York, 1962.

[38] R. T. Rockafellar, *State constraints in convex control problems of Bolza*, SIAM J. Control, 10 (1972), pp. 691–715.

[39] R. T. Rockafellar, *Favorable classes of Lipschitz-continuous functions in subgradient optimization*, in II ASA Collaborative Proc. Ser. CP-82-S8, Internat. Inst. Appl. Systems Anal., Laxenburg, 1982, pp. 125–143.

[40] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, Tokyo, 1970.

[41] D. Ward, *Calculus for parabolic second-order derivatives*, Set-Valued Anal., 1 (1993), pp. 213–246.

[42] D. Ward, *Metric Regularity and Second-Order Nonsmooth Calculus*, preprint.

[43] V. M. Zeidan and P. Zezza, *Coupled points in the calculus of variations and applications to periodic problems*, Trans. AMS, 315 (1989) pp. 323–335.

[44] V. M. Zeidan and P. Zezza, *Coupled points in optimal control theory*, IEEE Trans. Automat. Control, 36 (1991), pp. 1276–1281.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 457–477

ANALYSIS OF NEUMANN BOUNDARY OPTIMAL CONTROL PROBLEMS FOR THE STATIONARY BOUSSINESQ EQUATIONS INCLUDING SOLID MEDIA[*]

HYUNG-CHUN LEE[†] AND O. YU. IMANUVILOV[‡]

Abstract. This article deals with Neumann boundary optimal control problems associated with the Boussinesq equations including solid media. These problems are first put into an appropriate mathematical formulation. Then the existence of optimal solutions is proved. The use of Lagrange multiplier techniques is justified and an optimality system of equations is derived.

Key words. flow control, temperature control, Boussinesq equations, optimization

AMS subject classifications. 49J20, 76D05, 49B22

PII. S0363012998347110

1. Introduction.

In this article we study boundary optimal control problems for a steady natural convection fluid. The control is heat flux on a portion of the boundary.

We consider the nondimensional Boussinesq equations (including solid media) as follows:

$$-Pr\Delta\mathbf{u}+(\mathbf{u}\cdot\nabla)\mathbf{u}=-\nabla p+Pr\,Ra\,T\mathbf{e}_2+\mathbf{f}\quad\text{in }\Omega_f,\tag{1.1}$$

$$\nabla\cdot\mathbf{u}=0\quad\text{in }\Omega_f,\tag{1.2}$$

$$-\nabla\cdot(\kappa\nabla T)+(\mathbf{u}\cdot\nabla)T=Q,\quad\text{in }\Omega\tag{1.3}$$

with boundary conditions

$$\mathbf{u}=\mathbf{0}\quad\text{on }\partial\Omega_f,\quad\mathbf{u}\equiv\mathbf{0}\quad\text{in }\Omega-\Omega_f=\Omega_s,\tag{1.4}$$

$$T=0\quad\text{on }\Gamma_D,\quad\frac{\partial T}{\partial\mathbf{n}}=g\quad\text{on }\Gamma_C,\tag{1.5}$$

where the regular bounded open set Ω in $\mathbb{R}^2$ is made up of two subdomains Ω_f (fluid domain) and Ω_s (solid domain) separated by a C^∞, connected arc (a regular hypersurface) Σ, with the result that $\Omega=\Omega_f\cup\Sigma\cup\Omega_s$. *Moreover, we assume that* Σ *intersects* $\partial\Omega$ *in two junction points* x_-,x_+ *under nonzero angles. We have, in addition,* $\overline{\partial\Omega_f\cap\partial\Omega_s}=\overline{\Sigma}$. *In* (1.5), $\Gamma_D=\partial\Omega\backslash\Gamma_C$, *where* Γ_C *is a regular open subset of* $\partial\Omega$ *such that* $x_-,x_+\notin\overline{\Gamma}_C$. In (1.1)–(1.5), $\mathbf{u}$, p, and T denote the velocity, pressure, and temperature fields, respectively, with $\mathbf{f}$ a given body force, Q a given heat source, and

[*] Received by the editors November 8, 1998; accepted for publication (in revised form) May 12, 2000; published electronically September 7, 2000.
http://www.siam.org/journals/sicon/39-2/34711.html

[†] Department of Mathematics, Ajou University, Suwon, Korea 442-749 (hclee@madang.ajou.ac.kr). The research of this author was supported by KOSEF 97-07-01-01-01-3 and the BK21 project.

[‡] School of Mathematics, KIAS, Seoul 130-012, Korea. Current address: Department of Mathematics, Iowa State University, Ames, IA 50011 (vika@iastate.edu). The research of this author was supported by KIAS (grant KIAS-M97003).

control g. The vector $\mathbf{e}_2$ is a unit vector in the direction of gravitational acceleration and κ is a thermal conductivity parameter. In this article, we consider the case of $\kappa \equiv \kappa_f$ in Ω_f and $\kappa \equiv \kappa_s$ in Ω_s, where κ_f and κ_s are positive constants. The vector $\mathbf{n}$ denotes the outward unit normal to Ω, and Pr and Ra denote the Prandtl and Rayleigh numbers, respectively.

Next, we introduce the functionals

$$\mathcal{J}_1(\mathbf{u}, T, p, g) = \frac{1}{2}\int_{\Omega} |T - T_d|^2 \, d\mathbf{x} + \frac{\delta}{2}\int_{\Gamma_C} |g|^2 \, ds \tag{1.6}$$

and

$$\mathcal{J}_2(\mathbf{u}, T, p, g) = \frac{1}{2}\int_{\Omega} |\nabla \times \mathbf{u}|^2 \, d\mathbf{x} + \frac{\delta}{2}\int_{\Gamma_C} |g|^2 \, ds. \tag{1.7}$$

The optimal control problem we consider is to seek state variables $(\mathbf{u}, p, T)$ and a control g such that the functional (1.6) or (1.7) is minimized subject to (1.1)–(1.5), where T_d is some desired temperature distribution. The functional (1.6) effectively measures the difference between the temperature field T and a prescribed field T_d. The real goal of optimization is to minimize the first term appearing in the definition (1.6). The functional (1.7) measures the vorticity of the flow. The control of vorticity has significant applications in science and engineering, such as control of turbulence and control of crystal growth process. The second terms in the cost functionals (1.6) and (1.7) are added to limit the cost of controls. The positive penalty parameter δ can be used to change the relative importance of the two terms appearing in the definition of the functional.

In past years, considerable progress has been made in mathematical analyses and computations of optimal control problems for viscous flows. Optimal control problems for the viscous, incompressible Navier–Stokes equations have been studied very actively during the last ten years (see [8, 9, 10, 13, 14, 15, 21] and references therein). An optimal control problem for the thermally coupled incompressible Navier–Stokes equations by Neumann boundary heat control is considered in [12] in which the Navier–Stokes equations and heat equations are not fully coupled. In [16], an optimal control problem for a coupled solid/fluid temperature control is considered. Linear feedback control of Boussinesq equations is considered in [22]. Also, control problems for the time dependent Boussinesq equations and related problems are considered in [2, 4, 17, 18, 20]. In this article, we consider optimal control problems for the stationary Boussinesq equations including solid media.

The plan of the paper is as follows. In the remainder of this section, we introduce the notation that will be used throughout the paper. Then, in section 2, we give a precise statement of a weak formulation of the Boussinesq equations and prove that a sufficiently smooth solution to the Boussinesq equations exists. In section 3, we give a precise statement of the optimization problem and prove that an optimal solution exists. In section 4, we prove the existence of Lagrange multipliers and then use the method of Lagrange multipliers to derive an optimality system. Some remarks and further discussions are also given.

1.1. Notation. We introduce some function spaces and their norms, along with some related notations used in subsequent sections; for details see [1].

Let Ω be a bounded domain of $\mathbb{R}^2$ with a Lipschitz continuous boundary Γ. Let $L^2(\Omega)$ be the space of real-valued square integrable functions defined on Ω, and let

$\|\cdot\|_{L^2(\Omega)}$ be the norm in this space. We define the Sobolev space $H^m(\Omega)$ for the nonnegative integer m by

$$H^m(\Omega) \overset{def}{=} \left\{u \in L^2(\Omega) \mid D^\alpha u \in L^2(\Omega) \text{ for } 0 \le |\alpha| \le m\right\},$$

where D^α is the weak (or distributional) partial derivative and α is a multi-index. The norm $\|\cdot\|_{H^m(\Omega)}$ associated with $H^m(\Omega)$ is given by

$$\|u\|^2_{H^m(\Omega)} = \sum_{|\alpha|\le m} \|D^\alpha u\|^2_{L^2(\Omega)}.$$

Note that $H^0(\Omega) = L^2(\Omega)$. For the vector-valued functions, we define the Sobolev space $\mathbf{H}^m(\Omega)$ (in all cases, boldface indicates vector-valued) by

$$\mathbf{H}^m(\Omega) \overset{def}{=} \{\mathbf{u} = (u_1, u_2) \mid u_i \in H^m(\Omega) \text{ for } i = 1, 2\},$$

and its associated norm $\|\cdot\|_{\mathbf{H}^m(\Omega)}$ is given by

$$\|\mathbf{u}\|^2_{\mathbf{H}^m(\Omega)} = \sum_{i=1}^{2} \|u_i\|^2_{H^m(\Omega)}.$$

We also define particular subspaces:

$$L^2_0(\Omega) = \left\{f \in L^2(\Omega) : \int_\Omega f \, d\mathbf{x} = 0\right\}, \quad \mathbf{H}^1_0(\Omega) = \left\{\mathbf{u} \in \mathbf{H}^1(\Omega) : \mathbf{u} = \mathbf{0} \text{ on } \Gamma\right\},$$

and

$$H^1_D(\Omega) = \left\{S \in H^1(\Omega) : S = 0 \text{ on } \Gamma_D\right\}.$$

We make use of the well-known space $\mathbf{L}^4(\Omega)$ equipped with the norm $||\cdot||_{\mathbf{L}^4(\Omega)}$.

We also define the solenoidal spaces

$$\mathbf{V} \overset{def}{=} \left\{\mathbf{u} \in \mathbf{H}^1_0(\Omega_f) \,|\, \nabla \cdot \mathbf{u} = 0\right\}.$$

If Ω is bounded and has a Lipschitz continuous boundary (these are kinds of domains under consideration here), Sobolev's embedding theorem yields that $H^1(\Omega) \hookrightarrow\hookrightarrow L^4(\Omega)$, where $\hookrightarrow\hookrightarrow$ denotes compact embedding, i.e., a constant C exists such that

$$||u||_{L^4(\Omega)} \le C||u||_{H^1(\Omega)}. \tag{1.8}$$

Obviously a similar result holds for the spaces $\mathbf{H}^1(\Omega)$ and $\mathbf{L}^4(\Omega)$.

2. A weak formulation of the Boussinesq equations. We introduce the following bilinear and trilinear forms for $\mathbf{u}, \mathbf{v}$ and $\mathbf{w} \in \mathbf{H}^1(\Omega_f)$, $T, S \in H^1(\Omega)$:

$$\begin{aligned}
a_0(\mathbf{u}, \mathbf{v}) &= \int_{\Omega_f} \nabla\mathbf{u} : \nabla\mathbf{v} \, d\mathbf{x} \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathbf{H}^1(\Omega_f),\\
a_1(T, S) &= \int_\Omega \kappa\nabla T \cdot \nabla S \, d\mathbf{x} \quad \forall\, T, S \in H^1(\Omega),\\
b(\mathbf{v}, q) &= -\int_{\Omega_f} q\nabla \cdot \mathbf{v} \, d\mathbf{x} \quad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega_f),\ \forall\, q \in L^2(\Omega_f),\\
c_0(\mathbf{u}, \mathbf{w}, \mathbf{v}) &= \int_{\Omega_f} (\mathbf{u} \cdot \nabla)\mathbf{w} \cdot \mathbf{v} \, d\mathbf{x} \quad \forall\, \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{H}^1(\Omega_f),\\
c_1(\mathbf{u}, T, S) &= \int_{\Omega_f} (\mathbf{u} \cdot \nabla)T\, S \, d\mathbf{x} \quad \forall\, \mathbf{u} \in \mathbf{H}^1(\Omega_f),\ \forall\, T, S \in H^1(\Omega),
\end{aligned}$$

and

$$d(T,\mathbf{v}) = \int_{\Omega_f} T\mathbf{e}_2 \cdot \mathbf{v}\, d\mathbf{x} \quad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega_f),\ \forall\, T \in H^1(\Omega).$$

We first note that the bilinear forms $a_0(\cdot,\cdot)$ and $a_1(\cdot,\cdot)$ are clearly continuous, i.e.,

$$|a_0(\mathbf{u},\mathbf{v})| \leq ||\mathbf{u}||_{\mathbf{H}^1(\Omega_f)}||\mathbf{v}||_{\mathbf{H}^1(\Omega_f)} \tag{2.1}$$

and

$$|a_1(T,S)| \leq \max(\kappa_f,\kappa_s)||T||_{H^1(\Omega)}||S||_{H^1(\Omega)}. \tag{2.2}$$

We have the coercivity relations associated with $a_0(\cdot,\cdot)$ and $a_1(\cdot,\cdot)$:

$$a_0(\mathbf{u},\mathbf{u}) = ||\nabla\mathbf{u}||^2_{L^2(\Omega_f)} \geq C_1||\mathbf{u}||^2_{\mathbf{H}^1(\Omega_f)} \qquad \forall\, \mathbf{u} \in \mathbf{H}^1_0(\Omega_f) \tag{2.3}$$

and

$$a_1(T,T) \geq \min(\kappa_f,\kappa_s)||\nabla T||^2_{L^2(\Omega)} \geq C_2||T||^2_{H^1(\Omega)} \qquad \forall\, T,S \in H^1_D(\Omega), \tag{2.4}$$

which are direct consequences of Poincaré inequality.

2.1. A weak formulation of the equations. The weak form of the constraint equations (1.1)–(1.5) is then given as follows: seek $\mathbf{u} \in \mathbf{H}^1_0(\Omega_f)$, $p \in L^2_0(\Omega_f)$, and $T \in H^1_D(\Omega)$ such that

$$\nu\, a_0(\mathbf{u},\mathbf{v}) + c_0(\mathbf{u},\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = \alpha\, d(T,\mathbf{v}) + \langle \mathbf{f},\mathbf{v}\rangle \quad \forall\, \mathbf{v} \in \mathbf{H}^1_0(\Omega_f), \tag{2.5}$$

$$b(\mathbf{u},q) = 0 \quad \forall\, q \in L^2_0(\Omega_f), \tag{2.6}$$

and

$$a_1(T,S) + c_1(\mathbf{u},T,S) = \langle Q,S\rangle + \int_{\Gamma_C} \kappa\, g\, S\, ds \quad \forall\, S \in H^1_D(\Omega). \tag{2.7}$$

Throughout the mathematical discussions, for the sake of convenience we set $\nu = Pr$ and $\alpha = Pr \times Ra$, which are not to be confused with the physical quantities such as kinematic viscosity.

LEMMA 2.1. *Suppose* $0 < \min(\kappa_f,\kappa_s) \leq \max(\kappa_f,\kappa_s) < \infty$. *Then, for* $\mathbf{u},\mathbf{v},\mathbf{w} \in \mathbf{H}^1(\Omega_f)$ *and* $T,S \in H^1(\Omega)$ *there are constants* $C_{1,2,3,4}$ *such that*

$$|c_0(\mathbf{u},\mathbf{w},\mathbf{v})| \leq C_1||\mathbf{u}||_{\mathbf{H}^1(\Omega_f)}||\mathbf{v}||_{\mathbf{H}^1(\Omega_f)}||\mathbf{w}||_{\mathbf{H}^1(\Omega_f)}, \tag{2.8}$$

$$c_0(\mathbf{u},\mathbf{v},\mathbf{v}) = 0 \quad \textit{if}\ \mathbf{u} \in \mathbf{V}, \tag{2.9}$$

$$|c_1(\mathbf{u},T,S)| \leq C_2||\mathbf{u}||_{\mathbf{H}^1(\Omega_f)}||T||_{H^1(\Omega)}||S||_{H^1(\Omega)} \quad \forall\, \mathbf{u} \in \mathbf{V}, \tag{2.10}$$

$$c_1(\mathbf{u},T,T) = 0 \quad \textit{if} \quad \mathbf{u} \in \mathbf{V}, \tag{2.11}$$

and

$$|d(T,\mathbf{u})| \leq C_3||T||_{L^2(\Omega)}||\mathbf{u}||_{\mathbf{L}^2(\Omega_f)} \leq C_4||\nabla T||_{L^2(\Omega)}||\nabla\mathbf{u}||_{\mathbf{L}^2(\Omega_f)}. \tag{2.12}$$

Proof. The first and third inequalities follow from the Hölders inequalities and the continuous embeddings of $\mathbf{H}^1$ into $\mathbf{L}^4$ and $\mathbf{L}^2$ and H^1 into L^4 and L^2, respectively. We obtain

$$\begin{aligned}|c_0(\mathbf{u},\mathbf{w},\mathbf{v})| &\leq ||\mathbf{u}||_{\mathbf{L}^4(\Omega_f)}||\nabla\mathbf{v}||_{\mathbf{L}^2(\Omega_f)}||\mathbf{w}||_{\mathbf{L}^4(\Omega_f)}\\ &\leq C_1||\mathbf{u}||_{\mathbf{H}^1(\Omega_f)}||\mathbf{v}||_{\mathbf{H}^1(\Omega_f)}||\mathbf{w}||_{\mathbf{H}^1(\Omega_f)}\end{aligned}$$

and

$$\begin{aligned}|c_1(\mathbf{u},T,S)| &\leq ||\mathbf{u}||_{\mathbf{L}^4(\Omega_f)}||\nabla T||_{L^2(\Omega)}||\nabla S||_{L^2(\Omega)}\\ &\leq C_2||\mathbf{u}||_{\mathbf{H}^1(\Omega_f)}||T||_{H^1(\Omega)}||S||_{H^1(\Omega)}.\end{aligned}$$

The second and fourth equalities follow from Green's formulas

$$c_0(\mathbf{u},\mathbf{v},\mathbf{v}) = \frac{1}{2}(\mathbf{u},\nabla(\mathbf{v}\cdot\mathbf{v})) = \frac{1}{2}(\mathbf{n}\cdot\mathbf{u},\mathbf{v}\cdot\mathbf{v})_{\partial\Omega_f} = 0$$

and

$$c_1(\mathbf{u},T,T) = \frac{1}{2}\int_{\partial\Omega_f}(\mathbf{u}\cdot\mathbf{n})T^2\,ds = 0,$$

provided that $\nabla\cdot\mathbf{u} = 0$ and $\mathbf{u}\in\mathbf{V}$. The last inequality follows from the Cauchy Schwarz inequality. □

We now show that at least one solution always exists for data $g\in L^2(\Gamma_C)$, $Q\in L^2(\Omega)$, and $\mathbf{f}\in\mathbf{L}^2(\Omega_f)$. Further, that solution is unique for either small data or an equivalent restriction on the Rayleigh and Prandtl numbers. Stationary boundary value problems (1.1)–(1.5) were studied by many authors (see [3, 5] and references therein). Here, we adapt and/or extend their results with suitable modifications.

LEMMA 2.2 (Leray–Schauder). *Let E be a Banach space, and let $G:[0,1]\times E\to E$ be a continuous, compact map, such that $G(0,v) = v_0$ is independent of $v\in E$. Suppose that there exists $M<\infty$ such that, for all $(\sigma,x)\in[0,1]\times E$,*

$$G(\sigma,x) = x \implies ||x|| < M.$$

Then the map $G_1 : E\to E$ given by $G_1(v) = G(1,v)$ has a fixed point.

Proof. For the proof, see [7, Theorem 8.1, p. 57]. □

THEOREM 2.3. *For every $g\in L^2(\Gamma_C)$, $Q\in L^2(\Omega)$, and $\mathbf{f}\in\mathbf{L}^2(\Omega_f)$, the Boussinesq equations* (2.5)–(2.7) *have a solution $(\mathbf{u},T,p)\in\mathbf{V}\times H^1(\Omega)\times L^2_0(\Omega_f)$ satisfying the estimate*

$$\begin{aligned}&||\mathbf{u}||_{\mathbf{H}^1(\Omega_f)} + ||p||_{L^2(\Omega_f)} + ||T||_{H^1(\Omega)}\\ &\qquad\leq C\left(||\mathbf{f}||_{\mathbf{L}^2(\Omega_f)} + ||Q||_{L^2(\Omega)} + ||g||_{L^2(\Gamma_C)}\right).\end{aligned} \tag{2.13}$$

Proof. From (2.2), (2.4), (2.10), and (2.11), it follows that for $\mathbf{u}\in\mathbf{V}$, $a_1(\cdot,\cdot) + c_1(\mathbf{u},\cdot,\cdot)$ is a continuous and elliptic bilinear form on $H^1_0(\Omega)\times H^1_0(\Omega)$, and thus on $H^1_D(\Omega)\times H^1_D(\Omega)$. Thus, for given $g\in L^2(\Gamma_C)$ and $Q\in L^2(\Omega)$, by the Lax–Milgram lemma and trace theorems there is a unique solution $T \in H^1_D(\Omega)$ satisfying (2.7) and the estimate

$$||T||_{H^1(\Omega)} + ||T||_{L^2(\Gamma_C)} \leq C \left(||g||_{L^2(\Gamma_C)} + ||Q||_{L^2(\Omega)}\right). \tag{2.14}$$

Thus, we may define a mapping $F : \mathbf{V} \to H^1(\Omega)$ by $F(\mathbf{u}) = T$. The theorem will be proved if one can show that there is at least one $\mathbf{u} \in \mathbf{V}$ such that

$$\nu\, a_0(\mathbf{u}, \mathbf{v}) + c_0(\mathbf{u}, \mathbf{u}, \mathbf{v}) = \alpha\, d(F(\mathbf{u}), \mathbf{v}) + \langle \mathbf{f}, \mathbf{v}\rangle \quad \forall\, \mathbf{v} \in \mathbf{V}. \tag{2.15}$$

From inequality (2.3) it follows that $a_0(\cdot, \cdot)$ is a continuous and elliptic bilinear form on $\mathbf{V} \times \mathbf{V}$ and

$$\begin{aligned} &| - c_0(\mathbf{u}, \mathbf{u}, \mathbf{v}) + d(F(\mathbf{u}), \mathbf{v}) + \langle \mathbf{f}, \mathbf{v}\rangle| \\ &\qquad\qquad \leq (C_2||\mathbf{u}||^2_{\mathbf{H}^1(\Omega_f)} + \alpha\, C_4||F(\mathbf{u})||_{H^1(\Omega)} + ||\mathbf{f}||_{\mathbf{L}^2(\Omega_f)})\; ||\mathbf{v}||_{\mathbf{H}^1(\Omega_f)} \end{aligned}$$

for all $\mathbf{v} \in \mathbf{V}$ follows from (2.10) and (2.12). Thus we may define a mapping $G : \mathbf{V} \to \mathbf{V}$ by

$$\nu\, a_0(G(\mathbf{u}), \mathbf{v}) = -c_0(\mathbf{u}, \mathbf{u}, \mathbf{v}) + \alpha\, d(F(\mathbf{u}), \mathbf{v}) + \langle \mathbf{f}, \mathbf{v}\rangle \quad \forall\, \mathbf{v} \in \mathbf{V}. \tag{2.16}$$

Clearly, $\mathbf{u}$ is a solution of (2.15) if it is a solution of

$$G(\mathbf{u}) = \mathbf{u}. \tag{2.17}$$

Now we may apply the Leray–Schauder Principle to prove the existence of the solution to (2.17). First, we verify the compactness of G. Let $\mathbf{u}_1, \mathbf{u}_2 \in \mathbf{V}$. Set $\mathbf{w} = G(\mathbf{u}_2) - G(\mathbf{u}_1)$. Subtracting the equations obtained from (2.16) by substituting $\mathbf{u}_2$ and $\mathbf{u}_1$ for $\mathbf{u}$ and $\mathbf{w}$ for $\mathbf{v}$, we get

$$\begin{aligned} \nu\, a_0(\mathbf{w}, \mathbf{w}) = &-c_0(\mathbf{u}_2 - \mathbf{u}_1; \mathbf{u}_2, \mathbf{w}) \\ &\qquad + c_0(\mathbf{u}_1; \mathbf{u}_2 - \mathbf{u}_1, \mathbf{w}) + \alpha\, d(F(\mathbf{u}_2) - F(\mathbf{u}_1), \mathbf{w}). \end{aligned} \tag{2.18}$$

Now we estimate $||F(\mathbf{u}_2) - F(\mathbf{u}_1)||_{H^1(\Omega)}$. Substitute $\mathbf{u}_2$ and $\mathbf{u}_1$ in (2.7) and subtract to get

$$\begin{aligned} a_1(F(\mathbf{u}_2) - F(\mathbf{u}_1), S) = &- c_1(\mathbf{u}_2 - \mathbf{u}_1; F(\mathbf{u}_2), S) \\ &- c_1(\mathbf{u}_1; F(\mathbf{u}_2) - F(\mathbf{u}_1), S) \quad \forall\, S \in H^1_D(\Omega). \end{aligned} \tag{2.19}$$

Substituting $F(\mathbf{u}_2) - F(\mathbf{u}_1)$ for S and using (2.4), (2.10), and (2.11), we have

$$||\nabla F(\mathbf{u}_2) - \nabla F(\mathbf{u}_1)||_{L^2(\Omega)} \leq C(||g||_{L^2(\Gamma_C)} + ||Q||_{L^2(\Omega)})\; ||\mathbf{u}_2 - \mathbf{u}_1||_{\mathbf{L}^4(\Omega_f)}. \tag{2.20}$$

Thus

$$\begin{aligned} ||\nabla \mathbf{w}||_{\mathbf{L}^2(\Omega_f)} \leq \nu^{-1}(&||\mathbf{u}_2||_{\mathbf{L}^4(\Omega_f)} + ||\mathbf{u}_1||_{\mathbf{L}^4(\Omega_f)} \\ &+\alpha\, C(||g||_{L^2(\Gamma_C)} + ||Q||_{L^2(\Omega)}))\; ||\mathbf{u}_2 - \mathbf{u}_1||_{\mathbf{L}^4(\Omega_f)} \end{aligned}$$

follows from (2.18) and (2.20) using (2.3), (2.8), and (2.11). Since $\mathbf{H}^1_0(\Omega_f)$ is compactly embedded in $\mathbf{L}^4(\Omega_f)$ and hence so is $\mathbf{V}$, it follows that G is a continuous compact map.

Now we define $G(\sigma, \mathbf{v}) = \sigma G(\mathbf{v})$ for all $(\sigma, \mathbf{v}) \in [0, 1] \times \mathbf{V}$. Clearly, $G(0, \mathbf{v}) = \mathbf{0}$ is independent of $\mathbf{v}$.

Suppose $\sigma \in (0,1]$ and $\mathbf{v} \in \mathbf{V}$ satisfies $\sigma G(\mathbf{v}) = \mathbf{v}$. Then

$$\sigma^{-1}\, \nu\, a_0(\mathbf{v},\mathbf{v}) = -c_0(\mathbf{v};\mathbf{v},\mathbf{v}) + \alpha\, d(F(\mathbf{v}),\mathbf{v}) + \langle \mathbf{f},\mathbf{v}\rangle. \tag{2.21}$$

From the above fact, we have

$$\begin{aligned} ||\nabla \mathbf{v}||_{\mathbf{L}^2(\Omega_f)} &\leq \sigma\, \left(\frac{\alpha}{\nu}\, C_4 ||\nabla F(\mathbf{v})||_{L^2(\Omega)} + ||\mathbf{f}||_{\mathbf{L}^2(\Omega_f)}\right) \\ &\leq C\left(||g||_{L^2(\Gamma_C)} + ||Q||_{L^2(\Omega)} + ||\mathbf{f}||_{\mathbf{L}^2(\Omega_f)}\right), \end{aligned}$$

which completes the proof. □

We now prove a global uniqueness of the Boussinesq equations (2.5)–(2.7) for the case of small data.

THEOREM 2.4. *Let* $\mathbf{u}$ *and* $F(\mathbf{u}) = T$ *be a solution of* (2.5)–(2.7)*, and suppose* $N||\nabla \mathbf{u}||_{\mathbf{L}^2(\Omega_f)}$ $+\alpha\, M < \nu$*, where*

$$N = \sup\left\{c_0(\mathbf{u},\mathbf{v},\mathbf{w}) : ||\nabla \mathbf{u}||_{\mathbf{L}^2(\Omega_f)} = ||\nabla \mathbf{v}||_{\mathbf{L}^2(\Omega_f)} = ||\nabla \mathbf{w}||_{\mathbf{L}^2(\Omega_f)} = 1, \mathbf{u},\mathbf{v},\mathbf{w} \in \mathbf{V}\right\}$$

and

$$M = \sup\left\{\frac{d(F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u}-\mathbf{v})}{||\nabla \mathbf{u} - \nabla \mathbf{v}||^2_{\mathbf{L}^2(\Omega_f)}} : \mathbf{u} \neq \mathbf{v}, \mathbf{u},\mathbf{v} \in \mathbf{V}\right\}.$$

Then $\mathbf{u}$ *and* $F(\mathbf{u}) = T$ *is the unique solution of* (2.5)–(2.7).

Proof. Suppose $(\mathbf{w}, F(\mathbf{w}))$ is a solution of (2.5)–(2.7), where $\mathbf{w} \neq \mathbf{u}$; then

$$\nu\, a_0(\mathbf{u},\mathbf{v}) + c_0(\mathbf{u},\mathbf{u},\mathbf{v}) = \alpha\, d(F(\mathbf{u}),\mathbf{v}) + \langle \mathbf{f},\mathbf{v}\rangle \quad \forall\, \mathbf{v} \in \mathbf{V}$$

and

$$\nu\, a_0(\mathbf{w},\mathbf{v}) + c_0(\mathbf{w},\mathbf{w},\mathbf{v}) = \alpha\, d(F(\mathbf{w}),\mathbf{v}) + \langle \mathbf{f},\mathbf{v}\rangle \quad \forall\, \mathbf{v} \in \mathbf{V}.$$

Subtracting with $\mathbf{v} = \mathbf{u} - \mathbf{w}$ and using the fact that $c_0(\mathbf{w}, \mathbf{u}-\mathbf{w}, \mathbf{u}-\mathbf{w}) = 0$, we have

$$\nu\, a_0(\mathbf{u}-\mathbf{w}, \mathbf{u}-\mathbf{w}) = -c_0(\mathbf{u}-\mathbf{w}, \mathbf{u}, \mathbf{u}-\mathbf{w}) + \alpha\, d\left(F(\mathbf{u}) - F(\mathbf{w}), \mathbf{u}-\mathbf{w}\right).$$

Hence,

$$\begin{aligned} \nu ||\nabla(\mathbf{u}-\mathbf{w})||^2_{\mathbf{L}^2(\Omega)} &\leq \left(N ||\nabla \mathbf{u}||_{\mathbf{L}^2(\Omega)} + \alpha M\right) ||\nabla(\mathbf{u}-\mathbf{w})||^2_{\mathbf{L}^2(\Omega)} \\ &< \nu ||\nabla(\mathbf{u}-\mathbf{w})||^2_{\mathbf{L}^2(\Omega)}, \end{aligned}$$

which is a contradiction. Therefore, $\mathbf{w} = \mathbf{u}$. □

2.2. Regularity of solutions of the Boussinesq equations. We now examine the regularity of solutions of the Boussinesq equations (2.5)–(2.7).

THEOREM 2.5. *Suppose that the given data satisfies* $Q \in L^2(\Omega)$, $\mathbf{f} \in \mathbf{L}^2(\Omega_f)$. *Then if* $(\mathbf{u}, p, T) \in \mathbf{H}^1_0(\Omega_f) \times L^2_0(\Omega_f) \times H^1_D(\Omega)$ *denotes a solution of the problem* (2.5)–(2.7)*, we have that* $(\mathbf{u}, p, T) \in \mathbf{H}^1_0(\Omega_f) \cap \mathbf{H}^2(\Omega_f) \times L^2_0(\Omega_f) \cap H^1(\Omega_f) \times H^1_D(\Omega) \cap H^s(\Omega)$ *for all* $s < \frac{3}{2}$*. Moreover, there exists a continuous function* P_s *for each* s *such that*

$$\begin{aligned} ||\mathbf{u}||_{\mathbf{H}^2(\Omega_f)} + ||p||_{H^1(\Omega_f)} + ||T||_{H^s(\Omega)} & \\ \leq P_s\left(||\mathbf{f}||_{\mathbf{L}^2(\Omega_f)} + ||Q||_{L^2(\Omega)} + ||g||_{L^2(\Gamma_C)}\right). & \end{aligned} \tag{2.22}$$

Proof. First we prove for the temperature T the regularity result and estimate. Since in Theorem 2.3 we already obtained an a priori estimate for $(\mathbf{u}, T)$ in $\mathbf{H}_0^1(\Omega_f) \times H^1(\Omega)$, by a localization argument it suffices to prove an estimate for temperature only in a vicinity of Σ. By our assumption, in the neighborhood of junction points T equals zero on the boundary. Thus locally one can consider our transmission problem as one with zero Dirichlet boundary conditions. Thus by Theorem 5 in [24], we have

$$(2.23) \qquad ||T||_{H^s(\Omega)} \le C(s) \left(||Q||_{L^2(\Omega)} + ||g||_{L^2(\Gamma_C)} + ||\mathbf{f}||_{\mathbf{L}^2(\Omega_f)} + 1\right)^2$$

for all $s \in (0, 3/2)$.

We also note that by assumption the boundary $\partial\Omega_f$ consists of two smooth arcs which intersect two times with anglers $\theta_i \in (0, \pi)$. Then since $T_f \in H^1(\Omega_f)$ and $\mathbf{f} \in \mathbf{L}^2(\Omega_f)$, the regularity of $\mathbf{u}$ and p follows from well-known theories concerning the Navier–Stokes equations in polygons (see Theorem 7.3.3.4 in [11]). By (2.23) and a priori estimates for the Stokes system in a polygon (see [11]) there exists a continuous function P_s for each s such that

$$(2.24) \qquad ||\mathbf{u}||_{\mathbf{H}^2(\Omega_f)} \le P_s \left(||\mathbf{f}||_{\mathbf{L}^2(\Omega_f)} + ||Q||_{L^2(\Omega)} + ||g||_{L^2(\Gamma_C)}\right).$$

Thus, the proof is completed. □

REMARK 2.1. *The regularity of temperature T achieved in the previous theorem is optimal in the sense of scale of Sobolev spaces $H^s(\Omega)$. In fact, even if $\kappa_s = \kappa_f$, in general $T \notin H^{\frac{3}{2}}(\Omega)$ (see [23]).*

3. The optimization problem and the existence of optimal solutions.

3.1. The optimization problems. We state the optimal control problem. We look for a $(\mathbf{u}, T, p, g) \in \mathbf{H}_0^1(\Omega_f) \times H_D^1(\Omega) \times L_0^2(\Omega_f) \times \mathcal{V}$ such that the cost functional

$$(3.1) \qquad \textit{(Problem 1)} \quad \mathcal{J}_1(\mathbf{u}, T, p, g) = \frac{1}{2} \int_\Omega |T - T_d|^2 \, d\mathbf{x} + \frac{\delta}{2} \int_{\Gamma_C} |g|^2 \, ds$$

or

$$(3.2) \qquad \textit{(Problem 2)} \quad \mathcal{J}_2(\mathbf{u}, T, p, g) = \frac{1}{2} \int_\Omega |\nabla \times \mathbf{u}|^2 \, d\mathbf{x} + \frac{\delta}{2} \int_{\Gamma_C} |g|^2 \, ds,$$

subject to the constraints

$$(3.3) \qquad \nu \, a_0(\mathbf{u}, \mathbf{v}) + c_0(\mathbf{u}, \mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = \alpha \, d(T, \mathbf{v}) + \langle \mathbf{f}, \mathbf{v} \rangle \ \ \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega_f),$$

$$(3.4) \qquad b(\mathbf{u}, q) = 0 \quad \forall\, q \in L_0^2(\Omega_f),$$

and

$$(3.5) \qquad a_1(T, S) + c_1(\mathbf{u}, T, S) = \langle Q, S \rangle - (\kappa g, S)_{\Gamma_C} \quad \forall\, S \in H_D^1(\Omega),$$

where $\mathcal{V}$ is a nonempty, closed, and convex subset of $L^2(\Gamma_C)$.

The *admissibility set* $\mathcal{U}_{ad}$ is defined by

$$(3.6) \qquad \begin{aligned} \mathcal{U}_{ad} = \big\{(\mathbf{u}, T, p, g) \in \mathbf{H}_0^1(\Omega_f) \times H^1(\Omega) \times L_0^2(\Omega_f) \times \mathcal{V} :& \\ \mathcal{J}(\mathbf{u}, T, p, g) < \infty, \quad \text{and (3.3)–(3.5) are satisfied}&\big\}, \end{aligned}$$

where $\mathcal{J}(\mathbf{u},T,p,g)$ is $\mathcal{J}_1(\mathbf{u},T,p,g)$ or $\mathcal{J}_2(\mathbf{u},T,p,g)$ depending on minimization problems. Then $(\mathbf{u},T,p,g)\in\mathcal{U}_{ad}$ is called an optimal solution if there exists $\varepsilon>0$ such that

$$\mathcal{J}(\mathbf{u},T,p,g)\le\mathcal{J}(\mathbf{v},S,q,h)\quad\forall\,(\mathbf{v},S,q,h)\in\mathcal{U}_{ad},\tag{3.7}$$

satisfying

$$\|\mathbf{u}-\mathbf{v}\|_{\mathbf{H}^1(\Omega_f)}+\|T-S\|_{H^1(\Omega)}+\|p-q\|_{L^2(\Omega_f)}+\|g-h\|_{L^2(\Gamma_C)}<\varepsilon\,.\tag{3.8}$$

If for an optimal solution $(\mathbf{u},T,p,g)\in\mathcal{U}_{ad}$ the inequalities (3.7) and (3.8) hold true with $\varepsilon=+\infty$, we say that $(\mathbf{u},T,p,g)$ is the *global minimum.* The optimal control problem can now be formulated as a constrained minimization in a Hilbert space:

$$\min_{(\mathbf{v},S,q,h)\in\mathcal{U}_{ad}}\mathcal{J}(\mathbf{v},S,q,h).\tag{3.9}$$

Problems 1 and 2 can be analyzed in exactly the same manner. In this section, we treat the first problem in detail.

3.2. The existence of an optimal solution. The existence of an optimal solution can be proved based on the a priori estimates (2.13) and standard techniques.

THEOREM 3.1. *Let* $Q\in L^2(\Omega)$ *and* $\mathbf{f}\in\mathbf{L}^2(\Omega_f)$. *Then there is an optimal solution* $(\hat{\mathbf{u}},\hat{T},\hat{p},\,\hat{g})\in\mathcal{U}_{ad}$ *to* (3.9).

Proof. The set $\mathcal{U}_{ad}$ is apparently nonempty because of Lemma 2.2. Thus we may choose a minimizing sequence $\{\mathbf{u}^{(n)},T^{(n)},p^{(n)},g^{(n)}\}$ in $\mathcal{U}_{ad}$ such that

$$\lim_{n\to\infty}\mathcal{J}_1(\mathbf{u}^{(n)},T^{(n)},p^{(n)},g^{(n)})=\inf_{(\mathbf{v},S,q,z)\in\mathcal{U}_{ad}}\mathcal{J}_1(\mathbf{v},S,q,z).\tag{3.10}$$

By the definition of $\mathcal{U}_{ad}$, we have

$$\begin{aligned}\nu\,a_0(\mathbf{u}^{(n)},\mathbf{v})+c_0(\mathbf{u}^{(n)},\mathbf{u}^{(n)},\mathbf{v})+b(\mathbf{v},p^{(n)})&\\ =\alpha\,d(T^{(n)},\mathbf{v})+\langle\mathbf{f},\mathbf{v}\rangle&\quad\forall\,\mathbf{v}\in\mathbf{H}_0^1(\Omega_f),\end{aligned}\tag{3.11}$$

$$b(\mathbf{u}^{(n)},q)=0\quad\forall\,q\in L_0^2(\Omega_f),\tag{3.12}$$

and

$$a_1(T^{(n)},S)+c_1(\mathbf{u}^{(n)},T^{(n)},S)=\langle Q,S\rangle+\kappa_f(g^{(n)},S)\quad\forall\,S\in H_D^1(\Omega).\tag{3.13}$$

From (1.6) and (3.6), we easily see that $\{||g^{(n)}||_{L^2(\Gamma_C)}\}$ is uniformly bounded. Also, by (2.13) we have that the sequences $\{||\mathbf{u}^{(n)}||_{\mathbf{H}^1(\Omega_f)}\}$, $\{||T^{(n)}||_{H^1(\Omega)}\}$ and $\{||p^{(n)}||_{L^2(\Omega_f)}\}$ are uniformly bounded. We may then extract subsequences such that

$$\begin{aligned}g^{(n)}\rightharpoonup\hat{g}&\quad\text{in }L^2(\Gamma_C),\\ \mathbf{u}^{(n)}\rightharpoonup\hat{\mathbf{u}}\quad\text{in }\mathbf{H}_0^1(\Omega_f)&\quad\text{and}\quad\nabla\mathbf{u}^{(n)}\rightharpoonup\nabla\hat{\mathbf{u}}\quad\text{in }\mathbf{L}^2(\Omega_f),\\ T^{(n)}\rightharpoonup\hat{T}\quad\text{in }H_D^1(\Omega)&\quad\text{and}\quad\nabla T^{(n)}\rightharpoonup\nabla\hat{T}\quad\text{in }\mathbf{L}^2(\Omega),\\ p^{(n)}\rightharpoonup\hat{p}&\quad\text{in }L^2(\Omega_f),\\ \mathbf{u}^{(n)}\to\hat{\mathbf{u}}&\quad\text{in }\mathbf{L}^4(\Omega_f)\ \text{ and }\ \mathbf{L}^2(\Omega_f)\end{aligned}$$

for some $(\hat{\mathbf{u}},\hat{T},\hat{p},\hat{g})\in\mathbf{H}_0^1(\Omega_f)\times H_D^1(\Omega)\times L_0^2(\Omega_f)\times L^2(\Gamma_C)$. The last convergence result above follows from the compact embedding $\mathbf{H}^1(\Omega_f)\hookrightarrow\hookrightarrow\mathbf{L}^4(\Omega_f)$ and $\mathbf{H}^1(\Omega_f)\hookrightarrow\hookrightarrow$

$\mathbf{L}^2(\Omega_f)$. We may pass to the limit in (3.11)–(3.13) to determine that $(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$ satisfies (3.3)–(3.5). Indeed, the only troublesome term when one passes to the limit is the nonlinearity $c_0(\cdot, \cdot, \cdot)$. However, note that

$$c_0(\mathbf{u}^{(n)}, \mathbf{u}^{(n)}, \mathbf{v}) \\ = \int_{\partial\Omega_f} (\mathbf{u}^{(n)} \cdot \mathbf{n})\mathbf{u}^{(n)} \cdot \mathbf{v}\, ds - \int_{\Omega_f} (\mathbf{u}^{(n)} \cdot \nabla)\mathbf{v} \cdot \mathbf{u}^{(n)}\, d\mathbf{x} \quad \forall\, \mathbf{v} \in \mathcal{D}(\bar{\Omega}_f),$$

where $\mathcal{D}(\bar{\Omega}_f)$ is the space of test functions. Then, since $\mathbf{u}^{(n)} \to \hat{\mathbf{u}}$ in $\mathbf{L}^2(\Omega_f)$ and $\int_{\partial\Omega_f} (\mathbf{u}^{(n)} \cdot \mathbf{n})\mathbf{u}^{(n)} \cdot \mathbf{v}\, ds = 0$ for all n, we have that

$$\lim_{k\to\infty} c_0(\mathbf{u}^{(k)}, \mathbf{u}^{(k)}, \mathbf{v}) = -\int_{\Omega_f} (\hat{\mathbf{u}} \cdot \nabla)\, \mathbf{v} \cdot \hat{\mathbf{u}}\, d\mathbf{x} \qquad \forall\, \mathbf{v} \in \mathcal{D}(\bar{\Omega}_f).$$

Since $\mathcal{D}(\bar{\Omega}_f)$ is dense in $\mathbf{H}^1_0(\Omega_f)$, we have that for each $\hat{\mathbf{u}} \in \mathbf{H}^1_0(\Omega_f)$,

$$\lim_{k\to\infty} c_0(\mathbf{u}^{(k)}, \mathbf{u}^{(k)}, \mathbf{v}) = c_0(\hat{\mathbf{u}}, \hat{\mathbf{u}}, \mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega_f).$$

Thus we have shown that $(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$ indeed satisfies (3.3)–(3.5) so that $(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}) \in \mathcal{U}_{ad}$.

Finally, it is easy to see that $\mathcal{J}_1(\cdot, \cdot, \cdot, \cdot)$ is weakly lower-semicontinuous so that

$$\mathcal{J}_1(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}) = \inf_{(\mathbf{v}, S, q, z) \in \mathcal{U}_{ad}} \mathcal{J}_1(\mathbf{v}, S, q, z). \tag{3.14}$$

Thus an optimal solution belonging to $\mathcal{U}_{ad}$ exists. □

4. The existence of Lagrange multipliers and an optimality system. This section is devoted to obtaining an optimality system to (3.9). We wish to use the method of Lagrange multipliers to turn the constrained optimization problem (3.9) into an unconstrained one. We establish also that there exists an open and dense set of initial data such that the Lagrange multiplier with respect to functional (1.6) is not equal to zero. We first show that suitable Lagrange multipliers exist.

THEOREM 4.1. *Let $Q \in L^2(\Omega)$ and $\mathbf{f} \in \mathbf{L}^2(\Omega)$. Assume $(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}) \in \mathcal{U}_{ad}$ is an optimal solution to minimization problem* (3.9). *Then there exist Lagrange multipliers $(\lambda, \mathbf{w}, R, q) \in \mathbb{R}^1 \times \mathbf{V} \cap \mathbf{H}^2(\Omega_f) \times H^s(\Omega) \times H^1(\Omega_f)$ for all $s < \frac{3}{2}$ such that*

$$(\lambda, \mathbf{w}, R, q) \neq 0\,, \tag{4.1}$$

$$-\nabla \cdot (\kappa \nabla R) - \chi_{\Omega_f}(\hat{\mathbf{u}} \cdot \nabla)R - \alpha\, \chi_{\Omega_f}(\mathbf{w}, \mathbf{e}_2) + \lambda(\hat{T} - T_d) = 0 \ \textit{in}\ \Omega, \tag{4.2}$$

$$\frac{\partial R}{\partial \mathbf{n}}|_{\Gamma_C} = 0, \quad (\lambda\delta\hat{g} - \kappa R, g - \hat{g})_{L^2(\Gamma_C)} \geq 0 \quad \forall\, g \in \mathcal{V}, \quad R|_{\Gamma_D} = 0, \tag{4.3}$$

$$-\nu\Delta\mathbf{w} - (\hat{\mathbf{u}} \cdot \nabla)\mathbf{w} + B(\hat{\mathbf{u}}, \mathbf{w}) + R\nabla\hat{T} = \nabla q \ \ \textit{in}\ \Omega_f, \tag{4.4}$$

and

$$\nabla \cdot \mathbf{w} = 0, \ \ \mathbf{w}|_{\partial\Omega_f} = 0, \tag{4.5}$$

where $B(\hat{\mathbf{u}}, \mathbf{w}) = \big((\mathbf{w}, \frac{\partial\hat{\mathbf{u}}}{\partial x_1}), (\mathbf{w}, \frac{\partial\hat{\mathbf{u}}}{\partial x_2})\big)$. Moreover, if $\lambda = 0$ and $\mathcal{V} = L^2(\Gamma_C)$, then $R \neq 0$.

Proof. To prove the existence of Lagrange multipliers for the constrained minimization problem (3.9), we use the penalty method. Let us consider the auxiliary extremal problem : find $(\mathbf{u}, T, p, g) \in \mathbf{V} \times H^1(\Omega) \times L_0^2(\Omega_f) \times L^2(\Gamma_C)$ which minimizes the functional

$$\begin{aligned} J_\varepsilon(\mathbf{u}, T, p, g) &= \mathcal{J}_1(\mathbf{u}, T, p, g) \\ &+ \frac{1}{2\varepsilon}\| - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \alpha T\mathbf{e}_2 + \nabla p - \mathbf{f}\|^2_{\mathbf{L}^2(\Omega_f)} \\ &+ \frac{1}{2\varepsilon}\| - \nabla\cdot(\kappa\nabla T) + \chi_{\Omega_f}(\mathbf{u}\cdot\nabla)T - Q\|^2_{L^2(\Omega)} \\ &+ \frac{N}{2}\|\mathbf{u} - \hat{\mathbf{u}}\|^2_{\mathbf{L}^2(\Omega_f)} + \frac{N}{2}\|T - \hat{T}\|^2_{L^2(\Omega)} + \frac{N}{2}\|g - \hat{g}\|^2_{L^2(\Gamma_C)} \end{aligned} \tag{4.6}$$

with

$$T|_{\Gamma_D} = 0,\ \frac{\partial T}{\partial \mathbf{n}}|_{\Gamma_C} = g,\ g \in \mathcal{V},\ \mathbf{u}|_{\partial\Omega_f} = 0,\ \nabla\cdot\mathbf{u} = 0, \tag{4.7}$$

where $(\hat{\mathbf{u}}, \hat{p}, \hat{T}, \hat{g}) \in \mathbf{V} \cap \mathbf{H}^2(\Omega_f) \times H^1(\Omega_f) \cap L_0^2(\Omega_f) \times H_D^1(\Omega) \cap H^s(\Omega) \times L^2(\Gamma_C)$ for all $s < \frac{3}{2}$ is a solution to extremal problem (3.9), such that inequality (3.8) holds true with $\varepsilon = \hat{\varepsilon}$ and $N > 0, \varepsilon \in (0, 1)$ as parameters. The existence of this solution $(\hat{\mathbf{u}}, \hat{p}, \hat{T}, \hat{g})$ was established in Theorem 3.1. By a method similar to the one used in the proof of Theorem 3.1, one can prove that there exists a solution of the problem (4.6)–(4.7) $(\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon) \in \mathbf{V} \cap \mathbf{H}^2(\Omega_f) \times H_D^1(\Omega) \cap H^s(\Omega) \times H^1(\Omega_f) \cap L_0^2(\Omega_f) \times L^2(\Gamma_C)$ for all $s \in [1, \frac{3}{2})$. Moreover, from the fact that $J_\varepsilon(\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon) \leq J_\varepsilon(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}) = \mathcal{J}_1(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$ and the inequality (2.22), we have that

$$\begin{aligned} &\left\{(\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon)\right\}_{\varepsilon\in(0,1)} \quad \text{is bounded in} \\ &\qquad\qquad \mathbf{V} \cap \mathbf{H}^2(\Omega_f) \times H_D^1(\Omega) \cap H^s(\Omega) \times L_0^2(\Omega_f) \times L^2(\Gamma_C) \end{aligned} \tag{4.8}$$

for all $s \in [1, \frac{3}{2})$. Thus, from (4.6)–(4.8) for any $\hat{\varepsilon} > 0$, taking parameter N sufficiently large, we obtain

$$\|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|_{\mathbf{L}^2(\Omega_f)} + \|\hat{T}_\varepsilon - \hat{T}\|_{L^2(\Omega)} + \|\hat{g}_\varepsilon - \hat{g}\|_{L^2(\Gamma_C)} \leq \hat{\varepsilon}/2\,. \tag{4.9}$$

Denoting

$$\begin{aligned} \hat{f}_\varepsilon &= -\nu\Delta\hat{\mathbf{u}}_\varepsilon + (\hat{\mathbf{u}}_\varepsilon\cdot\nabla)\hat{\mathbf{u}}_\varepsilon - \alpha\hat{T}_\varepsilon\mathbf{e}_2 + \nabla\hat{p}_\varepsilon - \mathbf{f}, \\ \hat{Q}_\varepsilon &= -\nabla\cdot(\kappa\nabla\hat{T}_\varepsilon) + \chi_{\Omega_f}(\hat{\mathbf{u}}_\varepsilon\cdot\nabla)\hat{T}_\varepsilon - Q, \end{aligned}$$

we obviously have

$$(\hat{f}_\varepsilon, \hat{Q}_\varepsilon) \to (0, 0) \quad \text{in} \quad \mathbf{L}^2(\Omega_f) \times L^2(\Omega). \tag{4.10}$$

By fact (4.8) and the Sobolev imbedding theorem and interpolation theorem, we have

$$\begin{aligned} \|(\hat{\mathbf{u}}_\varepsilon\cdot\nabla)\hat{\mathbf{u}}_\varepsilon &- (\hat{\mathbf{u}}\cdot\nabla)\hat{\mathbf{u}}\|_{\mathbf{L}^2(\Omega_f)} \\ &\leq \|((\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}})\cdot\nabla)\hat{\mathbf{u}}_\varepsilon + (\hat{\mathbf{u}}\cdot\nabla)(\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}})\|_{\mathbf{L}^2(\Omega_f)} \\ &\leq C\|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|_{\mathbf{H}^{\frac{3}{4}}(\Omega_f)}(\|\hat{\mathbf{u}}_\varepsilon\|_{\mathbf{H}^2(\Omega_f)} + \|\hat{\mathbf{u}}\|_{\mathbf{H}^2(\Omega_f)}) \\ &\leq C\|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|^{\frac{1}{4}}_{\mathbf{V}}\|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|^{\frac{3}{4}}_{\mathbf{H}^2(\Omega_f)}(\|\hat{\mathbf{u}}_\varepsilon\|_{\mathbf{H}^2(\Omega_f)} + \|\hat{\mathbf{u}}\|_{\mathbf{H}^2(\Omega_f)}) \\ &\leq C\|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|^{\frac{1}{4}}_{\mathbf{V}}, \end{aligned} \tag{4.11}$$

where C is independent of $\epsilon \in (0,1)$. Note that

$$\begin{aligned}(4.12)\qquad & -\nu\Delta(\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}) + (\hat{\mathbf{u}}_\varepsilon \cdot \nabla)\hat{\mathbf{u}}_\varepsilon - (\hat{\mathbf{u}} \cdot \nabla)\hat{\mathbf{u}} \\ & \qquad - \alpha(\hat{T}_\varepsilon - \hat{T})\mathbf{e}_2 + \nabla(\hat{p}_\varepsilon - \hat{p}) = \hat{f}_\varepsilon \quad \text{in} \quad \Omega_f\end{aligned}$$

and

$$(4.13)\qquad (\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}})|_{\partial\Omega_f} = 0, \quad \nabla \cdot (\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}) = 0.$$

By (4.11) and a priori estimates for the Stokes problem (see [25]), we have

$$\begin{aligned}(4.14)\qquad & \|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|_{\mathbf{H}^2(\Omega_f)} + \|\hat{p}_\varepsilon - \hat{p}\|_{H^1(\Omega_f)} \\ & \qquad \le C(\|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|_{\mathbf{V}}^{\frac{1}{4}} + \|\hat{T}_\varepsilon - \hat{T}\|_{L^2(\Omega)} + \|\hat{f}_\varepsilon\|_{\mathbf{L}^2(\Omega)}).\end{aligned}$$

The inequalities (4.9), (4.10), and (4.14) imply that for any $\hat{\varepsilon} > 0$ there exists a $N(\hat{\varepsilon}) > 0$ such that

$$\begin{aligned}(4.15)\qquad & \|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|_{\mathbf{H}^1(\Omega_f)} + \|\hat{T}_\varepsilon - \hat{T}\|_{H^1(\Omega)} + \|\hat{p}_\varepsilon - \hat{p}\|_{L^2(\Omega_f)} + \|\hat{g}_\varepsilon - \hat{g}\|_{L^2(\Gamma_C)} \le \hat{\varepsilon}, \\ & \qquad \forall \varepsilon \in (0,1).\end{aligned}$$

Therefore, without loss of generality, taking if necessary a subsequence, one can prove that

$$(\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon) \rightharpoonup (\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g}) \quad \text{in } \mathbf{V} \cap \mathbf{H}^2(\Omega_f) \times H^1_D(\Omega) \times H^1(\Omega_f) \cap L^2_0(\Omega_f) \times L^2(\Gamma_C).$$

In the same way, as was done in the proof of Theorem 3.1, one can show that $(\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g}) \in \mathcal{U}_{ad}$. Moreover, inequality (4.15) and weak lower-semicontinuity of norms in Hilbert spaces imply

$$(4.16)\quad \|\tilde{\mathbf{u}} - \hat{\mathbf{u}}\|_{\mathbf{H}^1(\Omega_f)} + \|\tilde{T} - \hat{T}\|_{H^1(\Omega)} + \|\tilde{p} - \hat{p}\|_{L^2(\Omega_f)} + \|\tilde{g} - \hat{g}\|_{L^2(\Gamma_C)} \le \hat{\varepsilon}.$$

On the other hand, the inequality

$$\mathcal{J}_\varepsilon(\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon) \le \mathcal{J}_1(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$$

yields

$$(4.17)\qquad \mathcal{J}_1(\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon) \le \mathcal{J}_1(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}).$$

Since the functional $\mathcal{J}_1$ is weak lower-semicontinuous, we have

$$(4.18)\qquad \mathcal{J}_1(\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g}) \le \mathcal{J}_1(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}).$$

By (4.16) and (4.18), we have that $(\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g})$ is a solution of the optimal control problem (3.9).

Now if we assume that $(\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g}) \neq (\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$, then

$$\mathcal{J}_1(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}) - \mathcal{J}_1(\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g}) \ge \frac{1}{2}\|\tilde{\mathbf{u}} - \hat{\mathbf{u}}\|^2_{\mathbf{L}^2(\Omega_f)} + \frac{1}{2}\|\tilde{T} - \hat{T}\|^2_{L^2(\Omega)} > 0,$$

which contradicts the fact that $(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$ is the solution to problem (3.9). Thus, $(\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g}) = (\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$ and we have

$$(4.19)\quad (\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon) \rightharpoonup (\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}) \text{ in } \mathbf{H^2}(\Omega_f) \times H^1_D(\Omega) \times L^2(\Omega_f) \times L^2(\Gamma_C).$$

Moreover, by (4.18) and (4.19) we have

$$\lim_{\varepsilon\to 0}\mathcal{J}_1(\hat{\mathbf{u}}_\varepsilon,\hat{T}_\varepsilon,\hat{p}_\varepsilon,\hat{g}_\varepsilon)=\mathcal{J}_1(\hat{\mathbf{u}},\hat{T},\hat{p},\hat{g}). \tag{4.20}$$

Hence by (4.19) and (4.20) we have

$$\hat{g}_\varepsilon\to\hat{g}\ \text{ in } L^2(\Gamma_C). \tag{4.21}$$

On the other hand, the facts (4.10), (4.14), (4.19), and (4.21) imply

$$(\hat{\mathbf{u}}_\varepsilon,\hat{T}_\varepsilon,\hat{p}_\varepsilon,\hat{g}_\varepsilon)\to(\hat{\mathbf{u}},\hat{T},\hat{p},\hat{g})\text{ in }\mathbf{H^2}(\Omega_f)\times H^1_D(\Omega)\times L^2(\Omega_f)\times L^2(\Gamma_C) \tag{4.22}$$

and

$$(\hat{\mathbf{u}}_\varepsilon,\hat{p}_\varepsilon,\hat{T}_\varepsilon)\to(\hat{\mathbf{u}},\hat{p},\hat{T})\text{ in }\mathbf{V}\times H^1(\Omega_f)\times H^s(\Omega)\quad\forall s\in\left(1,\frac{3}{2}\right). \tag{4.23}$$

Let $\mathbf{v}\in\mathbf{H}^2(\Omega)\cap\mathbf{V}$, $p\in C^2(\overline{\Omega})$, and $S_1\in H^2(\Omega)$ be arbitrary functions such that $S_1=0$ on Γ_D and $\frac{\partial S_1}{\partial\mathbf{n}}=0$ on Γ_C, and let $S_2\in H^s(\Omega)$ $(s<3/2)$ be the solution of the boundary value problem

$$\begin{gathered}-\nabla\cdot(\kappa\nabla S_2)=0\quad\text{on }\Omega,\\ S_2=0\quad\text{on }\Gamma_D,\quad\frac{\partial S_2}{\partial\mathbf{n}}=g\quad\text{on }\Gamma_C,\end{gathered}$$

where g is an arbitrary element in $\mathcal{V}$.

Now we introduce a function P defined by

$$P(\lambda_1,\lambda_2,\lambda_3)=\mathcal{J}_\epsilon(\hat{\mathbf{u}}_\epsilon+\lambda_1\mathbf{v},\hat{T}_\epsilon+\lambda_2S_1+\lambda_3(S_2-\hat{T}_\epsilon),\hat{p}_\epsilon,\hat{g}_\epsilon+\lambda_3(g-\hat{g}_\epsilon)).$$

Clearly, the function $P\in C^2(\mathbb{R}^3)$ and attains its minimum at $(0,0,0)$ on the set $\{(\lambda_1,\lambda_2,\lambda_3)\in\mathbb{R}^3\mid\lambda_3\in[0,1]\}$. Thus, we have

$$\frac{\partial P}{\partial\lambda_1}(0,0,0)=0,\quad\frac{\partial P}{\partial\lambda_2}(0,0,0)=0,\quad\frac{\partial P}{\partial\lambda_3}(0,0,0)\ge 0. \tag{4.24}$$

From the equations and inequality in (4.24), we obtain the optimality system

$$R_\varepsilon=\frac{1}{\varepsilon}(-\nabla\cdot(\kappa\nabla\hat{T}_\varepsilon)+\chi_{\Omega_f}(\hat{\mathbf{u}}_\varepsilon\cdot\nabla)\hat{T}_\varepsilon-Q), \tag{4.25}$$

$$\mathbf{w}_\varepsilon=\frac{1}{\varepsilon}(-\nu\Delta\hat{\mathbf{u}}_\varepsilon+(\hat{\mathbf{u}}_\varepsilon\cdot\nabla)\hat{\mathbf{u}}_\varepsilon+\nabla\hat{p}_\varepsilon-\alpha\hat{T}_\varepsilon\mathbf{e}_2-\mathbf{f}), \tag{4.26}$$

$$\begin{aligned}-\nabla\cdot(\kappa\nabla R_\varepsilon)-\chi_{\Omega_f}(\hat{\mathbf{u}}_\varepsilon\ \cdot\ \nabla)R_\varepsilon-\alpha\,\chi_{\Omega_f}(\mathbf{w}_\varepsilon,\mathbf{e}_2)&\\ +(\hat{T}_\varepsilon-T_d)+N(\hat{T}_\varepsilon-\hat{T})=0\quad\text{in }\Omega,&\end{aligned} \tag{4.27}$$

$$\begin{gathered}\frac{\partial R_\varepsilon}{\partial\mathbf{n}}|_{\Gamma_C}=0,\quad R_\varepsilon|_{\Gamma_D}=0,\\ (\delta\hat{g}_\varepsilon+N(\hat{g}_\varepsilon-\hat{g})-\kappa R_\varepsilon,\quad g-\hat{g}_\varepsilon)_{L^2(\Gamma_C)}\ge 0\quad\forall\,g\in\mathcal{V},\end{gathered} \tag{4.28}$$

$$-\nu\,\Delta\mathbf{w}_\varepsilon-(\hat{\mathbf{u}}_\varepsilon\cdot\nabla)\mathbf{w}_\varepsilon+B(\hat{\mathbf{u}}_\varepsilon,\mathbf{w}_\varepsilon)+R_\varepsilon\nabla\hat{T}_\varepsilon+N(\hat{\mathbf{u}}_\varepsilon-\hat{\mathbf{u}})=\nabla q_\varepsilon\text{ in }\Omega_f, \tag{4.29}$$

and

$$\nabla\cdot\mathbf{w}_\varepsilon=0,\ \ \mathbf{w}_\varepsilon|_{\partial\Omega_f}=0, \tag{4.30}$$

where the first equality in (4.28) has a sense due to the estimate

$$\left\|\frac{\partial R_\varepsilon}{\partial\mathbf{n}}\right\|_{H^s(\partial\Omega)}\le C(s)\big(\|\nabla\cdot(\kappa\nabla R_\varepsilon)\|_{L^2(\Omega)}+\|R_\varepsilon\|_{H^1(\Omega)}\big)$$

whenever $R_\varepsilon\in H^1_D(\Omega),\ s<0$.

Setting $I_\varepsilon=(\|R_\varepsilon\|^2_{L^2(\Omega)}+\|\mathbf{w}_\varepsilon\|^2_{\mathbf{L}^2(\Omega_f)})^{\frac{1}{2}}$, we consider two cases.

(A) Let $\liminf_{\varepsilon\to+0}I_\varepsilon=+\infty$. Denote $\tilde R_\varepsilon=R_\varepsilon/I_\varepsilon,\tilde{\mathbf{w}}_\varepsilon=\mathbf{w}_\varepsilon/I_\varepsilon,\tilde q=q_\varepsilon/I_\varepsilon$. From (4.25)–(4.30) the triple $(\tilde{\mathbf{w}}_\varepsilon,\tilde R_\varepsilon,\tilde q_\varepsilon)$ satisfies the equations

$$\begin{aligned}-\nabla\cdot(\kappa\nabla\tilde R_\varepsilon)-\chi_{\Omega_f}(\hat{\mathbf{u}}_\varepsilon\cdot\nabla)\tilde R_\varepsilon-\alpha\,\chi_{\Omega_f}(\tilde{\mathbf{w}}_\varepsilon,\mathbf{e}_2)&\\ +\frac{\hat T_\varepsilon-T_d}{I_\varepsilon}+N\frac{\hat T_\varepsilon-\hat T}{I_\varepsilon}&=0\quad\text{in}\quad\Omega,\end{aligned}\tag{4.31}$$

$$\begin{gathered}\frac{\partial\tilde R_\varepsilon}{\partial\mathbf{n}}|_{\Gamma_C}=0,\quad \tilde R_\varepsilon|_{\Gamma_D}=0,\\ \left(\frac{\delta\hat g_\varepsilon+N(\hat g_\varepsilon-\hat g)}{I_\varepsilon}-\kappa\tilde R_\varepsilon,\quad g-\hat g_\varepsilon\right)_{L^2(\Gamma_C)}\ge 0\quad\forall g\in\mathcal{V},\end{gathered}\tag{4.32}$$

$$-\nu\,\Delta\tilde{\mathbf{w}}_\varepsilon-(\hat{\mathbf{u}}_\varepsilon\cdot\nabla)\tilde{\mathbf{w}}_\varepsilon+B(\hat{\mathbf{u}}_\varepsilon,\tilde{\mathbf{w}}_\varepsilon)+\tilde R_\varepsilon\nabla\hat T_\varepsilon+N\frac{\hat{\mathbf{u}}_\varepsilon-\hat{\mathbf{u}}}{I_\varepsilon}=\nabla\tilde q_\varepsilon\ \text{ in }\Omega_f, \tag{4.33}$$

and

$$\nabla\cdot\tilde{\mathbf{w}}_\varepsilon=0,\ \ \tilde{\mathbf{w}}_\varepsilon|_{\partial\Omega_f}=0. \tag{4.34}$$

By the definitions of $\tilde{\mathbf{w}}_\varepsilon$ and $\tilde R_\varepsilon$, we have $\|\tilde{\mathbf{w}}_\varepsilon\|_{\mathbf{L}^2(\Omega_f)}\le 1$ and $\|\tilde R_\varepsilon\|_{L^2(\Omega)}\le 1$. Thus, taking if necessary a subsequence, one can show that

$$(\tilde{\mathbf{w}}_\varepsilon,\tilde R_\varepsilon)\rightharpoonup(\tilde{\mathbf{w}},\tilde R)\quad\text{in }\mathbf{L}^2(\Omega_f)\times L^2(\Omega). \tag{4.35}$$

Taking the inner product of (4.31) with R_ε in $L^2(\Omega)$ and integrating the product by parts, we obtain

$$\begin{aligned}\int_\Omega\kappa|\nabla\tilde R_\varepsilon|^2d\mathbf{x}&\le\|\tilde R_\varepsilon\|^2_{L^2(\Omega)}\\ &+C(\|\hat T_\varepsilon-T_d\|^2_{L^2(\Omega_s)}/I^2_\varepsilon+\|\hat T_\varepsilon-\hat T\|^2_{L^2(\Omega)}/I^2_\varepsilon+\|\tilde{\mathbf{w}}_\varepsilon\|^2_{\mathbf{L}^2(\Omega_f)}).\end{aligned}\tag{4.36}$$

By the fact $\|\tilde R_\varepsilon\|_{L^2(\Omega)}\le 1$ and the inequalities (4.15) and (4.36), we can assume without loss of generality that

$$\|\tilde R_\varepsilon\|_{H^1(\Omega)}\le C, \tag{4.37}$$

where the constant C is independent of ε. Then again taking the inner product of (4.33) with $\tilde{\mathbf{w}}_\varepsilon$ in $\mathbf{L}^2(\Omega)$ and integrating the product by parts, we obtain

$$\begin{aligned}\nu \int_\Omega |\nabla \tilde{\mathbf{w}}_\varepsilon|^2 \, d\mathbf{x} &= -\int_\Omega (\tilde{R}_\varepsilon(\nabla \hat{T}_\varepsilon, \tilde{\mathbf{w}}_\varepsilon) + (\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}, \tilde{\mathbf{w}}_\varepsilon) + (B(\hat{\mathbf{u}}_\varepsilon, \tilde{\mathbf{w}}_\varepsilon), \tilde{\mathbf{w}}_\varepsilon)) \, d\mathbf{x} \\ &\le C\Bigg(\|\nabla \hat{T}_\varepsilon\|_{L^2(\Omega)} \|\tilde{R}_\varepsilon\|_{H^1(\Omega)} \|\tilde{\mathbf{w}}_\varepsilon\|_V + \|\hat{\mathbf{u}}_\varepsilon\|_V \|\tilde{\mathbf{w}}_\varepsilon\|^{\frac{1}{2}}_{\mathbf{L}^2(\Omega_f)} \|\tilde{\mathbf{w}}_\varepsilon\|^{\frac{3}{2}}_V \\ &\qquad + \frac{\|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|_{\mathbf{L}^2(\Omega_f)} \|\tilde{\mathbf{w}}_\varepsilon\|_{\mathbf{L}^2(\Omega_f)}}{I_\varepsilon} \Bigg). \end{aligned} \tag{4.38}$$

By (4.22), (4.35), and (4.37) the above estimate implies immediately

$$\|\tilde{\mathbf{w}}_\varepsilon\|_{\mathbf{V}} \le C, \tag{4.39}$$

where the constant C is independent of ε.

From the facts (4.35), (4.37), and (4.39), again taking if necessary a subsequence, we obtain

$$\begin{aligned} (\tilde{\mathbf{w}}_\varepsilon, \tilde{R}_\varepsilon) &\rightharpoonup (\tilde{\mathbf{w}}, \tilde{R}) \quad \text{in} \quad \mathbf{V} \times H^1(\Omega), \\ (\tilde{\mathbf{w}}_\varepsilon, \tilde{R}_\varepsilon) &\to (\tilde{\mathbf{w}}, \tilde{R}) \quad \text{in} \quad \mathbf{L}^2(\Omega_f) \times L^2(\Omega). \end{aligned} \tag{4.40}$$

Furthermore, by (4.8) and (4.40) the sequence

$$\left\{ -(\hat{\mathbf{u}}_\varepsilon \cdot \nabla)\tilde{\mathbf{w}}_\varepsilon + B(\hat{\mathbf{u}}_\varepsilon, \tilde{\mathbf{w}}_\varepsilon) + \tilde{R}_\varepsilon \nabla \hat{T}_\varepsilon + N \frac{\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}}{I_\varepsilon} \right\}_{\varepsilon \in (0,1)}$$

is bounded in $\mathbf{L}^2(\Omega_f)$, and thus we have

$$\tilde{\mathbf{w}}_\varepsilon \rightharpoonup \tilde{\mathbf{w}} \ \text{ in } \mathbf{H}^2(\Omega_f). \tag{4.41}$$

Once again, since (4.8), (4.40), and (4.41) imply the boundedness of the sequence

$$\left\{ -\chi_{\Omega_f} (\hat{\mathbf{u}} \cdot \nabla)\tilde{R}_\varepsilon - \alpha\, \chi_{\Omega_f} (\tilde{\mathbf{w}}_\varepsilon, \mathbf{e}_2) + \frac{\hat{T}_\varepsilon - T_d}{I_\varepsilon} + \frac{\hat{T}_\varepsilon - \hat{T}}{I_\varepsilon} \right\}_{\varepsilon \in (0,1)}$$

in the space $L^2(\Omega)$, by (4.22) and (4.40) we have (see [24])

$$\tilde{R}_\varepsilon \rightharpoonup \tilde{R} \text{ in } H^s(\Omega) \text{ as } \varepsilon \to 0 \quad \forall s \in \left(1, \frac{3}{2}\right). \tag{4.42}$$

Since $\|(\tilde{\mathbf{w}}_\varepsilon, \tilde{R}_\varepsilon)\|_{\mathbf{L}^2(\Omega_f) \times L^2(\Omega)} = 1$, it follows from (4.40) that

$$\|(\tilde{\mathbf{w}}, \tilde{R})\|_{\mathbf{L}^2(\Omega_f) \times L^2(\Omega)} = 1. \tag{4.43}$$

Thus passing to the limit in (4.31)–(4.34) as $\varepsilon \to +0$, keeping in mind (4.19), (4.41), and (4.42), we obtain the optimality system (4.1)–(4.5) with $\lambda = 0$:

$$-\nabla \cdot (\kappa \nabla \tilde{R}) - \chi_{\Omega_f} (\, \hat{\mathbf{u}} \cdot \nabla)\tilde{R} - \alpha\, \chi_{\Omega_f} (\tilde{\mathbf{w}}, \mathbf{e}_2) = 0 \text{ in } \Omega, \tag{4.44}$$

$$\frac{\partial \tilde{R}}{\partial \mathbf{n}}|_{\Gamma_C} = 0, \quad -(\kappa \tilde{R}, g - \hat{g})|_{L^2(\Gamma_C)} \ge 0 \quad \forall\, g \in \mathcal{V}, \quad \tilde{R}|_{\Gamma_D} = 0, \tag{4.45}$$

$$-\nu \Delta \tilde{\mathbf{w}} - (\hat{\mathbf{u}} \cdot \nabla)\tilde{\mathbf{w}} + B(\hat{\mathbf{u}}, \tilde{\mathbf{w}}) + \tilde{R} \nabla \hat{T} = \nabla \tilde{q} \ \text{ in } \Omega_f, \tag{4.46}$$

and

$$\nabla \cdot \tilde{\mathbf{w}} = 0, \ \ \tilde{\mathbf{w}}|_{\partial\Omega_f} = 0. \tag{4.47}$$

The necessary regularity of Lagrange multipliers follows from (4.22), (4.41), and (4.42).

Now we consider the case $\mathcal{V} = L^2(\Gamma_C)$. In that case the inequality

$$-(\kappa\tilde{R}, g - \hat{g})_{L^2(\Gamma_C)} \geq 0 \quad \forall\, g \in \mathcal{V}$$

implies $\tilde{R}|_{\Gamma_C} = 0$. Then by the uniqueness theorem for the Cauchy problem for the Laplace operator from (4.44) and (4.45) we have

$$\tilde{R} \equiv 0 \quad \text{in } \Omega_s.$$

Now let us show that $\tilde{R} \neq 0$ in Ω_f. If our statement is not true, then (4.44) yields

$$(\tilde{\mathbf{w}}(x), \mathbf{e}_2) = 0 \quad \text{in} \quad \Omega_f.$$

By (4.47) and the above equality there exists a vector $\mathbf{a}$ such that

$$\frac{\partial \tilde{\mathbf{w}}_i}{\partial \mathbf{a}} = 0 \quad \text{in } \Omega_f \ \forall\, i \in \{1,2\}. \tag{4.48}$$

Hence by (4.47) and (4.48) $\tilde{\mathbf{w}} \equiv 0$. But this contradicts (4.43).

(B) Let $\liminf_{\varepsilon\to 0} I_\varepsilon < +\infty$ or, put another way,

$$\|R_\varepsilon\|_{L^2(\Omega)} + \|\mathbf{w}_\varepsilon\|_{\mathbf{L}^2(\Omega_f)} \leq C. \tag{4.49}$$

Taking the inner product of (4.27) with R_ε in $L^2(\Omega)$ and integrating the product by parts, we obtain

$$\begin{aligned}\int_\Omega \kappa|\nabla R_\varepsilon|^2\, d\mathbf{x} \leq{}& \|R_\varepsilon\|^2_{L^2(\Omega)} \\ &+ C(\|\hat{T}_\varepsilon - T_d\|^2_{L^2(\Omega_s)} + \|\hat{T}_\varepsilon - \hat{T}\|^2_{L^2(\Omega)} + \|\mathbf{w}_\varepsilon\|^2_{\mathbf{L}^2(\Omega_f)}).\end{aligned} \tag{4.50}$$

By (4.49) and (4.50) we can assume without loss of generality that

$$R_\varepsilon \rightharpoonup R \quad \text{in } H^1(\Omega). \tag{4.51}$$

Then again taking the inner product of (4.29) with $\mathbf{w}_\varepsilon$ in $\mathbf{L}^2(\Omega)$ and integrating the product by parts, we obtain

$$\begin{aligned}\nu\int_\Omega |\nabla\mathbf{w}_\varepsilon|^2\, d\mathbf{x} &= -\int_\Omega (R_\varepsilon(\nabla\hat{T}_\varepsilon, \mathbf{w}_\varepsilon) + (\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}, \mathbf{w}_\varepsilon) + (B(\hat{\mathbf{u}}_\varepsilon, \mathbf{w}_\varepsilon), \mathbf{w}_\varepsilon))\, d\mathbf{x} \\ &\leq C(\|\nabla\hat{T}_\varepsilon\|_{L^2(\Omega)}\|R_\varepsilon\|_{H^1(\Omega)}\|\mathbf{w}_\varepsilon\|_V + \|\hat{\mathbf{u}}_\varepsilon\|_V\|\mathbf{w}_\varepsilon\|^{\frac{1}{2}}_{\mathbf{L}^2(\Omega_f)}\|\mathbf{w}_\varepsilon\|^{\frac{3}{2}}_V \\ &\qquad + \|\hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}}\|_{\mathbf{L}^2(\Omega_f)}\|\mathbf{w}_\varepsilon\|_{\mathbf{L}^2(\Omega_f)}).\end{aligned} \tag{4.52}$$

By (4.22), (4.49), and (4.51) the above estimate implies immediately

$$\mathbf{w}_\varepsilon \rightharpoonup \mathbf{w} \quad \text{in } \mathbf{V}. \tag{4.53}$$

Once again, using arguments similar to (4.35) and (4.40), we obtain

$$(\mathbf{w}_\varepsilon, R_\varepsilon) \rightharpoonup (\mathbf{w}, R) \quad \text{in} \quad \mathbf{H}^2(\Omega_f) \times H^s(\Omega) \quad \forall s \in \left(1, \frac{3}{2}\right). \tag{4.54}$$

By (4.22), (4.51), (4.53), and (4.54) passing to the limit in (4.27)–(4.30), we obtain optimality system (4.1)–(4.5) with $\lambda = 1$. The relations (4.22), (4.53), (4.54), and the equation (4.4) imply the necessary regularity of Lagrange multipliers. □

Remark 4.1. *If Lagrange multiplier $\lambda \neq 0$ in system (4.1)–(4.5), without loss of generality we can assume that $\lambda = 1$. Instead of $R, \mathbf{w}, q$ one can just introduce the new Lagrange multipliers $\tilde{R} = R/\lambda, \tilde{\mathbf{w}} = \mathbf{w}/\lambda, \tilde{q} = q/\lambda$. Obviously, $\tilde{\mathbf{w}}, \tilde{R}, \tilde{q}$ satisfy the system of equations*

$$(1, \tilde{\mathbf{w}}, \tilde{R}, \tilde{q}) \neq 0, \tag{4.55}$$

$$-\nabla \cdot (\kappa \nabla \tilde{R}) - \chi_{\Omega_f}(\hat{\mathbf{u}} \cdot \nabla)\tilde{R} - \alpha\, \chi_{\Omega_f}(\tilde{\mathbf{w}}, \mathbf{e}_2) + (\hat{T} - T_d) = 0 \ \text{in}\ \Omega, \tag{4.56}$$

$$\frac{\partial \tilde{R}}{\partial \mathbf{n}}|_{\Gamma_C} = 0, \quad (\delta \hat{g} - \kappa \tilde{R}, g - \hat{g})_{L^2(\Gamma_C)} \geq 0 \quad \forall\, g \in \mathcal{V}, \quad \tilde{R}|_{\Gamma_D} = 0, \tag{4.57}$$

$$-\nu \Delta \tilde{\mathbf{w}} - (\hat{\mathbf{u}} \cdot \nabla)\tilde{\mathbf{w}} + B(\hat{\mathbf{u}}, \tilde{\mathbf{w}}) + \tilde{R} \nabla \hat{T} = \nabla \tilde{q} \ \ \text{in} \quad \Omega_f, \tag{4.58}$$

and

$$\nabla \cdot \tilde{\mathbf{w}} = 0, \ \ \tilde{\mathbf{w}}|_{\partial \Omega_f} = 0. \tag{4.59}$$

Naturally, we are interested in the case when the optimality system has the Lagrange multiplier λ not equal to zero. Unfortunately, we could not prove that for an arbitrary $(\mathbf{f}, Q) \in \mathbf{L}^2(\Omega_f) \times L^2(\Omega)$ the optimality system (4.1)–(4.5) holds true with $\lambda \neq 0$. But below we will prove this fact for the dense set of $(\mathbf{f}, Q)$ in $\mathbf{L}^2(\Omega_f) \times L^2(\Omega)$. First, let us introduce the following definition.

Definition 4.1. *We say that pair $(\mathbf{f}, Q) \in \mathbf{L}^2(\Omega_f) \times L^2(\Omega)$ belongs to the set $\mathcal{O}$ if and only if for all global minimums to problem (3.9) (with right-hand sides in (1.1) and (1.3) equal to $\mathbf{f}$ and Q, respectively) the optimality system (4.1)–(4.5) holds true with $\lambda = 1$.*

Theorem 4.2. *The set $\mathcal{O}$ is dense in $\mathbf{L}^2(\Omega_f) \times L^2(\Omega)$.*

Proof. Let us consider the auxiliary extremal problem

$$\begin{aligned} \mathcal{I}_\varepsilon(\mathbf{u}, T, p, g) &= \mathcal{J}_1(\mathbf{u}, T, p, g) \\ &\quad + \frac{1}{2\varepsilon}\| - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} - \alpha T \mathbf{e}_2 + \nabla p - \mathbf{f}\|^2_{\mathbf{L}^2(\Omega_f)} \\ &\quad + \frac{1}{2\varepsilon}\| - \nabla \cdot (\kappa \nabla T) + \chi_{\Omega_f}(\mathbf{u} \cdot \nabla)T \ - Q\|^2_{L^2(\Omega)} \ \rightarrow \ \inf, \end{aligned} \tag{4.60}$$

and

$$T|_{\Gamma_D} = 0, \ \frac{\partial T}{\partial \mathbf{n}}|_{\Gamma_C} = g \in \mathcal{V}, \ \mathbf{u}|_{\partial \Omega_f} = 0, \ \nabla \cdot \mathbf{u} = 0. \tag{4.61}$$

By arguments similar to one in the proof of Theorem 4.1, one can show that there exists at least one solution $(\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon) \in \mathbf{V} \cap \mathbf{H}^2(\Omega_f) \times H^1(\Omega) \times L^2_0(\Omega_f) \times L^2(\Gamma_C)$ to problem (4.60) and (4.61). Let us denote by $(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$ a solution to problem (3.9). Since $\mathcal{I}_\varepsilon(\hat{\mathbf{u}}_\varepsilon, \hat{T}_\varepsilon, \hat{p}_\varepsilon, \hat{g}_\varepsilon) \leq \mathcal{I}_\varepsilon(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g}) = \mathcal{J}_1(\hat{\mathbf{u}}, \hat{T}, \hat{p}, \hat{g})$, we have

$$(\varepsilon R_\varepsilon, \varepsilon \mathbf{w}_\varepsilon) \rightarrow (0, 0) \quad \text{in} \quad L^2(\Omega) \times \mathbf{L}^2(\Omega_f), \tag{4.62}$$

where

$$R_\varepsilon = \frac{1}{\varepsilon}(-\nabla \cdot (\kappa \nabla \hat{T}_\varepsilon) + \chi_{\Omega_f}(\hat{\mathbf{u}}_\varepsilon \cdot \nabla)\hat{T}_\varepsilon - Q) \tag{4.63}$$

and

$$\mathbf{w}_\varepsilon = \frac{1}{\varepsilon}(-\nu\Delta\hat{\mathbf{u}}_\varepsilon + (\hat{\mathbf{u}}_\varepsilon \cdot \nabla)\hat{\mathbf{u}}_\varepsilon + \nabla\hat{p}_\varepsilon - \alpha\hat{T}_\varepsilon\mathbf{e}_2 - \mathbf{f}). \tag{4.64}$$

The optimality system to problem (4.60)–(4.61) is as follows:

$$-\nabla \cdot (\kappa\nabla R_\varepsilon) - \chi_{\Omega_f}(\hat{\mathbf{u}}_\varepsilon \cdot \nabla)R_\varepsilon - \alpha\chi_{\Omega_f}(\mathbf{w}_\varepsilon, \mathbf{e}_2) + (\hat{T}_\varepsilon - T_d) = 0 \text{ in } \Omega, \tag{4.65}$$

$$\frac{\partial R_\varepsilon}{\partial \mathbf{n}}|_{\Gamma_C} = 0, \ \ (\delta\hat{g}_\varepsilon - \kappa R_\varepsilon, g - \hat{g}_\varepsilon)_{L^2(\Gamma_C)} \geq 0 \quad \forall\, g \in \mathcal{V}, \ \ R_\varepsilon|_{\Gamma_D} = 0, \tag{4.66}$$

$$-\nu\,\Delta\mathbf{w}_\varepsilon - (\hat{\mathbf{u}}_\varepsilon \cdot \nabla)\mathbf{w}_\varepsilon + B(\hat{\mathbf{u}}_\varepsilon, \mathbf{w}_\varepsilon) + R_\varepsilon\nabla\hat{T}_\varepsilon + \hat{\mathbf{u}}_\varepsilon - \hat{\mathbf{u}} = \nabla q_\varepsilon \ \text{ in } \Omega_f, \tag{4.67}$$

and

$$\nabla \cdot \mathbf{w}_\varepsilon = 0, \ \ \mathbf{w}_\varepsilon|_{\partial\Omega_f} = 0. \tag{4.68}$$

Note that a global minimum to problem (4.60)–(4.61) is simultaneously the global minimum solution to problem (3.9); there, instead of $(Q, \mathbf{f})$, we have $(\varepsilon R_\varepsilon + Q, \varepsilon\mathbf{w}_\varepsilon + \mathbf{f})$. By (4.65)–(4.68) the Lagrange multiplier λ equals 1. On the other hand, let $(\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g})$ be another global minimum to problem (3.9); there, in the right-hand side of (2.5) and (2.7), instead of $(Q, \mathbf{f})$, we have $(\varepsilon R_\varepsilon + Q, \varepsilon\mathbf{w}_\varepsilon + \mathbf{f})$. Obviously, $(\tilde{\mathbf{u}}, \tilde{T}, \tilde{p}, \tilde{g})$ are also solutions to problem (4.60)–(4.61). Hence the pair $(\varepsilon R_\varepsilon + Q, \varepsilon\mathbf{w}_\varepsilon + \mathbf{f}) \in \mathcal{O}$. Then the statement of the our theorem follows from (4.62). □

Unfortunately, in the general case we cannot prove that the set $\mathcal{O}$ constructed in Theorem 4.2 is open. Below we consider the special case. Let us assume that

$$\mathcal{V} = X \subset L^2(\Gamma_C), \ \text{ where } X \text{ is a linear space, } \dim X < \infty.$$

We start from the following definitions.

Definition 4.2. *Let E and E_0 be Banach spaces and $\Omega \subset E$. The mapping $A : E \to E_0$ is called proper on Ω if the preimage $A^{-1}K \cap \Omega$ of a compact $K \subset E_0$ is compact in Ω for any choice of a compact $K \subset E_0$.*

Definition 4.3. *A linear operator $L : E \to E_0$ is called a Fredholm operator with index k, $k \in \mathbf{Z}$, if the image subspace $L(E) \subset E_0$ is closed and has a finite codimension* $\mathrm{codim} L(E) = \dim(E_0/L(E))$, *and the kernel $Ker\, L \subset E$ has a finite dimension* $\dim Ker\, L$. *The number k is called the index of L : $k = \dim Ker\, L - \mathrm{codim} L(E)$.*

Definition 4.4. *Let A be an operator defined in an open domain $O \subset E$ of a Banach space $E, A : O \to E_0$, where E_0 is another Banach space. It is supposed that A is continuously differentiable on O and $A'(u)$ is a Fredholm operator with the index k for any $u \in O$. Such a mapping is called Fredholm with the index k on O.*

Definition 4.5. *Let $B \subset O$. A point $y \in E_0$ is called a regular value of A on B, where A is the operator defined in Definition* 4.4, *in two cases:*

1. *if $y \notin A(B)$,*
2. *if $y \in A(B)$ and for any $u \in A^{-1}y \cap B$ the differential $A'(u)$ at the point u maps E onto E_0. All the values which are not regular are called critical values.*

The following Sard–Smale theorem was proved in [6].

Theorem 4.3 (Sard–Smale). *Let A be a mapping of class $C^r, r \in \mathbf{N}$, on a domain $O \subset E$. Let A be Fredholm on O with the index $k, r \geq k$. Let $B \subset O$ be a closed set, and let A be proper on B. Then the set D of all regular values of A on B is an open and dense set in E_0.*

The following theorem was proved in [19].

THEOREM 4.4 (Kato). *Let E and E_0 be Banach spaces, and let operator $K = T + S : E \to E_0$ be a sum of continuous linear Fredholm operator T and a compact linear operator S. Then K is the Fredholm operator with $Ind\, K = Ind\, T$.*

Let us introduce the Banach spaces

$$\begin{aligned} E &= \mathbf{V} \cap \mathbf{H}^2(\Omega_f) \times H^1(\Omega_f) \cap L_0^2(\Omega_f) \times Y, \\ \tilde{E} &= \mathbf{V} \cap \mathbf{H}^2(\Omega_f) \times H^1(\Omega_f) \cap L_0^2(\Omega_f) \times \tilde{Y}, \text{ and} \\ E_0 &= \mathbf{L}^2(\Omega_f) \times L^2(\Omega), \end{aligned}$$

where

$$Y = \left\{ T \in H_D^1(\Omega), -\nabla \cdot (\kappa \nabla T) \in L^2(\Omega), \frac{\partial T}{\partial \mathbf{n}} \in X \right\}$$

and

$$\tilde{Y} = \left\{ T \in H_D^1(\Omega), -\nabla \cdot (\kappa \nabla T) \in L^2(\Omega), \frac{\partial T}{\partial \mathbf{n}}|_{\Gamma_C} = 0 \right\}$$

are Banach spaces equipped with the norm

$$\|T\|_Y = \|T\|_{\tilde{Y}} = \|T\|_{H^1(\Omega)} + \|\nabla \cdot (\kappa \nabla T)\|_{L^2(\Omega)} + \left\| \frac{\partial T}{\partial \mathbf{n}} \right\|_{L^2(\Gamma_C)}.$$

We introduce the mapping $A : E \to E_0$ defined by

$$(4.69)\ A(u, p, T) = (-\nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \alpha T \mathbf{e}_2, -\nabla \cdot (\kappa \nabla T) + (\mathbf{u} \cdot \nabla) T).$$

PROPOSITION 4.5. *The mapping $A \in C^\infty(E, E_0)$ is a Fredholm mapping on E with the index $k = \dim X$. Moreover, for any $r > 0$ the mapping A is proper on $B_r = \{x \in E_0, \|x\|_{E_0} \le r\}$.*

Proof. We can write out the mapping A in the form

$$A(\mathbf{u}, p, T) = A_1(\mathbf{u}, p, T) + A_2(\mathbf{u}, p, T),$$

where $A_1(\mathbf{u}, p, T) = (-\nu \Delta \mathbf{u} + \nabla p - \alpha T \mathbf{e}_2, -\nabla \cdot (\kappa \nabla T)) : E \to E_0$ is the linear continuous operator and

$$A_2(\mathbf{u}, p, T) = ((\mathbf{u} \cdot \nabla)\mathbf{u}, \chi_{\Omega_f} (\mathbf{u} \cdot \nabla) T)$$

is the bilinear operator. By the estimate

$$\|A_2(\mathbf{u}, p, T)\|_{\mathbf{L}^2(\Omega_f)} \le C \left(\|\mathbf{u}\|_{\mathbf{L}^4(\Omega_f)} \|\nabla \mathbf{u}\|_{\mathbf{L}^4(\Omega_f)} + \|\mathbf{u}\|_{\mathbf{L}^\infty(\Omega_f)} \|\nabla T\|_{L^2(\Omega_f)} \right)$$

this bilinear operator is continuous. This implies (see [26]) $A \in C^\infty(E, E_0)$. One can write out the derivative of this operator at the point $(\mathbf{u}_0, p_0, T_0)$ as follows:

$$A_2'(\mathbf{u}_0, p_0, T_0)[\mathbf{u}, p, T] = ((\mathbf{u}_0 \cdot \nabla)\mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u}_0, (\mathbf{u}_0 \cdot \nabla) T + (\mathbf{u} \cdot \nabla) T_0).$$

Also, the estimate

$$\begin{aligned} &\|A_2'(\mathbf{u}_0, p_0, T_0)[\mathbf{u}, p, T]\|_{\mathbf{L}^2(\Omega_f) \times L^2(\Omega)} \\ &\qquad \le C\big(\|\mathbf{u}_0\|_{\mathbf{L}^4(\Omega_f)} \|\nabla \mathbf{u}\|_{\mathbf{L}^4(\Omega_f)} + \|\mathbf{u}\|_{\mathbf{L}^4(\Omega_f)} \|\nabla \mathbf{u}_0\|_{\mathbf{L}^4(\Omega_f)} \\ &\qquad\quad + \|\mathbf{u}\|_{\mathbf{L}^\infty(\Omega_f)} \|\nabla T_0\|_{L^2(\Omega_f)} + \|\mathbf{u}_0\|_{\mathbf{L}^\infty(\Omega_f)} \|\nabla T\|_{L^2(\Omega_f)} \big) \end{aligned}$$

implies that for each $(\mathbf{u}_0, p_0, T_0) \in E$ the derivative $A_2'(\mathbf{u}, p, T)$ is the compact operator. Note that $A'(0)(\mathbf{u}, p, T) = A_1(\mathbf{u}, p, T)$. Since $Im\, A_1 = E_0$, the index of this operator equals $\dim X$. Applying the Theorem 4.4, we obtain that A is a Fredholm mapping with the index $k = \dim X$ as desired.

Now let us prove that the mapping A is proper. Let K be an arbitrary compact set in E_0. Then $A^{-1}K \cap B_r$ is a bounded set in E. On the other hand, $A_2(A^{-1}K \cap B_r)$ is a compact set, and, obviously, $K_1 = K + A_2(A^{-1}K \cap B_r)$ is also a compact set. Then $A^{-1}K \cap B_r \subset A_1^{-1}K_1 \cap B_r \subset M_1 + M_2$, where

$$M_1 = \{(\mathbf{u}, p, T) \in \tilde{E}, A_1(\mathbf{u}, p, T) \in K_1\}$$

and

$$M_2 = \left\{(0, 0, T), T \in B, \nabla \cdot \kappa \nabla T \equiv 0, g \in X, \frac{\partial T}{\partial \mathbf{n}}|_{\Gamma_C} = g, \|g\|_{L^2(\Gamma_C)} \le r\right\}.$$

Since M_2 is a closed finite dimensional set, it is compact. On the other hand, since the operator A_1 is an isomorphism between $\tilde{E}$ and E_0, the set M_1 is also compact. This implies immediately that $M_1 + M_2$ is compact. The proof of the theorem is finished. □

Now we have the following theorem.

THEOREM 4.6. *Let $\mathcal{V}$ be a linear finite dimensional space in $L^2(\Gamma_C)$. Then the set $\mathcal{O}$ contains a set $\mathcal{O}_1$ which is open and dense in $\mathbf{L}^2(\Omega_f) \times L^2(\Omega)$.*

Proof. We prove this theorem by contradiction. Let $\mathrm{M} \subset E_0$ be the set of $(\mathbf{f}, Q) \in \mathbf{L}^2(\Omega_f) \times L^2(\Omega)$ such that there exists a solution of the system (4.1)–(4.5) with $\lambda = 0$ and $\mathbf{Int}\,\overline{\mathrm{M}} \neq \emptyset$. Let us show that M belongs to the set of critical values of the mapping (4.69). By Theorem 2.3, we have that $A(E) = E_0$. But by definition, for all $(\mathbf{f}, Q) \in \mathrm{M}$, there exists $(\hat{\mathbf{u}}, \hat{p}, \hat{T}) \in E$ such that $A(\hat{\mathbf{u}}, \hat{p}, \hat{T}) = (\mathbf{f}, Q)$ and $(\hat{\mathbf{u}}, \hat{p}, \hat{T})$ is a solution to optimal problem (3.9) with the optimality system (4.1)–(4.5) and $\lambda = 0$. Hence $\mathrm{Ker}\, A'^*(\hat{\mathbf{u}}, \hat{p}, \hat{T})(\cdot) \neq \emptyset$ and $\mathrm{Im}\, A'(\hat{\mathbf{u}}, \hat{p}, \hat{T}) \neq E_0$. This implies that M belongs to the set of critical values of the mapping A. On the other hand, by Proposition 4.5, $A \in C^\infty(E, E_0)$ is a Fredholm mapping with index $k = \dim X$. Finally, let $\mathcal{P}$ be an open bounded subset of $\mathbf{Int}\,\overline{\mathrm{M}}$. Then there exists a ball $B_r \subset E$ such that for all $(\mathbf{f}, Q) \in \mathcal{P}$ all global minimum to problem (3.9) belong to B_r. By Proposition 4.5 the mapping A is proper on B_r. Hence, by the Sard–Smale theorem, the set of regular values is dense in E_0. We obtain a contradiction, and thus the proof is completed. □

REFERENCES

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

[2] G. V. ALEKSEEV, *Solvability of stationary boundary control problems for heat convection equations*, Siberian Math. J., 39 (1998), pp. 844–858.

[3] F. ABERGEL AND F. CASAS, *Some optimal control problems of multistate equations appearing in fluid mechanics*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 223–247.

[4] F. ABERGEL AND R. TEMAM, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynamics, 1 (1990), pp. 303–325.

[5] J. BOLAND AND W. LAYTON, *Error analysis for finite element methods for steady natural convection problems*, Numer. Funct. Anal. and Optim., 11 (1990), pp. 449–483.

[6] A.V. BABIN AND M.I. VISHIK, *Attractors of Evolution Equations*, North-Holland, Amsterdam, 1992.

[7] K. DEIMLING, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, Heidelberg, 1985.

[8] M. DESAI AND K. ITO, *Optimal controls of Navier–Stokes equations*, SIAM J. Control Optim., 32 (1994), pp. 1428–1446.

[9] H. O. Fattorini and S. S. Sritharan, *Optimal controls for viscous flow problems*, Proc. Roy. Soc. London Ser. A, 439 (1992), pp. 81–102.

[10] A. V. Fursikov, *Properties of solutions to some extrimal problems related to the Navier-Stokes system*, Mat. Sb., 118 (1982), pp. 323–349.

[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman Advanced Publishing Program, Boston, 1985.

[12] M. Gunzburger, L. Hou, and T. Svobodny, *Heating and cooling control of temperature distributions along boundaries of flow domains*, J. Math. Systems Estim. Control, 3 (1993), pp. 147–172.

[13] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with distributed and Neumann controls*, Math. Comp., 57 (1991), pp. 123–151.

[14] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.

[15] M. Gunzburger, L. Hou, and T. Svobodny, *Boundary velocity control of incompressible flow with an application to viscous drag reduction*, SIAM J. Control Optim., 30 (1992), pp. 167–181.

[16] M. Gunzburger and H.-C. Lee, *Analysis, approximation, and computation of a coupled solid/fluid temperature control problem*, Comput. Methods Appl. Mech. Engrg., 118 (1994), pp. 133–152.

[17] K. Ito and S. S. Ravindran, *Optimal control of thermally convected fluid flows*, SIAM J. Sci. Comput., 19 (1998), pp. 1847–1869.

[18] K. Ito, J. S. Scroggs, and H. T. Tran, *Optimal control of thermally coupled Navier-Stokes equation*, in Optimal Design and Control (Blacksburg, VA, 1994), Progr. Systems Control Theory 19, Birkhauser Boston, Boston, MA, 1995, pp. 199–214.

[19] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1976.

[20] H.-C. Lee and B. C. Shin, *Piecewise optimal distributed controls for 2D Boussinesq equations*, Math. Methods Appl. Sci., 23 (2000), pp. 227–254.

[21] J. L. Lions, *Control of Distributed Singular System*, Gauthier-Villars, Paris, 1985.

[22] F.-L. Liu and D. L. Russell, *Solutions of the Boussinesq equation on a periodic domain*, J. Math. Anal. Appl., 194 (1995), pp. 78–102.

[23] G. Savaré, *Regularity and perturbation results for mixed second order elliptic problems*, Comm. Partial Differential Equations, 22 (1997), pp. 869–899.

[24] G. Savaré, *Elliptic equations in Lipshitz domains*, J. Funct. Anal., 152 (1998), pp. 176–201.

[25] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1979.

[26] V. Tikhomirov, *Fundamental Principles of the Theory of Extremal Problems*, Wiley, Chichester, UK, 1982.

SIAM J. Control Optim.
Vol. 39, No. 2, pp. 478–491

A NECESSARY AND SUFFICIENT CONDITION FOR BOUNDED-INPUT BOUNDED-STATE STABILITY OF NONLINEAR SYSTEMS*

ANDREA BACCIOTTI† AND LUISA MAZZI†

Abstract. The main purpose of this work is to prove a converse theorem for bounded-input bounded-state stability of nonlinear systems in the framework of Liapunov's second method. The construction gives rise to an upper semicontinuous time-dependent Liapunov function. In certain cases, the monotonicity conditions can be checked by means of contingent directional derivatives.

Key words. Liapunov functions, external stability, nonlinear systems

AMS subject classifications. 93D25, 93C10, 93D05

PII. S0363012999350572

Introduction. In this work we address the problem of characterizing uniform bounded-input bounded-state (UBIBS) stability for time-varying, continuous time, finite-dimensional nonlinear systems

$$\dot{x} = f(t, x, u), \qquad t \geq 0, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m, \tag{0.1}$$

in the framework of Liapunov's second method. Roughly speaking, a system is said to be bounded-input bounded-state stable if for each initial pair (t_0, x_0) and each admissible input $u = u(t)$, the corresponding solution $x = \varphi(t)$ of (0.1) remains bounded for all $t \geq t_0$. In order to conceive a more formal and precise definition, we insert a short digression about the time-invariant linear case. First, let us specify some terminology and notation.

Throughout this paper, an *admissible input* is a measurable, essentially bounded function $u = u(t) : [0, +\infty) \to \mathbb{R}^m$. The euclidean norm of a finite-dimensional vector v is denoted by $|v|$. If $u(\cdot)$ is an admissible input, we shall write

$$\|u(\cdot)\| = \operatorname*{ess\,sup}_{t \geq 0} |u(t)|.$$

Let I be an interval, and let $u(\cdot)$ be an admissible input. A (Carathéodory) *solution* of (0.1) on I corresponding to the input $u(\cdot)$ is a function $\varphi(t)$ which is absolutely continuous on every compact subinterval of I and which satisfies the differential equation $\dot{x} = f(t, x, u(t))$ almost everywhere (a.e.) on I. When we want to emphasize the dependence of a solution $\varphi(t)$ on the admissible input $u(\cdot)$ and a given initial pair (t_0, x_0), we write $x = \varphi(t; t_0, x_0, u(\cdot))$.

Let a linear system of the form

$$\dot{x} = Ax + Bu \tag{0.2}$$

be given.

It is well known that A is Hurwitz (if and) only if there exist positive constants γ_1, γ_2, and α such that for each initial state x_0 and each admissible input

*Received by the editors January 13, 1999; accepted for publication (in revised form) February 1, 2000; published electronically September 7, 2000.

http://www.siam.org/journals/sicon/39-2/35057.html

†Dipartimento di Matematica, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy (bacciotti@polito.it, mazzi@polito.it).

$u : [0, +\infty) \to \mathbb{R}^m$ one has

$$|\varphi(t; 0, x_0, u(\cdot))| \leq \gamma_1 |x_0| e^{-\alpha t} + \gamma_2 \cdot \|u(\cdot)\| \tag{0.3}$$

for each $t \geq 0$. In particular, (0.3) implies the weaker condition

$$|\varphi(t; 0, x_0, u(\cdot))| \leq \gamma_1 |x_0| + \gamma_2 \cdot \|u(\cdot)\| \tag{0.4}$$

for each $t \geq 0$. This last condition has a characterization in terms of the Kalman canonical decomposition of (0.2).

Proposition 0.1. *Let the linear system* (0.2) *be given and let A_{11} and A_{22} be, respectively, the controllable part and the uncontrollable part of A which appear in its Kalman decomposition. There exist $\gamma_1, \gamma_2 > 0$ such that* (0.4) *holds for each x_0 and each $u(\cdot)$, if and only if*

(i) *the real part of all the eigenvalues of A_{11} is negative;*

(ii) *the real part of all the eigenvalues of A_{22} is nonpositive and all the eigenvalues with zero real part are simple.*

In fact, if the uncontrollable part is stable, the corresponding solutions remain bounded. Hence, the controllable part can be viewed as a subsystem with Hurwitz matrix and bounded input, so that (0.3) applies.

The estimates (0.3) and (0.4) turn out to be very useful in studying the behavior of the state response of a system with respect to bounded inputs, as classical classroom examples show. Both (0.3) and (0.4) can be extended to nonlinear systems. In this context, it is convenient to introduce the formalism of the so-called functions of class $\mathcal{K}$. Although functions of class $\mathcal{K}$ are a classical tool in stability theory, the details of the definition may vary from author to author. We adopt the following definition.

Definition 0.1. *A real function defined on $[0, +\infty)$ is said to be of class $\mathcal{K}^\infty$ if it is continuous, strictly increasing, $\alpha(0) \geq 0$, and $\lim_{r \to +\infty} \alpha(r) = +\infty$.*

The definitions of class $\mathcal{K}^\infty$ available in the literature often include the requirement $\alpha(0) = 0$. Since in this paper we are interested in the behavior for large x, we also include functions with $\alpha(0) > 0$.

The extension of (0.3) to the nonlinear case gives rise to the well-known notion of *input-to-state stability* introduced by Sontag ([11]; see also [7, p. 501] for a definition in the time-varying setting) and thoroughly studied in the last decade. In particular, it has been proved that at least in the time-invariant case, input-to-state stability can be characterized by means of $\mathcal{C}^\infty$ Liapunov functions which satisfy certain growth conditions and a monotonicity condition expressed in differential form (see [12]).

In this paper we are mainly interested in the extension of (0.4) to the nonlinear case.

Definition 0.2. *We say that system* (0.1) *satisfies the UBIBS stability property if there exist maps $\gamma_1, \gamma_2 \in \mathcal{K}^\infty$ such that, for each initial pair (t_0, x_0), each admissible input $u : [0, +\infty) \to \mathbb{R}^m$, and each $t \geq t_0$,*

$$|\varphi(t; t_0, x_0, u(\cdot))| \leq \gamma_1(|x_0|) + \gamma_2(\|u(\cdot)\|).$$

This definition has several equivalent formulations, reported in the following proposition, whose proof is omitted. (The equivalence between (i) and (ii) is proved in [1] for the time-invariant case; the equivalence between (ii) and (iii) can be obtained by analogous arguments.)

PROPOSITION 0.2. *The following statements are equivalent:*

(i) *System* (0.1) *has the UBIBS stability property.*

(ii) *For each $R > 0$ there exists $S > 0$ such that for each initial pair (t_0, x_0) and each essentially bounded input $u : [0, +\infty) \to \mathbb{R}^m$*

$$|x_0| \leq R, \quad \|u(\cdot)\| \leq R, \quad t \geq t_0 \quad \Longrightarrow \quad |\varphi(t; t_0, x_0, u(\cdot))| \leq S.$$

(iii) *There exists $\Sigma \in \mathcal{K}^\infty$ such that for each $t_0 > 0$, each $R > 0$, each $x_0 \in \mathbb{R}^n$, and each admissible input $u(\cdot) : [0, +\infty) \to \mathbb{R}^m$ one has*

$$|x_0| \leq R, \quad \|u(\cdot)\| \leq R, \quad t \geq t_0 \quad \Longrightarrow \quad |\varphi(t; t_0, x_0, u(\cdot))| \leq \Sigma(R).$$

Remark 0.1. From (ii) it is clear in particular that the term "uniform" in Definition 0.2 is intended to have a double interpretation: the dependence of S on R is affected neither by the choice of x_0 (within the ball of radius R) nor by the choice of t_0.

Notice also that for the map Σ in (iii), we necessarily have $\Sigma(r) \geq r$ for each $r \geq 0$. Indeed from (iii) it follows in particular that

$$|\varphi(t_0; t_0, x_0, u(\cdot))| = |x_0| \leq \Sigma(R)$$

for each $|x_0| \leq R$.

The main contribution of this paper is presented in section 1. We prove that (0.1) is UBIBS stable if and only if there exists a Liapunov function satisfying some suitable uniform growth conditions and the monotonicity condition. Our characterization is valid under very general assumptions. The "if" part is easy and basically known (see, for instance, [4]). The "only if" part is actually a converse Liapunov theorem for UBIBS stability, and it requires a delicate and original procedure. The principal difficulty is the construction of the Liapunov function in the presence of inputs. In section 1, we also point out that in general an UBIBS stable system has no continuous Liapunov function.

At this point, a comparison with some earlier, related results and some comments on the regularity issue are necessary. It is evident that for systems without input, UBIBS stability reduces to *Lagrange stability* (sometimes referred to as *uniform boundedness of solutions*). A characterization of Lagrange stability by means of nonnecessarily continuous Liapunov functions is immediate and can be deduced, for instance, from [15, Ch. V] and [2]. On the other hand, it is well known that continuous Liapunov functions for Lagrange stability may not exist (see [8], [4]). More regular Liapunov functions can be obtained either by strengthening the notion of stability (see again [8], [4]) or for systems with Lipschitz continuous right-hand side [15, Ch. V].

Time-invariant systems deserve a particular mention. In this case, one would expect Liapunov functions independent of t. However, a famous counterexample [5, p. 87] shows that in general continuous time-invariant Liapunov functions for Lagrange stable time-invariant systems do not exist, even if the right-hand side is $\mathcal{C}^\infty$. To this respect, it is worth mentioning Yorke's (rarely quoted) paper [14], where it is proved that for any stable, time-invariant, Lipschitz continuous system there exists a lower semicontinuous Liapunov function.

Keeping in mind this last result, we come back to system (0.1) and UBIBS stability. In section 2, we prove that if a suitable Lipschitz condition is fulfilled, then the Liapunov function constructed in section 1 is upper semicontinuous.

Finally, in section 3 we show that, provided that $V(t,x)$ is upper semicontinuous, it is possible to recognize the monotonicity property by means of a differential inequality involving the so-called generalized contingent directional derivative. In fact, this requires further restrictions on both f and the class of admissible inputs.

Dealing with Liapunov functions and UBIBS stability, it is worth mentioning the pioneering paper [13]: while in the present paper we characterize UBIBS stability by means of a single Liapunov function, in [13] the authors use a whole family $V_a(t,x)$ of them. The parameter a is related to the bound of the input variable.

1. A necessary and sufficient condition. In this section, we need the following hypotheses on (0.1).

Assumption (H). For any admissible input $u(t)$, for any pair $(t_0, x_0) \in \mathbb{R}^+ \times \mathbb{R}^n$, there exists a unique solution $\varphi(t; t_0, x_0, u(\cdot))$, defined for all $t \geq 0$.

We shall systematically employ the reformulation of the UBIBS stability property (iii) of Proposition 0.2. We denote by $H^\eta = \{x \in \mathbb{R}^n \ : \ |x| > \eta\}$ for $\eta > 0$.

THEOREM 1.1. *Assume that* (0.1) *satisfies assumption* (H). *System* (0.1) *has the UBIBS stability property if and only if there exist* $\eta > 0$*, functions* $a, b, \gamma \in \mathcal{K}^\infty$*, and a function* $V(t,x) : [0,+\infty) \times H^\eta \to \mathbb{R}$ *such that*

(i) $a(|x|) \leq V(t,x) \leq b(|x|)$ *for each* $t \geq 0$ *and each* $x \in H^\eta$*;*

(ii) $\gamma(r) \geq r + \eta$ *for each* $r \geq 0$*;*

(iii) *for each* $R > 0$*, for each input* $u(\cdot)$*, and for each solution* $\varphi(t)$ *of* (0.1) *corresponding to the input* $u(\cdot)$*,*

$$t_1 < t_2 \quad \Longrightarrow \quad V(t_1, \varphi(t_1)) \geq V(t_2, \varphi(t_2)) \tag{1.1}$$

provided that $|u(t)| \leq R$ *and* $|\varphi(t)| \geq \gamma(R)$ *for all* $t \in [t_1, t_2]$.

A function $V(t,x)$ satisfying the properties listed above is called a UBIBS-*weak Liapunov function* or, simply, a weak Liapunov function. For the sake of brevity, we shall refer to property (iii) of Theorem 1.1 by saying that V is *decreasing along the solutions* of (0.1).

We remark that the Liapunov function $V(t,x)$ we get in Theorem 1.1 is allowed to be discontinuous.

Proof of the sufficient part. Without loss of generality, we assume that $a(0) = 0$, so that $a^{-1} \in \mathcal{K}^\infty$. We actually show that (iii) of Proposition 0.2 is fulfilled with

$$\Sigma(R) = (a^{-1} \circ b \circ \gamma)(R) + \gamma(R).$$

First, we notice that according to this definition, $\Sigma \in \mathcal{K}^\infty$, and in fact, for each $R > 0$, we have $\Sigma(R) > \gamma(R) > R$.

Now, fix $R > 0$ and take an arbitrary initial instant $t_0 \geq 0$, an initial state x_0 with $|x_0| \leq R$, and an input $u(\cdot) : [t_0, +\infty) \to \mathbb{R}^m$ such that $\|u(\cdot)\| \leq R$. Assume that for some $t_2 > t_0$, it may happen that

$$|\varphi(t_2; t_0, x_0, u(\cdot))| > S = \Sigma(R).$$

Since $\Sigma(R) > \gamma(R) > R$ and $\varphi(t; t_0, x_0, u(\cdot))$ is continuous, there exists an instant $t_1 \in (t_0, t_2)$ such that $|\varphi(t_1; t_0, x_0, u(\cdot))| = \gamma(R)$, while $|\varphi(t; t_0, x_0, u(\cdot))| \geq \gamma(R)$ for $t \in [t_1, t_2]$.

For the sake of simplicity, we set $x_1 = \varphi(t_1; t_0, x_0, u(\cdot))$ and $x_2 = \varphi(t_2; t_0, x_0, u(\cdot))$. Since a is strictly increasing we have

$$a^{-1}(b(\gamma(R))) < \Sigma(R) = S \quad \Longrightarrow \quad b(\gamma(R)) < a(S).$$

In conclusion,

$$V(t_1, x_1) \leq b(|x_1|) = b(\gamma(R)) < a(S) < a(|x_2|) \leq V(t_2, x_2),$$

while $|u(t)| \leq R$ and $|\varphi(t; t_0, x_0, u(\cdot))| \geq \gamma(R)$ for $t \in [t_1, t_2]$. This is a contradiction to (1.1). □

Proof of the necessary part. Of course, the necessary part of Theorem 1.1 is more difficult, since it requires the construction of $V(t, x)$. We begin by introducing some notation and by proving one lemma.

For $t > 0, x \in \mathbb{R}^n$ $(x \neq 0)$, we define a subset of admissible inputs:

$$U(t, x) := \{u(\cdot) : \exists c > 0 \text{ such that } |u(s)| \leq \min_{0 \leq \theta \leq t} |\varphi(\theta; t, x, u(\cdot))| - c \quad \forall s \geq 0\}.$$

The definition of $U(t, x)$ can be extended to the case $t = 0$ by taking $U(0, x) = \{0\}$, where 0 denotes the zero input.

LEMMA 1.1. *Assume that* (0.1) *has the UBIBS property. We set* $\eta = \Sigma(1)$. *Then,* $U(t, x) \neq \emptyset$ *for all* $x \in H^\eta$, *for all* $t \geq 0$.

Proof. We prove that $u(\cdot) \equiv 0 \in U(t, x)$ for all $t > 0$ and $x \in H^\eta$. The case $t = 0$ follows trivially from the definition. When $t \neq 0$, we consider a solution $\varphi(\cdot; t, x, 0)$. Since it is continuous, $|\varphi(\cdot; t, x, 0)|$ has a minimum at a point $\tau \in [0, t]$. Let us assume by contradiction that $\varphi(\tau; t, x, 0) = 0$. By uniqueness of solutions, $x = \varphi(t; \tau, 0, 0)$. Then, according to the UBIBS stability property (applied with $R = 1$), we should have $|x| \leq \Sigma(1)$. But this is impossible, since $|x| > \eta = \Sigma(1)$.

Thus, it follows that $\min_{0 \leq \theta \leq t} |\varphi(\theta; t, x, 0)| > 0$, and there exists $c > 0$ such that $0 \equiv |u(s)| < \min_{0 \leq \theta \leq t} |\varphi(\theta; t, x, 0)| - c$ for each $s \geq 0$. □

We are now ready to define a function $V(t, x)$. As in Lemma 1.1, let $\eta = \Sigma(1)$. For each $t \geq 0$ and each $x \in H^\eta$, we set

$$V(t, x) := \inf_{u(\cdot) \in U(t,x)} \min_{s \in [0,t]} |\varphi(s; t, x, u(\cdot))|. \tag{1.2}$$

The proof that this function actually fulfills all the requirements of Theorem 1.1 will be accomplished through several steps.

Step 1. There exists $b \in \mathcal{K}^\infty$ such that $V(t, x) \leq b(|x|)$ for $x \in H^\eta$.

The proof of Step 1 is obvious: indeed, from (1.2) it turns out immediately that $V(t, x) \leq |x|$.

Step 2. There exists $a \in \mathcal{K}^\infty$ such that $a(|x|) \leq V(t, x)$ for $x \in H^\eta$.

To begin with, we remark that $\Sigma^{-1}(r)$ is defined for $r \in [\Sigma(0), +\infty)$, where $\Sigma(0) < \eta$. Moreover, Σ^{-1} is continuous, strictly increasing on its domain, and $\lim_{r \to +\infty} \Sigma^{-1}(r) = +\infty$. We prove that the inequality

$$\Sigma^{-1}(|x|) \leq V(t, x) \tag{1.3}$$

holds for $x \in H^\eta$. According to (1.2), for $t \geq 0, x \in H^\eta$, and every fixed $\varepsilon > 0$ there exist $u(\cdot) \in U(t, x)$ and $\tau \in [0, t]$ such that

$$|\varphi(\tau; t, x, u(\cdot))| < V(t, x) + \varepsilon.$$

Let $\xi = \varphi(\tau; t, x, u(\cdot))$ and set $R = V(t, x) + \varepsilon$. Therefore

$$\min_{s \in [0,t]} |\varphi(s; t, x, u(\cdot))| < R.$$

Since $u(\cdot) \in U(t,x)$, we have $|u(s)| \leq R$ for each $s \geq 0$ as well. Applying the definition of UBIBS stability, we obtain

$$|x| = |\varphi(t;\tau,\xi,u(\cdot))| \leq \Sigma(R) = \Sigma(V(t,x)+\varepsilon),$$

which yields

$$\Sigma^{-1}(|x|) \leq V(t,x)+\varepsilon.$$

Inequality (1.3) follows since the choice of ε was arbitrary and independent of the pair (t,x). To complete the proof of Step 2, it is sufficient to take any $a \in \mathcal{K}^\infty$ such that $a(r) = \Sigma^{-1}(r)$, for $r > \eta$.

Let us now set $\gamma(r) = \Sigma(r) + \eta$ for $r \geq 0$. By virtue of Remark 0.1, we trivially have the following.

Step 3. γ fulfills (ii) of Theorem 1.1.

Thus, it remains to prove the following step.

Step 4. $V(t,x)$ is decreasing along the trajectories of (0.1).

Let us fix $R > 0$ and let $0 \leq t_1 < t_2$. Pick up any admissible input $u(\cdot)$ such that $|u(t)| \leq R$ for each $t \in [t_1,t_2]$, and let $\varphi(t)$ be any solution of (0.1), defined for $t \in [t_1,t_2]$, corresponding to the input $u(\cdot)$ and lying on the closed region $\overline{H^{\gamma(R)}}$. Finally, we set $x_1 = \varphi(t_1)$ and $x_2 = \varphi(t_2)$. Since $\gamma(R) > \eta$, the values $V(t_1,x_1)$ and $V(t_2,x_2)$ are well defined.

Claim A.

$$V(t_2,x_2) \leq \min_{t_1 \leq t \leq t_2} |\varphi(t)|.$$

We prove only the case $t_1 > 0$, because the case $t_1 = 0$ is a simplified version of this one. Let us consider the admissible input:

$$u_0(t) = \begin{cases} u(t), & t \in [t_1,t_2], \\ 0, & t \notin [t_1,t_2]. \end{cases}$$

Obviously, $|u_0(t)| \leq R$ for each $t \geq 0$. On the other hand,

$$\varphi(t;t_2,x_2,u_0(\cdot)) = \varphi(t)$$

for $t \in [t_1,t_2]$, so that

$$|\varphi(t;t_2,x_2,u_0(\cdot))| \geq \gamma(R) > \Sigma(R) \geq R \tag{1.4}$$

for $t \in [t_1,t_2]$. Assume that there exists $\theta < t_1$ such that

$$|\varphi(\theta;t_2,x_2,u_0(\cdot))| \leq R.$$

According to the UBIBS stability property, this should imply $|x_2| \leq \Sigma(R)$, a contradiction to (1.4). Hence, we see that

$$|\varphi(t;t_2,x_2,u_0(\cdot))| > R$$

even when $t < t_1$.

The minimum of $|\varphi(\cdot;t_2,x_2,u_0(\cdot))|$ on $[0,t_2]$ exists and, of course, it is strictly greater than R. Hence, there exists some $c > 0$ such that

$$|u_0(t)| \leq R < \min_{s \in [0,t_2]} |\varphi(s;t_2,x_2,u_0(\cdot))| - c$$

for $t \geq 0$. This implies that $u_0(\cdot) \in U(t_2,x_2)$.

We are now ready to prove Claim A. Indeed,

$$\begin{aligned}V(t_2,x_2) &= \inf_{u\in U(t_2,x_2)} \min_{0\le t\le t_2} |\varphi(t;t_2,x_2,u(\cdot))|\\ &\le \min_{0\le t\le t_2} |\varphi(t;t_2,x_2,u_0(\cdot))| \le \min_{t_1\le t\le t_2} |\varphi(t)|\end{aligned}$$

as required.

To carry on the proof of Step 4, we shall proceed now by contradiction. Assume that the opposite inequality holds:

$$V(t_1,x_1) < V(t_2,x_2).$$

By definition, there exist $\tau \le t_1$, $u_1(\cdot) \in U(t_1,x_1)$, and $\xi_1 \in \mathbb{R}^n$ such that $\xi_1 = \varphi(\tau;t_1,x_1,u_1(\cdot))$ and

$$V(t_1,x_1) \le |\xi_1| < V(t_2,x_2). \tag{1.5}$$

Claim B.

$$\min_{t\le t_1} |\varphi(t;t_1,x_1,u_1(\cdot))| > R.$$

Otherwise, for some $\tilde t \le t_1$ we should have

$$|\tilde\xi| = |\varphi(\tilde t;t_1,x_1,u_1(\cdot))| \le R.$$

Since $u_1(\cdot)\in U(t_1,x_1)$, we should have

$$|u_1(t)| < |\tilde\xi| \le R$$

for $t \ge 0$, as well. Then, the UBIBS stability property should imply $|x_1| \le \Sigma(R)$, and this is a contradiction (recall that $x_1 = \varphi(t_1) = \varphi(t_1;t_1,x_1,u_1(\cdot)) \in \overline{H^{\gamma(R)}}$ and $\gamma(R) \ge \Sigma(R)$).

Claim B is thus proved. We now define

$$u_2(t) = \begin{cases} u(t), & t\in[t_1,t_2],\\ u_1(t), & t\notin[t_1,t_2].\end{cases}$$

Of course,

$$\varphi(t;t_2,x_2,u_2(\cdot)) = \begin{cases} \varphi(t;t_1,x_1,u_1(\cdot)), & t\in[0,t_1),\\ \varphi(t), & t\in[t_1,t_2].\end{cases}$$

Claim C.

$$u_2(\cdot)\in U(t_2,x_2).$$

We need to show that

$$|u_2(t)| \le \min_{0\le s\le t_2} |\varphi(s;t_2,x_2,u_2(\cdot))| - c$$

for some $c>0$ and $t\ge 0$. From (1.5) and Claim A it follows that

$$\min_{0\le s\le t_1} |\varphi(s;t_1,x_1,u_1(\cdot))| < V(t_2,x_2) \le \min_{t_1\le s\le t_2} |\varphi(s)|.$$

In other words, $|\varphi(\cdot;t_2,x_2,u_2(\cdot))|$ reaches its minimum on the interval $[0,t_1]$. Hence,

$$\min_{0\le s\le t_2}|\varphi(s;t_2,x_2,u_2(\cdot))| = \min_{0\le s\le t_1}|\varphi(s;t_1,x_1,u_1(\cdot))|. \tag{1.6}$$

Since $u_1(\cdot)\in U(t_1,x_1)$, there exists $c_1>0$ such that

$$|u_1(t)|\le \min_{0\le s\le t_1}|\varphi(s;t_1,x_1,u_1(\cdot))|-c_1$$

for all $t\ge 0$. But $u_1(\cdot)$ and $u_2(\cdot)$ coincide for $t<t_1$ and $t>t_2$. Taking into account (1.6) we therefore have

$$|u_2(t)|\le \min_{0\le s\le t_2}|\varphi(s;t_2,x_2,u_2(\cdot))|-c_1 \tag{1.7}$$

for $t<t_1$ and $t>t_2$. On the other hand, for $t\in[t_1,t_2]$ we have $|u_2(t)|=|u(t)|\le R$. According to Claim B and the continuity of solutions, there exists $c_2>0$ such that

$$\min_{0\le s\le t_1}|\varphi(s;t_1,x_1,u_1(\cdot))|-c_2>R$$

or, from (1.6),

$$\min_{0\le s\le t_2}|\varphi(s;t_2,x_2,u_2(\cdot))|-c_2>R\ge|u(t)|.$$

Taking $c=\min\{c_1,c_2\}$, we finally see that (1.7) holds even when $t\in[t_1,t_2]$. This shows that $u_2(\cdot)\in U(t_2,x_2)$ and Claim C is proved.

We can now get the conclusion. Claim C implies that

$$V(t_2,x_2)\le \min_{0\le s\le t_2}|\varphi(s;t_2,x_2,u_2(\cdot))|.$$

On the other hand, it is clear that

$$\xi_1=\varphi(\tau;t_1,x_1,u_1(\cdot))=\varphi(\tau;t_2,x_2,u_2(\cdot))$$

and that $V(t_2,x_2)\le|\xi_1|$. Comparing this last conclusion with (1.5), we obtain a contradiction. The proof of Step 4 and of the necessary part of Theorem 1.1 is completed. □

Remark 1.1. In general, it is not possible to construct a continuous Liapunov function $V(t,x)$ for any UBIBS-stable system (0.1). As a counterexample, we can take the scalar system $\dot{x}=f(x)+bu$, where $f(x)$ is the function defined in [8, p. 269], and $b=0$. It is clear that the system is UBIBS-stable. If $V(t,x)$ is a continuous weak Liapunov function, $V(t,x)$ also should be a continuous Liapunov function for the Lagrange-stable equation $\dot{x}=f(x)$. But it is well known that such a function cannot exist (see [4]).

Remark 1.2. The proof of Theorem 1.1 can be carried out even if the set $U(t,x)$ is replaced by

$$U_0(t,x)=\{u(\cdot)\ \text{ such that } |u(s)|\le \min_{0\le\theta\le t}|\varphi(\theta;t,x,u(\cdot))|\ \forall s\le 0\}$$

(see [3], where a preliminary version of Theorem 1.1 was showed). In this case, Lemma 1.1 becomes obvious and the proof of Step 4 is simplified as well. However, if the function V is defined by means of the set $U_0(t,x)$, the development of the next section becomes impossible. For this reason, we preferred to adopt from the beginning the stronger construction based on the set $U(t,x)$.

Remark 1.3 (time-invariant systems). If we consider a time-invariant system

$$\dot{x} = f(x,u)$$

and we assume the existence and uniqueness of solutions for all $t \in \mathbb{R}$, we are able to obtain a time-invariant Liapunov function $V(x)$ satisfying Theorem 1.1.

In this case we define the set of admissible inputs as

$$U(x) = \Big\{ u(\cdot) \ : \ |u(s)| \le \inf_{\theta \le 0} |\varphi(\theta; x, u(\cdot))| \Big\}.$$

Note that here, according to Remark 1.2, we have taken $c = 0$.

The Liapunov function $V(x)$ can be defined as

$$V(x) = \inf_{u \in U(x)} \ \inf_{\theta \le 0} |\varphi(\theta; x, u(\cdot))|.$$

With these definitions, the proof of Theorem 1.1 works, with minor modifications.

2. An upper semicontinuous Liapunov function. The goal of this section is to prove that, under additional assumptions, the Liapunov function constructed in the previous section is at least semicontinuous. To begin with, the next lemma shows that the definition of the set $U(t,x)$ introduced in the previous section can be reformulated in an apparently stronger form.

LEMMA 2.1. *For any pair* (t,x),

$$\begin{aligned} U(t,x) = \{\ u(\cdot) :\ & \exists\delta > 0,\ \exists c > 0 \text{ such that} \\ & |u(s)| \le \min\nolimits_{0 \le \theta \le t+\delta} |\varphi(\theta; t, x, u(\cdot))| - c \ \ \forall s \ge 0\ \}. \end{aligned}$$

The proof is an easy consequence of continuity of solutions, and it is left to the reader.

As already suggested, in order to proceed we need to strengthen the hypotheses on system (0.1). So, from now on, we shall consider system (0.1) under the following additional assumption.

Assumption (L).

(i) $f(t,x,u)$ is locally bounded.

(ii) $f(t,x,u)$ is locally Lipschitz with respect to x. More precisely, we require that for each initial pair (t_0,x_0) and each admissible input $u(t)$, there exists a compact neighborhood Ω of (t_0,x_0) in $\mathbb{R}^+ \times \mathbb{R}^n$ and a positive function $L(t)$ such that $L(t)$ is integrable and

$$|f(t,x,u(t)) - f(t,y,u(t))| \le L(t)|x-y| \tag{2.1}$$

for each pair $(t,x), (t,y) \in \Omega$.

The main result of this section is based on Lemma 2.2. The arguments of its proof are classical (see, for instance, [9, section 5]). A short sketch of the proof is reported, for the reader's convenience. Let

$$\mathcal{B}_0(\delta) := \{(t,x) \in [0,+\infty) \times \mathbb{R}^n \ : \ |t-t_0| < \delta,\ |x-x_0| < \delta\},$$

where (t_0,x_0) is a fixed point of $[0,+\infty) \times \mathbb{R}^n$.

LEMMA 2.2. *If* (0.1) *satisfies Assumptions* (H) *and* (L), *for any admissible input* $u(\cdot)$ *there exist* $K > 0$, $\rho \in (0,1)$ *such that, for all* $\delta \in (0,\rho)$ *and all* $(t,x) \in \mathcal{B}_0(\delta)$,

$$|\varphi(s;t,x,u(\cdot)) - \varphi(s;t_0,x_0,u(\cdot))| \leq K\delta \quad \forall s \in [0, t_0+1].$$

Proof. Let t_0, x_0, and the admissible input $u(t)$ be fixed. The solution $\varphi(s;t_0,x_0,u(\cdot))$ has compact image I on the interval $[0,t_0+1]$. Let N be the tubular neighborhood of radius 1 of I, and let $L(t)$ be a function such that (2.1) holds for any $(t,x),(t,y) \in N$. Moreover, let M be an upper bound for the norm of $f(t,x,u(t))$ on N. Finally, let ρ be such that

$$\rho \leq \frac{1}{2(1+M)}\ e^{-\int_0^{t_0+1} L(s)ds} < 1$$

and

$$\mathcal{B}_0(\rho) \subseteq N.$$

We claim that for each $(t,x) \in \mathcal{B}_0(\rho)$, $\varphi(s;t,x,u(\cdot)) \in N$ for each $s \in [0,t_0+1]$. We study separately the intervals $[0,t_0]$ and $[t_0,t_0+1]$, and we see the proof in the first case; in the second one, the proof is analogous, and it is left to the reader.

By contradiction, let us assume that there exists $s \in [0,t_0]$ such that $\varphi(s;t,x,u(\cdot)) \notin N$. Let

$$\tau = \sup\{s \in [0,t_0] \ : \ \varphi(s;t,x,u(\cdot)) \notin N\}.$$

Since (2.1) holds in the interval $[\tau,t_0]$, we have

$$\begin{aligned} &|\varphi(\tau;t,x,u(\cdot)) - \varphi(\tau;t_0,x_0,u(\cdot))| \\ &\leq |x-x_0| + \left|\int_t^\tau |f(\theta;\varphi(\theta;t,x,u(\cdot)),u(\theta)) - f(\theta;\varphi(\theta;t_0,x_0,u(\cdot)),u(\theta))|\ d\theta\right| \\ &\quad + \left|\int_{t_0}^t f(\theta;\varphi(\theta;t_0,x_0,u(\cdot)),u(\theta))\ d\theta\right| \\ &\leq |x-x_0| + \left|\int_t^\tau L(\theta)\,|\varphi(\tau;t,x,u(\cdot)) - \varphi(\tau;t_0,x_0,u(\cdot))|\,d\theta\right| + M\,|t-t_0| \end{aligned}$$

since $|x-x_0| + M\,|t-t_0| \leq (1+M)\rho$, and by Gronwall's inequality we have

$$\leq (1+M)\rho e^{\left|\int_t^\tau L(\theta)\ d\theta\right|} \leq (1+M)\rho e^{\int_0^{t_0+1} L(\theta)\ d\theta} \leq \frac{1}{2}.$$

Therefore, $\varphi(\tau;t,x,u(\cdot)) \in N$ together with $\varphi(\tau-s;t,x,u(\cdot))$ for any small enough s, and this is a contradiction to the choice of τ.

Let us now consider any $\delta \in (0,\rho)$. For $(t,x) \in \mathcal{B}_0(\delta)$, the image of the corresponding solution lies in N, so that (2.1) applies, with the same $L(t)$ used before. By repeating the argument based on Gronwall's inequality, we get

$$|\varphi(s;t,x,u(\cdot)) - \varphi(s;t_0,x_0,u(\cdot))| \leq K\delta \quad \forall s \in [0,t_0+1],$$

where $K = (1+M)e^{\int_0^{t_0+1} L(s)ds}$. □

THEOREM 2.1. *If the input system* (0.1) *satisfies the UBIBS property and Assumptions* (H) *and* (L)*, then there exists an UBIBS-weak upper semicontinuous Liapunov function.*

Proof. We show that, under Assumption (L), the Liapunov function defined in Theorem 1.1 is upper semicontinuous.

Step 1. For any pair (t_0, x_0), if $u \in U(t_0, x_0)$, there exists $\delta_0 > 0$ such that $u \in U(t, x)$ for all $(t, x) \in \mathcal{B}_0(\delta_0)$.

By Lemma 2.1, there exist $\delta_1 > 0$, $c > 0$ such that

$$|u(s)| \leq \min_{0 \leq \theta \leq t_0 + \delta_1} |\varphi(\theta; t_0, x_0, u(\cdot))| - c \quad \forall s \geq 0. \tag{2.2}$$

Let $K > 0$ be as in Lemma 2.2. Then, for each $\delta \in (0, \rho)$ we have

$$|\varphi(\theta; t_0, x_0, u(\cdot))| \leq |\varphi(\theta; t, x, u(\cdot))| + K\delta, \qquad \theta \in [0, t_0 + 1] \tag{2.3}$$

for any $(t, x) \in \mathcal{B}_0(\delta)$.

Let us choose $\delta_0 < \min\{\rho, \delta_1, \frac{c}{K}\}$, and let us set $c' = c - K\delta_0 > 0$. By (2.2) and (2.3), if $(t, x) \in \mathcal{B}_0(\delta_0)$, then

$$\begin{aligned}
|u(s)| &\leq \min_{0 \leq \theta \leq t_0 + \delta_1} |\varphi(\theta; t_0, x_0, u(\cdot))| - c \\
&\leq \min_{0 \leq \theta \leq t_0 + \delta_0} |\varphi(\theta; t_0, x_0, u(\cdot))| - c \\
&\leq \min_{0 \leq \theta \leq t_0 + \delta_0} |\varphi(\theta; t, x, u(\cdot))| - c + K\delta_0 \\
&\leq \min_{0 \leq \theta \leq t} |\varphi(\theta; t, x, u(\cdot))| - c + K\delta_0 \\
&\leq \min_{0 \leq \theta \leq t} |\varphi(\theta; t, x, u(\cdot))| - c' \quad \forall s \geq 0.
\end{aligned}$$

This means that $u(s) \in U(t, x)$ for all $(t, x) \in \mathcal{B}_0(\delta_0)$.

Step 2. $V(t, x)$ is upper semicontinuous.

Let V be defined at (t_0, x_0). By construction, for all $\varepsilon > 0$ there exist $v \in U(t_0, x_0)$ and $\tau \in [0, t_0]$ such that

$$|\varphi(\tau; t_0, x_0, v(\cdot))| \leq V(t_0, x_0) + \frac{1}{2}\varepsilon. \tag{2.4}$$

By Step 1, there exists $\delta_0 > 0$ such that $v \in U(t, x)$ for all $(t, x) \in \mathcal{B}_0(\delta_0)$. Therefore, by the definition of $V(t, x)$,

$$V(t, x) \leq |\varphi(\tau; t, x, v(\cdot))| \quad \forall (t, x) \in \mathcal{B}_0(\delta_0). \tag{2.5}$$

By Lemma 2.2, there exists $K = K(v) > 0$ such that, for all $\delta < \delta_0$ and all $(t, x) \in \mathcal{B}_0(\delta)$,

$$|\varphi(\tau; t, x, v(\cdot))| \leq |\varphi(\tau; t_0, x_0, v(\cdot))| + K\delta. \tag{2.6}$$

Summing up (2.4), (2.5), and (2.6) we get

$$\begin{aligned}
V(t, x) \leq |\varphi(\tau; t, x, v(\cdot))| &\leq |\varphi(\tau; t_0, x_0, v(\cdot))| + K\delta \\
&\leq V(t_0, x_0) + \frac{1}{2}\varepsilon + K\delta.
\end{aligned}$$

If we choose $\delta < \min\left(\delta_0, \frac{\varepsilon}{2K}\right)$, we get

$$V(t, x) < V(t_0, x_0) + \varepsilon$$

for all $(t, x) \in \mathcal{B}_0(\delta)$. $\square$

3. Monotonicity and generalized derivatives. As far as Liapunov-like functions are concerned, the regularity issue is strictly related to the monotonicity condition (iii) of Theorem 1.1. For instance, if the function $V(t,x)$ turns out to be differentiable for each $t>0$ and for $x \in H^\eta$, condition (iii) of Theorem 1.1 can be reformulated in the following way:

$$\forall R>0,\ \forall x \in H^\eta,\ \forall u \in \mathbb{R}^m,$$
$$|u| \le R,\ |x| \ge \gamma(R) \Longrightarrow \frac{\partial V}{\partial t}(t,x) + \nabla_x V(t,x)\cdot f(t,x,u) \le 0, \quad \text{a.e. } t \ge 0. \tag{3.1}$$

The obvious advantage of (3.1), with respect to (iii) of Theorem 1.1, is that it can be checked without explicitly solving the differential equation. We are interested in getting some characterization of (iii) in terms of a differential inequality even with the less regular Liapunov function we are able to obtain.

It is well known that when $V(t,x)$ is locally Lipschitz, its monotonicity along a trajectory $x=\varphi(t)$ can be checked by looking at the sign of one of the Dini derivatives of $V(t,\varphi(t))$ (see [15]). Unfortunately, the Liapunov function we get is only upper semicontinuous. Therefore, we need to adjust previous results to our weaker assumptions. We start with the following lemma, which establishes a relation between monotonicity of semicontinuous functions of a single variable and Dini derivatives. The lemma can be obtained as a consequence of [6, Theorem 1.4], or as an easy generalization of the argument used in [10, p. 347].

LEMMA 3.1. *Let $\psi : [a,b] \to \mathbb{R}$ be upper semicontinuous. Then*

(i) *ψ is nonincreasing on $[a,b]$ if and only if $\overline{D^-}\psi(t) \le 0$ for each $t \in [a,b]$;*

(ii) *ψ is nondecreasing on $[a,b]$ if and only if $\underline{D^+}\psi(t) \ge 0$ for each $t \in [a,b]$.*

This result, together with the use of contingent derivatives, allows us to get a characterization of monotonicity of $V(t,x)$ along trajectories by means of a differential inequality. We recall the following definition.

DEFINITION 3.1. *Given a function $g : \mathbb{R}^N \to \mathbb{R}$, the upper left contingent generalized directional derivative of g at z, with respect to $v \in \mathbb{R}^N$, is*

$$\overline{D_K^-}g(z,v) = \limsup_{\substack{h\to 0^- \\ w\to v}} \frac{g(z+hw)-g(z)}{h}.$$

When g is locally Lipschitz, $\overline{D_K^-}g(z,v)$ is equal to the corresponding directional Dini derivative, but in general $\overline{D_K^-}g(z,v)$ is greater than the corresponding Dini derivative.

Since we are interested in the case $N=1+n$, $z=(t,x)$, and $v=(1,w)$, we shall use the notation $\overline{D_K^-}g(t,x,w)$ instead of $\overline{D_K^-}g((t,x),(1,w))$.

In order to state the main result of this section, we introduce appropriate assumptions.

Assumption (C).

(i) $f(t,x,u)$ is continuous.

(ii) The class of admissible inputs is restricted to the one of continuous maps $u : [0,+\infty) \to \mathbb{R}^m$.

Under these hypotheses, any solution $\varphi(t)$ corresponding to an admissible input is a classical solution, and it satisfies (0.1) everywhere on its domain.

THEOREM 3.1. *Let* (0.1) *satisfy Assumption* (C), *and let* $V(t,x)$ *be an upper semicontinuous function satisfying the following property:* $\exists\gamma \in \mathcal{K}^\infty$ *such that,* $\forall R > 0$, $\forall x \in \mathbb{R}^n$, $\forall u \in \mathbb{R}^m$, $\forall t \geq 0$,

$$|u| \leq R,\ |x| \geq \gamma(R) \quad \Longrightarrow \quad \overline{D_k^-} V(t,x,f(t,x,u)) \leq 0.$$

Then V *is decreasing along the solutions of* (0.1)*; that is, condition* (iii) *of Theorem* 1.1 *is fulfilled.*

Proof. Let $\psi(t) = V(t,\varphi(t))$, where $\varphi(t)$ is as in (iii) of Theorem 1.1 and $u(t)$ is the corresponding input. By Lemma 3.1, it is sufficient to prove that $\overline{D^-}\psi(t) \leq 0$ for each $t \in [t_1,t_2]$. We have

$$\begin{aligned}\overline{D^-}\psi(t) &= \limsup_{h\to 0^-} \frac{\psi(t+h)-\psi(t)}{h}\\ &= \limsup_{h\to 0^-} \frac{V(t+h,\varphi(t+h)) - V(t,\varphi(t))}{h}\\ &= \limsup_{h\to 0^-} \frac{V(t+h,\varphi(t)+h\dot\varphi(t)+o(h)) - V(t,\varphi(t))}{h}.\end{aligned}$$

We write $o(h) = h\alpha(h)$, where $\lim_{h\to 0}\alpha(h) = 0$. Then

$$\overline{D^-}\psi(t) = \limsup_{h\to 0^-} \frac{V(t+h,\varphi(t)+h[\dot\varphi(t)+\alpha(h)]) - V(t,\varphi(t))}{h}.$$

Notice that $w(h) = \dot\varphi(t) + \alpha(h) \to \dot\varphi(t)$ for $h \to 0$. Thus, if we set $x = \varphi(t)$ and $v = \dot\varphi(t) = f(t,x,u(t))$, we get

$$\begin{aligned}\overline{D^-}\psi(t) &\leq \limsup_{\substack{h\to 0^-\\ w\to v}} \frac{V(t+h,x+hw) - V(t,x)}{h}\\ &= \overline{D_k^-}V(t,x,v) = \overline{D_k^-}V(t,x,f(t,x,u(t))).\end{aligned}$$

The conclusion is immediate. □

REFERENCES

[1] V. ANDRIANO, A. BACCIOTTI, AND G. BECCARI, *Global stability and external stability of dynamical systems*, J. Nonlinear Anal., 28 (1997), pp. 1167–1185.

[2] E. ARZARELLO AND A. BACCIOTTI, *On stability and boundedness for Lipschitzian differential inclusions: The converse of Lyapunov's theorems*, Set-Valued Anal., 5 (1997), pp. 377–390.

[3] A. BACCIOTTI, *A Liapunov-like characterization of bounded-input bounded-state stability*, in Proceedings of IFAC-NOLCOS Conference, Enschede, 1998, pp. 493–498.

[4] A. BACCIOTTI AND L. ROSIER, *Liapunov and Lagrange stability: Inverse theorems for discontinuous systems*, Math. Control Signals Systems, 11 (1998), pp. 101–128.

[5] N.P. BATHIA AND G.P. SZEGÖ, *Stability Theory of Dynamical Systems*, Springer-Verlag, Berlin, 1970.

[6] F.H. CLARKE, YU. S. LEDYAEV, R.J. STERN, AND P.R. WOLENSKI, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[7] M. KRSTIĆ, I. KANNELLAKOPOULOS, AND P. KOKOTOVIĆ, *Nonlinear and adaptive control design*, Wiley-Interscience, New York, 1995.

[8] J. KURZWEIL AND I. VRKOČ, *The converse theorem of Lyapunov and Persidskij concerning the stability of motion*, Czech. Math. J., 82 (1957), pp. 254–272 (in Russian).

[9] Y. LIN, E.D. SONTAG, AND Y. WANG, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[10] N. Rouche, P. Habets, and M. Laloy, *Stability Theory by Liapunov's Direct Method*, Springer-Verlag, New York, 1977.

[11] E.D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.

[12] E.D. Sontag and Y. Wang, *Characterization of the input-to-state stability property*, System Control Lett., 24 (1995), pp. 351–359.

[13] P.P. Varaiya and R. Liu, *Bounded-input bounded-output stability of nonlinear time-varying differential systems*, SIAM J. Control, 4 (1966), pp. 698–704.

[14] J.A. Yorke, *Differential inequalities and non-Lipschitz scalar functions*, Math. Systems Theory, 4 (1970), pp. 140–153.

[15] T. Yoshizawa, *Stability Theory by Liapunov's Second Method*, Mathematical Society of Japan, Tokyo, 1966.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 492–511

MAXIMAL SOLUTIONS IN DECENTRALIZED SUPERVISORY CONTROL*

ARD OVERKAMP† AND JAN H. VAN SCHUPPEN‡

Abstract. The decentralized supervisory control problem is to construct for a discrete-event system a set of supervisors each observing only part of the system and each controlling only part of the events such that the interconnection of the system and the supervisors meets control objectives of safety and liveness. Definitions are provided of the concepts of a maximal solution, of a Nash equilibrium, and of a strong Nash equilibrium for a set of supervisors with as order relation the inclusion relation on the set of closed-loop languages. The main result is that a set of supervisors is a maximal solution if and only if it is a strong Nash equilibrium. A procedure to determine a Nash equilibrium is described and illustrated by an example. There is no guarantee that the procedure halts in finite time. However, in the case that it halts in finite time, then it is proven that a Nash equilibrium is obtained.

Key words. discrete-event system, decentralized supervisory control, maximal solution, Nash equilibrium

AMS subject classification. 93C30

PII. S0363012997321139

1. Introduction.

The purpose of this paper is to show how the concept of a Nash equilibrium can be used to obtain maximal solutions of supervisors for decentralized control of discrete-event systems.

Decentralized supervisory control problems arise very naturally in protocol design problems for computer and communication networks but also occur in transportation and manufacturing problems. The network may be modeled as a discrete-event system. The physical separation between the sender and the receiver implies that observations of the operation of the network are available only locally. The problem is then to synthesize a set of controllers in a protocol problem—one at the sender end and one at the receiver end of the communication channel. The interconnection of the network with the supervisors has then to meet control objectives of safety and liveness according to a specification.

Decentralized control is a conceptually difficult problem. Results are available mainly for decentralized control of linear systems and of stochastic systems (see the survey [18]). Fundamental results are partly based on the analogy with game, dynamic game, and team problems. The decentralized supervisory control problem was formulated by R. Cieslak et al. (see [2]), in which the alternating bit protocol is used as an example. The authors presented a necessary and sufficient condition for the existence of a controller for which the closed-loop language equals a specified language. A generalization of this result to the closed-loop language was fit between an upper and lower bound, and an analysis of the set of supervisors was derived by K. Rudie and W. M. Wonham (see [14, 17]). These authors also studied the protocol synthesis problem (see [15, 16]). P. Kozak and W. M. Wonham proposed another solution procedure based on projection of the supremal supervisor (see [4]). Recent work on

*Received by the editors May 7, 1997; accepted for publication (in revised form) February 28, 2000; published electronically September 7, 2000.
http://www.siam.org/journals/sicon/39-2/32113.html

†ORTEC Consultants B.V., P.O. Box 490, 2800 AL Gouda, The Netherlands (aoverkamp@ortec.nl).

‡CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (J.H.van.Schuppen@cwi.nl).

decentralized supervisory control of nondeterministic systems using prioritized synchronization is presented by R. Kumar and M.A. Shayman (see [5]). The synthesis procedures proposed so far do not satisfy the need of engineering decentralized control problems. The performance of the resulting controllers is in general too conservative.

The approach of this paper is based on an analogy with dynamic game problems. The restriction is imposed to consider a model with only two supervisors. The concept of a maximal solution of a pair of supervisors is defined with respect to the inclusion relation on the set of languages of the closed-loop system. Because the set of pairs of supervisors is a large discrete set, there may be many such pairs. The determination of a maximal pair of supervisors is achieved indirectly. The concept of a strong Nash equilibrium of a pair of supervisors is introduced based on analogy with game theory. It is shown that a pair of supervisors that is a strong Nash equilibrium is also a maximal solution and conversely. A procedure is proposed to compute a strong Nash equilibrium of a pair of supervisors. The procedure is illustrated with an example.

A description of the paper by section follows. Section 2 contains a definition of a discrete-event system that differs slightly from the case usually considered in the literature, the formulation of the decentralized supervisory control problem, and the definition of maximal solution and Nash equilibrium. The result, that a pair of supervisors that is a strong Nash equilibrium is also a maximal solution and conversely, is established in section 3. The procedure for a Nash equilibrium is stated in section 4. Section 5 contains conclusions.

The results of this paper were announced in the conference paper [11] and form part of the thesis [10, Chap. 6] of the first author.

2. Problem formulation.

2.1. Framework.

A simple framework will be introduced that allows us to concentrate on the decentralized aspects of the control problem.

Throughout this paper denote the global set of events by Σ, the global set of controllable events by Σ_{c}, the uncontrolled system by G, and the specification by E. The discrete-event system will be modeled as a finite state automaton with the notation $G = (\Sigma, Q, \delta, q_0)$, with Q the discrete state space, $\delta : \Sigma \times Q \to Q$ the transition function, and q_0 the initial state. For a string $s \in \Sigma^*$ denote by $\bar{s} \subseteq \Sigma^*$ the set of prefixes of this string.

Definition 2.1. *A* supervisor *or* discrete-event controller *is defined by a triple*

$$S = (\Sigma(S),\ \Sigma_{\mathrm{c}}(S),\ \gamma(S)),$$

where

$$\Sigma(S) \subseteq \Sigma, \quad \Sigma_{\mathrm{c}}(S) \subseteq \Sigma(S), \quad \gamma(S) : p_s(\mathrm{L}(G)) \to 2^{\Sigma_{\mathrm{c}}(S)},$$

and p_s is the projection from Σ to $\Sigma(S)$.

Define the controlled language of supervisor S with respect to G or, for short, the language *of S as*

$$\mathrm{L}(S/G) = \{s \in \mathrm{L}(G) : \forall v\sigma \in \bar{s}, \sigma \notin \gamma(S, p_s(v))\}.$$

Note that $\mathrm{L}(S/G) \subseteq \Sigma^*$.

Let $\mathcal{C}(\Sigma_a)$ denote the set of all supervisors S with event set $\Sigma(S) = \Sigma_a$ and controllable event set $\Sigma_c(S) = \Sigma_c \cap \Sigma_a$. The function $\gamma(S)$ will be called the control law *of supervisor S. Note that $\gamma(S, s)$ is defined for all $s \in p_s(\mathrm{L}(G))$.*

The control law $\gamma(S)$ maps each trace $s \in p_s(\mathrm{L}(G))$ onto the set of disabled events. In the literature, often the set of enabled events is specified [13]. Both approaches are equivalent.

In the definition above the set of controllable events is taken to be contained in the set of events observable by the supervisor. In general it is possible that a supervisor can influence events it cannot observe. In [10, Sect. 5.2] it is shown how in this situation a control problem can be remodeled such that all controllable events are observable. As that reference is not widely available, the approach is briefly sketched. The idea is based on flags. Controllable, unobservable events are usually implemented with flags. If a flag is set, then the event can execute. If the flag is cleared, then the event is disabled. The plant is remodelled such that it includes the events that set and clear the flags. These so-called flag events are observable and controllable. In this remodeled plant the original events are no longer controllable as they are enabled and disabled via the flag events. If the flag events are, via projection, removed from the language of the remodeled plant, then the language of the original plant is obtained.

Attention will be focused on the decentralized aspects of the supervisory control problem. Marking, nondeterminism, or failure semantics will not be considered. The argument for a simple framework also justifies the restriction to only two supervisors. The authors are confident that in the future the results can be extended to more general frameworks and more supervisors.

The basic supervisory control problem needs to be redefined for the new framework. Note that supervisors, as stated in Definition 2.1, can disable only controllable events. So they are always complete. It is not necessary to add a completeness requirement as is done in [13].

Definition 2.2. *Consider a discrete-event system and a legal language $L(E) \subset \Sigma^*$. The basic supervisory control problem (BSCP) is to find a supervisor S, such that $\mathrm{L}(S/G) \subseteq \mathrm{L}(E)$.*

Ramadge and Wonham showed that there exists a unique supremal solution to this control problem. This supremal can be effectively computed [13]. It is characterized by a language called the supremal controllable sublanguage contained in $\mathrm{L}(G) \cap \mathrm{L}(E)$. As the notion of controllability will not be used any further, we refer the interested reader to the given reference for more information. The only aspect of controllability that will be used in this chapter is that the supremal controllable language can be effectively computed.

Definition 2.3. *Let $K^\uparrow$ be the supremal controllable sublanguage contained in $\mathrm{L}(G) \cap \mathrm{L}(E)$. The* supremal supervisor*, denoted by $S^\uparrow$, is defined by*

$$\gamma(S^\uparrow, s) = \begin{cases} \{\sigma \in \Sigma_c : s\sigma \in \mathrm{L}(G) \text{ and } s\sigma \notin K^\uparrow\}, & \text{if } s \in K^\uparrow, \\ \emptyset, & \text{otherwise.} \end{cases}$$

It is not difficult to show that $\mathrm{L}(S^\uparrow/G) = K^\uparrow$. As $S^\uparrow$ is supremal it holds for all supervisors S which solve the given BSCP, that $\mathrm{L}(S/G) \subseteq \mathrm{L}(S^\uparrow/G)$.

In this paper it will be assumed that the BSCP is already solved and that the supremal supervisor $S^\uparrow$ is given. It is sufficient to find a supervisor that implements $S^\uparrow$, with respect to the implementation relation defined below. Proposition 2.7 shows that this is a valid approach. A supervisor implements the supremal supervisor if and only if the supervisor solves the BSCP.

Definition 2.4. *Let S_a, S_b be two supervisors such that $\Sigma(S_a) = \Sigma(S_b)$. Supervisor S_a* implements *S_b, denoted by $S_a \sqsubseteq S_b$, if*

$$\gamma(S_b, s) \subseteq \gamma(S_a, s) \quad \forall\, s \in p(\mathrm{L}(S_a/G)),$$

where p is the projection on $\Sigma(S_a) = \Sigma(S_b)$.

Supervisor S_a implements S_b if it disables at least as much as S_b.

LEMMA 2.5. *Let S_a, S_b be two supervisors such that $\Sigma(S_a) = \Sigma(S_b)$.*

$$S_a \sqsubseteq S_b \Rightarrow \mathrm{L}(S_a/G) \subseteq \mathrm{L}(S_b/G).$$

The proof of the preceding lemma and that of Proposition 2.7 are simple and may be found in [10, Chap. 6].

The following example will show why the converse of Lemma 2.5 does not hold.

Example 2.6. Let G be the system such that $\mathrm{L}(G) = \{\varepsilon, \mathtt{a}\}$. Define S_a by $\gamma(S_a, \varepsilon) = \emptyset$ and $\gamma(S_a, \mathtt{a}) = \emptyset$. Define S_b by $\gamma(S_b, \varepsilon) = \emptyset$ and $\gamma(S_a, \mathtt{a}) = \{\mathtt{a}\}$. Then $\mathrm{L}(S_a/G) = \{\varepsilon, \mathtt{a}\} = \mathrm{L}(S_b/G)$, but $\gamma(S_b, \mathtt{a}) \not\subseteq \gamma(S_a, \mathtt{a})$. So $S_a \not\sqsubseteq S_b$. □

In [10, Thm. 2.17] it was shown that in the failure-semantics-based framework a supervisor solves the BSCP if and only if it implements the supremal supervisor. Proposition 2.7 states the same result for the framework of this paper. Because the proof is analogous to that of [10, Thm. 2.17], it is omitted.

PROPOSITION 2.7. *Let the uncontrolled system G, the specification E, and the set of controllable events Σ_c be given. Let $S^\uparrow$ be the supremal supervisor of the BSCP.*

$$\forall S \in \mathcal{C}(\Sigma),\ S \sqsubseteq S^\uparrow \iff \mathrm{L}(S/G) \subseteq \mathrm{L}(S^\uparrow/G) \iff \mathrm{L}(S/G) \subseteq \mathrm{L}(E).$$

In the rest of this paper we will consider control problems that place extra constraints on the supervisor besides the ones given in the BSCP. Proposition 2.7 states that we can first solve the BSCP to get the supremal supervisor $S^\uparrow$. Next we can look for supervisors that satisfy the extra constraints and that implement $S^\uparrow$. In this last step we can concentrate on the extra requirement. As we are mainly interested in the extra requirements imposed by the decentralized nature of the control problem, we will assume that the first step is already solved and that the supremal supervisor $S^\uparrow$ is given.

DEFINITION 2.8. *The* basic supervisory synthesis problem *(BSSP) is to find a supervisor $S \in \mathcal{C}(\Sigma(S^\uparrow))$ such that $S \sqsubseteq S^\uparrow$.*

Often in the literature supervisors are defined as languages instead of control maps. We choose to use control maps as they allow us to divide the control problem into two steps. In the first step the supremal supervisor is synthesized. In this step the controllability condition plays an important role. In the second step we can concentrate on the decentralized aspect of the control problem. Proposition 2.7 shows that we do not have to consider the controllability condition in this step. If supervisors are defined as languages, then also the problem can be divided into two parts. The synthesis problem of the second part is then defined as follows: find a supervisor S such that $\mathrm{L}(S/G) \subseteq \mathrm{L}(S^\uparrow/G)$ and $\mathrm{L}(S/G)$ is controllable. It is necessary to check for controllability, as $\mathrm{L}(S/G) \subseteq \mathrm{L}(S^\uparrow/G)$ does not imply that $\mathrm{L}(S/G)$ is controllable. So in the second step we still have to consider controllability. Using control maps, we can forget about controllability in the second step and concentrate on the decentralized aspects of the control problem. It is not too difficult to adapt the results of this paper to a language-based approach.

2.2. Decentralized supervisory synthesis problem. Up until now we have only looked at supervisors that can observe the whole event set and that enable or disable all controllable events. Now we will look at the decentralized control problem where we have two supervisors, each observing a part of the event set, and each controlling only part of the controllable events (see Figure 1). The two supervisors

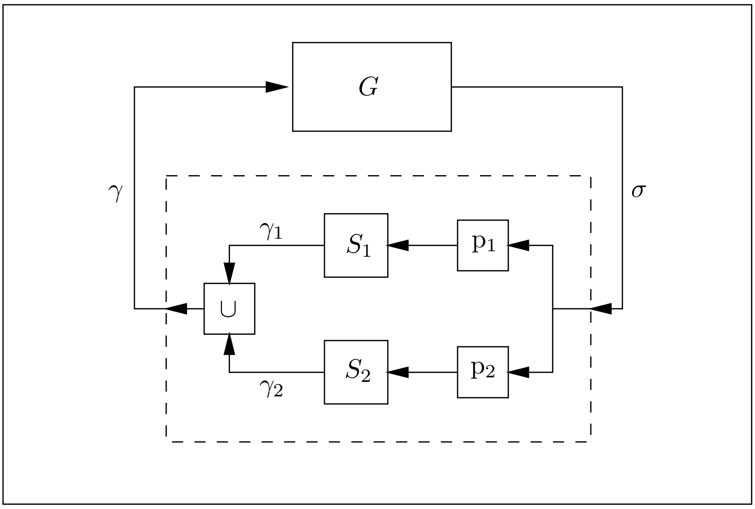


Fig. 1. *The decentralized supervisory control problem.*

together have to control G such that the language of the controlled system is contained in the language of E. Note that the specification is given for the whole controllable system. This is usually referred to as a global specification [14, 17]. If the specification can be decomposed into two local specifications, one for each supervisor, then the decentralized control problem can be reduced to two independent supervisory control problems. In each of these local control problems a single supervisor is synthesized. This control problem has already been solved by F. Lin and W. M. Wonham [7]. In what follows we will assume that the specification is global and cannot be decomposed into local specifications.

As stated before, we will assume the BSCP is already solved and the supremal supervisor $S^{\uparrow}$ is known. By Proposition 2.7 it is sufficient to find a decentralized implementation of $S^{\uparrow}$ to solve the decentralized supervisory control problem.

First it will be defined how two decentralized supervisors co-operate. An event is disabled by the combination of the two supervisors if it is disabled by at least one of them.

Definition 2.9. *Let S_1 and S_2 be two supervisors. The* composition *of S_1 and S_2 is denoted $S_1 \wedge S_2$ and defined by*

$$\begin{aligned}\Sigma(S_1 \wedge S_2) &= \Sigma_1 \cup \Sigma_2,\\ \Sigma_{\mathrm{c}}(S_1 \wedge S_2) &= \Sigma_{\mathrm{c}}(S_1) \cup \Sigma_{\mathrm{c}}(S_2),\\ \gamma(S_1 \wedge S_2, s) &= \gamma(S_1, \mathrm{p}_1(s)) \cup \gamma(S_2, \mathrm{p}_2(s)) \qquad \forall\, s \in \mathrm{p}_{1,2}(\mathrm{L}(G)),\end{aligned}$$

where p_1 *denotes the projection on* $\Sigma(S_1)$, p_2 *denotes the projection on* $\Sigma(S_2)$, *and* $\mathrm{p}_{1,2}$ *denotes the projection on* $\Sigma(S_1 \wedge S_2)$.

Proposition 2.10. $\mathrm{L}(S_1 \wedge S_2/G) \;=\; \mathrm{L}(S_1/G) \cap \mathrm{L}(S_2/G)$.

Proof. The reasoning follows from Definition 2.1.

$$\begin{aligned}& s \in \mathrm{L}(S_1 \wedge S_2/G)\\ &\iff s \in \mathrm{L}(G)\, \forall v\sigma \in \overline{s},\ \sigma \notin \gamma(S_1 \wedge S_2, \mathrm{p}_{1,2}(v))\\ &\iff s \in \mathrm{L}(G)\, \forall v\sigma \in \overline{s},\ \sigma \notin \gamma(S_1, \mathrm{p}_1(v)),\ \sigma \notin \gamma(S_2, \mathrm{p}_2(v))\\ &\iff s \in \mathrm{L}(S_1/G),\ s \in \mathrm{L}(S_2/G)\\ &\iff s \in \mathrm{L}(S_1/G) \cap \mathrm{L}(S_2/G). \qquad \Box\end{aligned}$$

Definition 2.11. *Consider the discrete-event system specified before and a global specification. Let the supremal supervisor $S^{\uparrow}$ be given. Let $\Sigma_1, \Sigma_2 \subseteq \Sigma$ be two event sets such that $\Sigma_1 \cup \Sigma_2 = \Sigma$. The decentralized supervisory synthesis problem (DSSP) is to find a pair of supervisors $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ such that*

$$S_1 \wedge S_2 \sqsubseteq S^{\uparrow}.$$

In this definition we made two important assumptions. The one is that $\Sigma_1 \cup \Sigma_2 = \Sigma$. The other is that, according to the definition of $\mathcal{C}(\Sigma_i)$, the set of controllable events of supervisor S_i, $\Sigma_{\mathrm{c},i}$, is equal to $\Sigma_i \cap \Sigma_{\mathrm{c}}$ for $i = 1, 2$.

Consider the case where $\Sigma_1 \cup \Sigma_2 \subsetneqq \Sigma$. If $\Sigma_{\mathrm{c}} \subseteq \Sigma_1 \cup \Sigma_2$, then we can compute the supremal supervisor under partial observation, with observation alphabet $\Sigma_1 \cup \Sigma_2$. See [2, 6] and section [10, Sect. 5.1]. Equivalently to Proposition 2.7, it can be shown that a supervisor implements this supremal supervisor if and only if it solves the control problem under partial observation. We can assume that this control problem is already solved and that the supremal supervisor under partial observation is given. So this control problem can be reduced to the DSSP.

If $\Sigma_{\mathrm{c}} \nsubseteq \Sigma_1 \cup \Sigma_2$, then the control problem can be remodeled in such a way that all controllable events are observable. See [10, Sect. 5.2].

The other assumption is that $\Sigma_{\mathrm{c},i} = \Sigma_i \cap \Sigma_{\mathrm{c}}$, $i = 1, 2$. That is, the controllable events of supervisor S_i are observable by S_i, and an event that is controllable by $S^{\uparrow}$ and observable by S_i is also controllable by S_i. This is the same constraint as given by Rudie [14, 17] under which decomposability of the closed-loop language is necessary and sufficient for the existence of a decentralized solution. It is argued that in most communication problems these constraints are satisfied. Again, as we want to keep the model simple, we do not consider systems that fail to satisfy this constraint. The authors hope that in the future these constraints can be relaxed.

2.3. Maximal solutions. Traditionally in discrete-event control, supervisors are synthesized that restrict the uncontrolled system as little as possible. A solution is considered optimal if the language of the system controlled by this optimal supervisor is larger than the languages of all other solutions.

Definition 2.12. *Consider the DSSP of Definition* 2.11. *A pair of supervisors* $(S_1^{\uparrow}, S_2^{\uparrow}) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ *is called an* optimal decentralized solution *if it is a solution, i.e.,*

$$S_1^{\uparrow} \wedge S_2^{\uparrow} \sqsubseteq S^{\uparrow}, \tag{1}$$

and for all pairs $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$

$$S_1 \wedge S_2 \sqsubseteq S^{\uparrow} \Rightarrow \mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(S_1^{\uparrow} \wedge S_2^{\uparrow}/G). \tag{2}$$

Recall from [14, 17] the definition of decomposability. A language $K \subseteq \mathrm{L}(G)$ is called *decomposable* if

$$K = \mathrm{p}_1^{-1}(\mathrm{p}_1(K)) \cap \mathrm{p}_2^{-1}(\mathrm{p}_2(K)) \cap \mathrm{L}(G). \tag{3}$$

Rudie showed that, under the given assumptions, there exists a decentralized solution, $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$, such that the language of the controlled system, $\mathrm{L}(S_1 \wedge S_2/G)$, is equal to a given language $K \subseteq \mathrm{L}(G)$ if and only if K is decomposable. The set of decomposable languages is not closed under arbitrary unions. It is therefore not guaranteed that this set contains a unique supremal element. This implies that in general the optimal decentralized solution does not exist. There may exist several, mutually incomparable, maximal solutions.

DEFINITION 2.13. *Consider the DSSP of Definition* 2.11. *A pair of supervisors* $(S_1^{\square}, S_2^{\square}) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ *is called a* maximal decentralized solution *if it is a solution, i.e.,*

$$S_1^{\square} \wedge S_2^{\square} \sqsubseteq S^{\uparrow},$$

and there does not exist a pair $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ *such that*

$$S_1 \wedge S_2 \sqsubseteq S^{\uparrow} \text{ and } \mathrm{L}(S_1^{\square} \wedge S_2^{\square}/G) \subsetneq \mathrm{L}(S_1 \wedge S_2/G).$$

The set of decomposable languages is closed under arbitrary intersections. It therefore contains a unique infimal element. Rudie posed the following control problem. Given lower bound $\mathrm{L}(A) \subseteq \Sigma^*$ and upper bound $\mathrm{L}(E) \subseteq \Sigma^*$, find a pair $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$, such that

$$\mathrm{L}(A) \subseteq \mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(E).$$

She showed there exists a solution to this control problem if and only if the infimal decomposable language containing $\mathrm{L}(A)$ is contained in $\mathrm{L}(E)$. Although this infimal is useful to solve the existence question, it often does not give a satisfactory solution. The following example shows that it is in general not trivial to define the lower bound $\mathrm{L}(A)$.

Example 2.14. Consider the alternating bit protocol [14, 17, 19]. This protocol achieves the reliable transmission of messages across an unreliable connection. To achieve this, the sender attaches to each message an extra bit containing either a zero or a one. The protocol can start with either a zero or a one attached to the first message. Consequently, the message with either a one or a zero attached is disabled initially. If the lower bound allows a zero attached to the first message, then the protocol cannot disable this message. It cannot choose the option where a one is attached to the first message. The lower bound $\mathrm{L}(A)$ should allow for both options. Therefore it cannot contain either of the options as this would exclude the other option. The only lower bound that allows both options is the empty language. Unfortunately the infimal decomposable language derived from the empty language does not give a satisfactory solution. See also [10, Sect. 2.5]. □

Another suggestion presented in [14, 17] was to look for the suboptimal solution characterized by the strong decomposability condition. A language $K \subseteq \mathrm{L}(G)$ is called *strongly decomposable* (with respect to Σ_1 and Σ_2) if

$$K = \left(\mathrm{p}_1^{-1}(\mathrm{p}_1(K)) \cup \mathrm{p}_2^{-1}(\mathrm{p}_2(K))\right) \cap \mathrm{L}(G). \tag{4}$$

This condition is closed under arbitrary unions. So the supremal strongly decomposable language exists. Recall from [6] the definition of normality. A language $K \subseteq \mathrm{L}(G)$ is called *normal* (with respect to $\Sigma_{\mathrm{o}} \subseteq \Sigma$) if

$$K = \mathrm{p}_{\mathrm{o}}^{-1}(\mathrm{p}_{\mathrm{o}}(K)) \cap \mathrm{L}(G). \tag{5}$$

Normality of a language K is a sufficient condition for the existence of supervisor that can observe events in Σ_{o} and that achieves K as language of the controlled system.

PROPOSITION 2.15. *If* $K \subseteq \mathrm{L}(G)$ *is strongly decomposable with respect to* Σ_1 *and* Σ_2*, then* K *is normal with respect to* Σ_1 *and normal with respect to* Σ_2.

Proof. The inclusion $K \subseteq \mathrm{p}_i^{-1}(\mathrm{p}_i(K)) \cap \mathrm{L}(G)$ is satisfied for all languages contained in $\mathrm{L}(G)$. So, it is sufficient to prove $K \supseteq \mathrm{p}_i^{-1}(\mathrm{p}_i(K)) \cap \mathrm{L}(G)$. By the definition of strong decomposability

$$\begin{aligned} K &= \big(\mathrm{p}_1^{-1}(\mathrm{p}_1(K)) \,\cup\, \mathrm{p}_2^{-1}(\mathrm{p}_2(K))\big) \,\cap\, \mathrm{L}(G) \\ &= \big(\mathrm{p}_1^{-1}(\mathrm{p}_1(K)) \,\cap\, \mathrm{L}(G)\big) \,\cup\, \big(\mathrm{p}_2^{-1}(\mathrm{p}_2(K)) \,\cap\, \mathrm{L}(G)\big) \\ &\supseteq \mathrm{p}_i^{-1}(\mathrm{p}_i(K)) \cap \mathrm{L}(G) \text{ for } i = 1, 2. \qquad \square \end{aligned}$$

The consequence of this proposition is that, if language K is strongly decomposable, then one supervisor, either $S_1 \in \mathcal{C}(\Sigma_1)$ or $S_2 \in \mathcal{C}(\Sigma_2)$, can obtain K as language of the controlled system. The other supervisor is not needed. Obviously, strong decomposability is too strong a restriction for decentralized control problems.

It can be concluded that the existing results for decentralized supervisory control problems do not satisfy the needs from control engineering.

In this paper a characterization of maximal solutions for decentralized control problems will be derived. Is it useful to look for maximal solutions? If a solution is maximal, then this does not imply that it is a good solution. For instance, a maximal solution may allow a lot of unimportant traces and disable all important ones. Another solution which allows less unimportant traces but more important ones may be considered a better solution. However, the authors believe there are some good reasons to investigate the characteristics of maximal solutions. The first and most important reason is that it gives us valuable insight into the fundamental properties of decentralized control problems. This insight may be used to derive algorithms that can synthesize "good" (in whatever sense) solutions, whether they are maximal or not.

Another reason why the authors believe maximality is important is that these "good" solutions will probably be maximal. So, although maximality of a solution does not imply that this solution is useful, a solution that is useful (good in some sense) will most likely be maximal. If a characterization of all maximal solutions can be given, then all "good" solutions will satisfy this characterization. So this characterization limits the class of solutions in which the good ones can be found.

Suppose a solution is given, but it is not fully satisfactory. One can ask the question whether the solution can be extended to obtain a better one. This is possible only if the given solution is not yet maximal. So also in this case a characterization of the maximal solutions will be useful.

2.4. Projections. In [4], Kozak and Wonham propose projections of the supremal supervisor as a solution to the decentralized control or synthesis problem.

DEFINITION 2.16. *The* projection of the supremal supervisor *to event set* $\Sigma_a \subseteq \Sigma(S^\uparrow)$ *is denoted by* $\mathrm{proj}(S^\uparrow, \Sigma_a)$. *It is defined for all* $s_a \in \mathrm{p_a}(\mathrm{L}(G))$ *by*

$$\begin{aligned} &\gamma(\mathrm{proj}(S^\uparrow, \Sigma_a), s_a) \\ &\quad = \left\{\sigma \in \Sigma_c \cap \Sigma_a : \exists s \in \mathrm{p}_a^{\text{-}1}(s_a) \cap \mathrm{L}(S^\uparrow/G) \textit{such that } \sigma \in \gamma(S^\uparrow, s)\right\}. \end{aligned}$$

PROPOSITION 2.17 ([4], Lem. 5.1).

$$\mathrm{proj}(S^\uparrow, \Sigma_1) \wedge \mathrm{proj}(S^\uparrow, \Sigma_2) \;\sqsubseteq\; S^\uparrow. \tag{6}$$

Kozak and Wonham call $\mathrm{proj}(S^\uparrow, \Sigma_1) \wedge \mathrm{proj}(S^\uparrow, \Sigma_2)$ the *fully decentralized solution.* In general the infimal decomposable solution of Rudie and the projected solution of Kozak and Wonham are incomparable. However, if the given lower bound, $\mathrm{L}(A)$, is

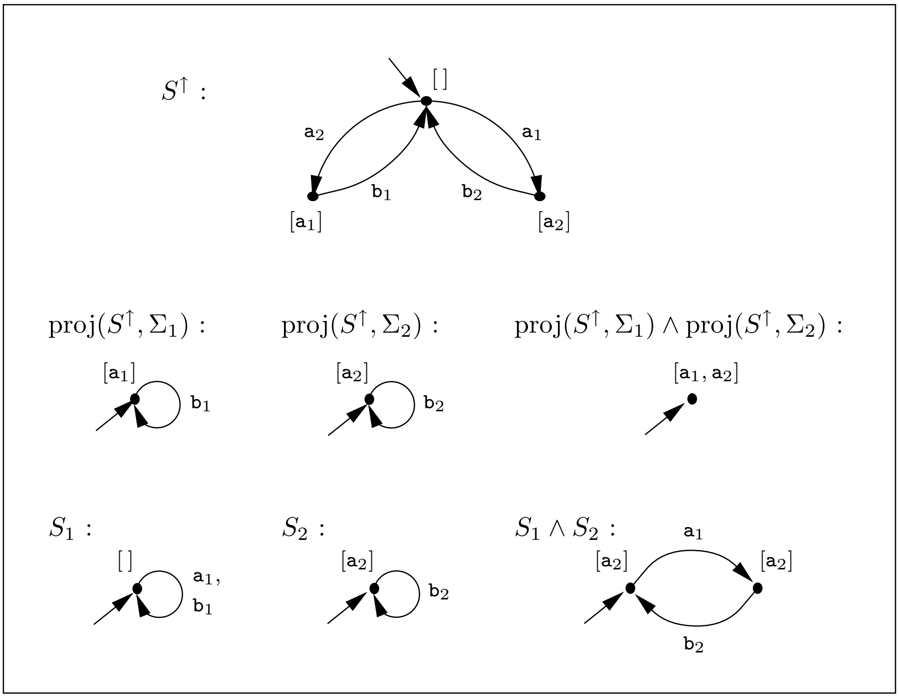

$S^{\uparrow}$:
[]
a2
a1
b1
b2
[a1]
[a2]
proj($S^{\uparrow}, \Sigma_1$) :
proj($S^{\uparrow}, \Sigma_2$) :
proj($S^{\uparrow}, \Sigma_1$) ∧ proj($S^{\uparrow}, \Sigma_2$) :
[a1]
b1
[a2]
b2
[a1, a2]
S_1 :
S_2 :
$S_1 \wedge S_2$:
[]
a1,
b1
[a2]
b2
a1
[a2]
[a2]
b2

FIG. 2. *The fully decentralized solution is in general not maximal.*

the empty trace, then the projected solution is larger than the infimal decomposable solution. But, even if $\Sigma_1 \cap \Sigma_2 = \emptyset$, the fully decentralized solution is in general not maximal. Consider the following example.

Example 2.18. Consider the supremal supervisor and the fully decentralized solution given in Figure 2. In this example $\Sigma_1 = \{\mathtt{a}_1, \mathtt{b}_1\}$, $\Sigma_2 = \{\mathtt{a}_2, \mathtt{b}_2\}$, and $\Sigma_c = \{\mathtt{a}_1, \mathtt{a}_2\}$. The pair $(\mathrm{proj}(S^{\uparrow}, \Sigma_1), \mathrm{proj}(S^{\uparrow}, \Sigma_2))$ is not maximal, because the pair (S_1, S_2) results in a strictly larger controlled language.

Supervisor $\mathrm{proj}(S^{\uparrow}, \Sigma_1)$ disables event $\mathtt{a}_1$ because the uncontrolled system can execute event $\mathtt{a}_2$, after which event $\mathtt{a}_1$ must be disabled. However, as supervisor S_2 disables $\mathtt{a}_2$ it is not necessary for supervisor S_1 to disable $\mathtt{a}_1$. The pair of supervisors obtained by projection from the supremal supervisor is in general not maximal because the supervisors only take into account the control actions of the supremal supervisor. They do not consider the control law of the other supervisor. In order to obtain a maximal solution it is necessary that the supervisors take into account the control law of the other supervisor. So, to synthesize supervisor S_1 one should already know the control law of supervisor S_2, and to synthesize S_2 one should already know the control law of supervisor S_1. It is this cyclic dependency that makes the synthesis of decentralized controllers such a hard problem.

3. Nash equilibria and maximal solutions.

Decentralized stochastic control has been studied extensively. It is related to game and team theory (see [1, 3, 8, 12]). In these fields of research a so-called cost function is used. This cost function maps a decentralized control law to a real number. A solution is considered optimal if it has the lowest cost. Using cost functions, all solutions can be compared with each other. In the field of decentralized supervisory control, solutions are compared by the

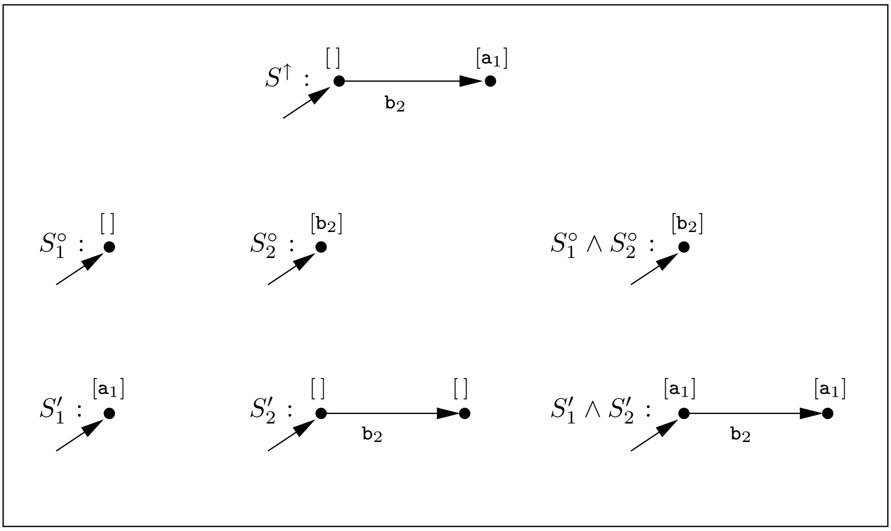


FIG. 3. *The pair* $(S_1^{\circ}, S_2^{\circ})$ *is a Nash equilibrium, yet it is not maximal.*

language of the controlled system. This ordering is not complete. Some solutions may not be comparable.

In game and team theory the notion of Nash equilibrium plays an important role. It will be shown that Nash equilibria are also important for decentralized supervisory control. A pair of supervisors forms a Nash equilibrium if a supervisor cannot improve the controlled language when the other supervisor is kept fixed and conversely.

DEFINITION 3.1. *Consider the DSSP of Definition* 2.11. *A pair of supervisors* $(S_1^{\circ}, S_2^{\circ}) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ *is called a* Nash equilibrium *if it is a solution, i.e.,*

$$S_1^{\circ} \wedge S_2^{\circ} \sqsubseteq S^{\uparrow},$$

and

$$\begin{aligned} &\forall S_2 \in \mathcal{C}(\Sigma_2) S_1^{\circ} \wedge S_2 \sqsubseteq S^{\uparrow} \Rightarrow \ \mathrm{L}(S_1^{\circ} \wedge S_2/G) \subseteq \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G), \ \textit{and} \\ &\forall S_1 \in \mathcal{C}(\Sigma_1) S_1 \wedge S_2^{\circ} \sqsubseteq S^{\uparrow} \Rightarrow \ \mathrm{L}(S_1 \wedge S_2^{\circ}/G) \subseteq \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G). \end{aligned}$$

In game theory, controllers have conflicting optimization criteria, whereas in team theory all controllers try to optimize the same cost criterion. Note that in the above definition the closed-loop language is analogous to the cost function in a team or game problem. The notion of Nash equilibrium has been introduced in game theory. In team theory it is also known as a person-by-person optimal solution.

In team theory, under certain convexity conditions, a set of controllers is maximal if and only if it is a Nash equilibrium [12]. This equivalence is quite useful because it is relatively easier to determine a Nash equilibrium than a maximum.

The following example shows that for discrete-event systems the Nash equilibrium condition is not sufficient to guarantee maximality.

Example 3.2. Consider the supremal supervisor $S^{\uparrow}$ and the decentralized implementation $(S_1^{\circ}, S_2^{\circ})$ given in Figure 3. $\Sigma_1 = \{\mathtt{a}_1\}$, $\Sigma_2 = \{\mathtt{b}_2\}$. All events are controllable. It is not difficult to check that the pair $(S_1^{\circ}, S_2^{\circ})$ is a Nash equilibrium. However, it is not maximal, because the pair (S_1', S_2') is a solution with a strictly larger controlled language. □

For discrete-event systems we need the stronger condition of a strong Nash equilibrium to guarantee maximality of a pair of supervisors.

DEFINITION 3.3. *Consider the DSSP of Definition* 2.11. *A pair of supervisors* $(S_1^\circ, S_2^\circ) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ *is called a* strong Nash equilibrium *if it is a Nash equilibrium, and for all* $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$,

$$\mathrm{L}(S_1 \wedge S_2/G) = \mathrm{L}(S_1^\circ \wedge S_2^\circ/G) \Rightarrow (S_1, S_2) \text{ is a Nash equilibrium.} \tag{7}$$

An intuitive interpretation of the need for the concept of a strong Nash equilibrium follows. The aim of the paper is to obtain a characterization of a maximal decentralized solution in terms of a person-by-person characterization as in game and team theory. Example 3.2 shows that there exists a pair of supervisors that is a Nash equilibrium but not a maximal solution. The condition for a pair of supervisors (S_1°, S_2°) to be a Nash equilibrium is phrased solely in terms of the closed-loop language $L(S_1^\circ \wedge S_2^\circ/G)$. Because of this formulation, it appears that it is necessary that any pair of languages that achieves the same closed-loop language is also a Nash equilibrium. If such a pair was not a Nash equilibrium, one would be able to construct a pair of supervisors with a strictly larger closed-loop language. This in turn would contradict maximality. The next theorem shows that the concept of a strong Nash equilibrium is appropriate.

By Proposition 2.7, $\mathrm{L}(S_1 \wedge S_2/G) = \mathrm{L}(S_1^\circ \wedge S_2^\circ/G)$ and $S_1^\circ \wedge S_2^\circ \sqsubseteq S^\uparrow$ together imply that $S_1 \wedge S_2 \sqsubseteq S^\uparrow$.

THEOREM 3.4. *A pair of supervisors* $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ *is maximal if and only if it is a strong Nash equilibrium.*

Proof (strong Nash ⇒ Maximal). Assume $(S_1^\circ, S_2^\circ) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ is strong Nash but not maximal. Then there exists a pair $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ such that $S_1 \wedge S_2 \sqsubseteq S^\uparrow$ and $\mathrm{L}(S_1^\circ \wedge S_2^\circ/G) \subsetneqq \mathrm{L}(S_1 \wedge S_2/G)$. Define $S_1^\square \in \mathcal{C}(\Sigma_1)$ by

$$\gamma(S_1^\square, s_1) = \gamma(S_1^\circ, s_1) \cup \gamma(S_1, s_1) \quad \forall s_1 \in \mathrm{p}_1(\mathrm{L}(G)).$$

We will prove the following points.

1. $\mathrm{L}(S_1^\square/G) = \mathrm{L}(S_1^\circ/G) \cap \mathrm{L}(S_1/G)$,
2. $S_1^\square \wedge S_2^\circ \sqsubseteq S^\uparrow$,
3. $S_1^\square \wedge S_2 \sqsubseteq S^\uparrow$,
4. $\mathrm{L}(S_1^\circ \wedge S_2^\circ/G) = \mathrm{L}(S_1^\square \wedge S_2^\circ/G)$,
5. $\mathrm{L}(S_1^\square \wedge S_2^\circ/G) \subseteq \mathrm{L}(S_1^\square \wedge S_2/G)$.

(Point 1.) This point will be proven by complete induction. The initial step follows from $\varepsilon \in \mathrm{L}(S_1^\square/G)$ and $\varepsilon \in \mathrm{L}(S_1^\circ/G) \cap \mathrm{L}(S_1/G)$. For the inductive step let $s \in \mathrm{L}(S_1^\square/G)$ and $s \in \mathrm{L}(S_1^\circ/G) \cap \mathrm{L}(S_1/G)$. Then

$$\begin{aligned} s\sigma \in \mathrm{L}(S_1^\square/G) &\iff s\sigma \in \mathrm{L}(G),\ \sigma \notin \gamma(S_1^\square, \mathrm{p}_1(s)) \\ &\iff s\sigma \in \mathrm{L}(G),\ \sigma \notin \gamma(S_1^\circ, \mathrm{p}_1(s)),\ \sigma \notin \gamma(S_1, \mathrm{p}_1(s)) \\ &\iff s\sigma \in \mathrm{L}(S_1^\circ/G),\ s\sigma \in \mathrm{L}(S_1/G) \\ &\iff s\sigma \in \mathrm{L}(S_1^\circ/G) \cap \mathrm{L}(S_1/G). \end{aligned}$$

It follows that $\mathrm{L}(S_1^\square/G) = \mathrm{L}(S_1^\circ/G) \cap \mathrm{L}(S_1/G)$.

(Points 2 and 4.) From point 1 and Proposition 2.10, it follows that

$$\begin{aligned}\mathrm{L}(S_1^{\Box} \wedge S_2^{\circ}/G) &= \mathrm{L}(S_1^{\circ}/G) \cap \mathrm{L}(S_1/G) \cap \mathrm{L}(S_2^{\circ}/G)\\ &= \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G) \cap \mathrm{L}(S_1/G)\\ &= \big[\text{because } \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G) \subseteq \mathrm{L}(S_1 \wedge S_2/G) \text{ and}\\ &\quad\ \text{by Proposition 2.10 } \mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(S_1/G)\big]\\ &\quad\ \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G).\end{aligned}$$

This proves point 4. Point 2 follows from $S_1^{\circ} \wedge S_2^{\circ} \sqsubseteq S^{\uparrow}$ and Proposition 2.7.

(Point 3.) From point 1 and Proposition 2.10, it follows that

$$\begin{aligned}\mathrm{L}(S_1^{\Box} \wedge S_2/G) &= \mathrm{L}(S_1^{\circ}/G) \cap \mathrm{L}(S_1/G) \cap \mathrm{L}(S_2/G)\\ &\subseteq \mathrm{L}(S_1/G) \cap \mathrm{L}(S_2/G) = \mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(S^{\uparrow}/G).\end{aligned}$$

So, by Proposition 2.7, $S_1^{\Box} \wedge S_2 \sqsubseteq S^{\uparrow}$.

(Point 5.) From point 4, it follows that

$$\begin{aligned}\mathrm{L}(S_1^{\Box} \wedge S_2^{\circ}/G) &= \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G)\\ &= \big[\text{because } \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G) \subseteq \mathrm{L}(S_1 \wedge S_2/G)\big]\\ &\quad\ \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G) \cap \mathrm{L}(S_1 \wedge S_2/G)\\ &= \mathrm{L}(S_1^{\circ}/G) \cap \mathrm{L}(S_2^{\circ}/G) \cap \mathrm{L}(S_1/G) \cap \mathrm{L}(S_2/G)\\ &\subseteq \mathrm{L}(S_1^{\Box}/G) \cap \mathrm{L}(S_2/G) = \mathrm{L}(S_1^{\Box} \wedge S_2/G).\end{aligned}$$

As the pair $(S_1^{\circ}, S_2^{\circ})$ is strong Nash, it follows from point 4 that $(S_1^{\Box}, S_2^{\circ})$ is Nash. So, by point 3 and the definition of Nash, $\mathrm{L}(S_1^{\Box} \wedge S_2/G) \subseteq \mathrm{L}(S_1^{\Box} \wedge S_2^{\circ}/G)$. Then, from points 4 and 5, $\mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G) = \mathrm{L}(S_1^{\Box} \wedge S_2^{\circ}/G) = \mathrm{L}(S_1^{\Box} \wedge S_2/G)$. As $(S_1^{\circ}, S_2^{\circ})$ is strong Nash, the pair $(S_1^{\Box}, S_2)$ is Nash. So

$$\mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(S_1^{\Box} \wedge S_2/G) = \mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}/G).$$

But this contradicts our assumption that $\mathrm{L}(S_1^{\circ} \wedge S_2^{\circ}) \subsetneqq \mathrm{L}(S_1 \wedge S_2/G)$. We can conclude that if $(S_1^{\circ}, S_2^{\circ})$ is strong Nash, then it is maximal.

(Maximal ⇒ Strong Nash.) Assume $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ is maximal but not strong Nash. Then there exists a pair $(S_1', S_2') \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ such that $\mathrm{L}(S_1' \wedge S_2'/G) = \mathrm{L}(S_1 \wedge S_2/G)$ and (S_1', S_2') is not Nash. So

$$\exists S_1'' \in \mathcal{C}(\Sigma_1) \,\text{such that}\, S_1'' \wedge S_2' \sqsubseteq S^{\uparrow}\text{and } \mathrm{L}(S_1'' \wedge S_2'/G) \not\subseteq \mathrm{L}(S_1' \wedge S_2'/G)$$

or

$$\exists S_2'' \in \mathcal{C}(\Sigma_2) \text{ such that } S_1' \wedge S_2'' \sqsubseteq S^{\uparrow} \text{ and } \mathrm{L}(S_1' \wedge S_2''/G) \not\subseteq \mathrm{L}(S_1' \wedge S_2'/G).$$

Assume, without loss of generality, that such an S_2'' exists. Let $S_2^{\Box} \in \mathcal{C}(\Sigma_2)$ be defined by

$$\gamma(S_2^{\Box}, s_2) = \begin{cases} \gamma(S_2', s_2) \cap \gamma(S_2'', s_2), & \text{if } s_2 \in \mathrm{p}_2(\mathrm{L}(S_2'/G)) \text{ and } s_2 \in \mathrm{p}_2(\mathrm{L}(S_2''/G)),\\ \gamma(S_2', s_2), & \text{if } s_2 \in \mathrm{p}_2(\mathrm{L}(S_2'/G)) \text{ and } s_2 \notin \mathrm{p}_2(\mathrm{L}(S_2''/G)),\\ \gamma(S_2'', s_2), & \text{if } s_2 \notin \mathrm{p}_2(\mathrm{L}(S_2'/G)) \text{ and } s_2 \in \mathrm{p}_2(\mathrm{L}(S_2''/G)),\\ \Sigma_{2,\mathrm{c}}, & \text{otherwise.}\end{cases}$$

We will prove the following points.

1. $\mathrm{L}(S_2^{\square}/G) = \mathrm{L}(S_2'/G) \cup \mathrm{L}(S_2''/G)$,
2. $S_1' \wedge S_2^{\square} \sqsubseteq S^{\uparrow}$,
3. $\mathrm{L}(S_1' \wedge S_2'/G) \subseteq \mathrm{L}(S_1' \wedge S_2^{\square}/G)$,
4. $\mathrm{L}(S_1' \wedge S_2''/G) \subseteq \mathrm{L}(S_1' \wedge S_2^{\square}/G)$.

(Point 1.) This point will be proven by complete induction. The initial step follows from $\varepsilon \in \mathrm{L}(S_2^{\square}/G)$ and $\varepsilon \in \mathrm{L}(S_2'/G) \cup \mathrm{L}(S_2''/G)$. For the inductive step let $s \in \mathrm{L}(S_2^{\square}/G)$ and $s \in \mathrm{L}(S_2'/G) \cup \mathrm{L}(S_2''/G)$. Trace s can be in one of the three sets $\mathrm{L}(S_2'/G) \cap \mathrm{L}(S_2''/G)$, $\mathrm{L}(S_2'/G) - \mathrm{L}(S_2''/G)$, or $\mathrm{L}(S_2''/G) - \mathrm{L}(S_2'/G)$. If $s \in \mathrm{L}(S_2'/G) \cap \mathrm{L}(S_2''/G)$, then

$$\begin{aligned} s\sigma \in \mathrm{L}(S_2^{\square}/G) &\iff s\sigma \in \mathrm{L}(G) \wedge \sigma \notin \gamma(S_2^{\square}, \mathrm{p}_2(s)) \\ &\iff s\sigma \in \mathrm{L}(G) \ \wedge \ \big(\sigma \notin \gamma(S_2', \mathrm{p}_2(s)) \vee \sigma \notin \gamma(S_2'', \mathrm{p}_2(s))\big) \\ &\iff s\sigma \in \mathrm{L}(S_2'/G) \vee s\sigma \in \mathrm{L}(S_2''/G) \\ &\iff s\sigma \in \mathrm{L}(S_2'/G) \cup \mathrm{L}(S_2''/G). \end{aligned}$$

If $s \in \mathrm{L}(S_2'/G)$ but $s \notin \mathrm{L}(S_2''/G)$, then

$$\begin{aligned} s\sigma \in \mathrm{L}(S_2^{\square}/G) &\iff s\sigma \in \mathrm{L}(G) \wedge \sigma \notin \gamma(S_2^{\square}, \mathrm{p}_2(s)) \\ &\iff s\sigma \in \mathrm{L}(G) \wedge \sigma \notin \gamma(S_2', \mathrm{p}_2(s)) \\ &\iff s\sigma \in \mathrm{L}(S_2'/G) \\ &\iff s\sigma \in \mathrm{L}(S_2'/G) \cup \mathrm{L}(S_2''/G). \end{aligned}$$

A similar reasoning holds if $s \in \mathrm{L}(S_2''/G)$ but $s \notin \mathrm{L}(S_2'/G)$. Hence, it follows that $\mathrm{L}(S_2^{\square}/G) = \mathrm{L}(S_2'/G) \cup \mathrm{L}(S_2''/G)$.

(Points 2, 3, and 4.) From point 1 and Proposition 2.10, it follows that

$$\begin{aligned} \mathrm{L}(S_1' \wedge S_2^{\square}/G) &= \mathrm{L}(S_1'/G) \ \cap \ \big(\mathrm{L}(S_2'/G) \cup \mathrm{L}(S_2''/G)\big) \\ &= \big(\mathrm{L}(S_1'/G) \cap \mathrm{L}(S_2'/G)\big) \ \cup \ \big(\mathrm{L}(S_1'/G) \cap \mathrm{L}(S_2''/G)\big) \\ &= \mathrm{L}(S_1' \wedge S_2'/G) \cup \mathrm{L}(S_1' \wedge S_2''/G). \end{aligned}$$

This directly proves points 3 and 4. Point 2 follows from $S_1' \wedge S_2' \sqsubseteq S^{\uparrow}$, $S_1' \wedge S_2'' \sqsubseteq S^{\uparrow}$, and Proposition 2.7.

As (S_1, S_2) is maximal, so is (S_1', S_2'). Then, by point 3, $\mathrm{L}(S_1' \wedge S_2'/G) = \mathrm{L}(S_1' \wedge S_2^{\square}/G)$. From point 4 it follows that $\mathrm{L}(S_1' \wedge S_2''/G) \subseteq \mathrm{L}(S_1' \wedge S_2'/G)$. But this contradicts our assumption that $\mathrm{L}(S_1' \wedge S_2''/G) \nsubseteq \mathrm{L}(S_1' \wedge S_2'/G)$. Hence it can be concluded that if (S_1, S_2) is maximal, then it is strong Nash. $\square$

Consider two pairs of supervisors to be *control equivalent* if their controlled languages are equal, or

$$(S_1, S_2) \equiv (S_3, S_4) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2) \text{ if } L(S_1 \wedge S_2/G) = L(S_3 \wedge S_4/G).$$

Then a pair of supervisors is maximal if and only if all control equivalent pairs are Nash equilibria. Let the *control equivalence class* corresponding with the language $K \subseteq \mathrm{L}(G)$ be the set of pairs for which the controlled language is equal to K. A prefix closed and decomposable language can be considered maximal if and only if all pairs in its corresponding control equivalence class are Nash equilibria.

If the event sets Σ_1 and Σ_2 are disjoint, then a weaker condition can be found to characterize maximal solutions. Define $(\widehat{S}_1, \widehat{S}_2)$ as the pair of most restrictive supervisors in the control equivalence class of (S_1, S_2).

Definition 3.5. *Let $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$. The supervisor $\widehat{S}_1 \in \mathcal{C}(\Sigma_1)$ is defined by*

$$\gamma(\widehat{S}_1, s_1) = \{\sigma \in \Sigma_c(S_1) : s_1\sigma \notin \mathrm{p}_1(\mathrm{L}(S_1 \wedge S_2/G))\} \quad \forall s_1 \in \mathrm{p}_1(\mathrm{L}(G)).$$

The supervisor $\widehat{S}_2 \in \mathcal{C}(\Sigma_2)$ is defined analogously.

Supervisor $\widehat{S}_1$ can be seen as the most restrictive supervisor of all supervisors S_1' for which there exists a supervisor S_2' such that $\mathrm{L}(S_1' \wedge S_2'/G) = \mathrm{L}(S_1 \wedge S_2/G)$. That is, if such an S_1' disables event σ after trace s, then $s\sigma$ is not an element of $\mathrm{L}(S_1'/G) \supseteq \mathrm{L}(S_1' \wedge S_2'/G) = \mathrm{L}(S_1 \wedge S_2/G)$. So $\widehat{S}_1$ will also disable this event.

First it needs to be proven that $(\widehat{S}_1, \widehat{S}_2)$ is a solution and that it is control equivalent with (S_1, S_2).

Proposition 3.6. *Let $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ be a decentralized solution implementing $S^\uparrow$, and let $(\widehat{S}_1, \widehat{S}_2)$ be defined as above. Then*

1. *$\widehat{S}_1 \wedge \widehat{S}_2 \sqsubseteq S^\uparrow$, and*
2. *$\mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G) = \mathrm{L}(S_1 \wedge S_2/G)$.*

Proof (point 2, $\mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G) \subseteq \mathrm{L}(S_1 \wedge S_2/G)$). First we will prove by induction that $\mathrm{L}(\widehat{S}_1/G) \subseteq \mathrm{L}(S_1/G)$. The initial step follows from $\varepsilon \in \mathrm{L}(\widehat{S}_1/G)$ and $\varepsilon \in \mathrm{L}(S_1/G)$. For the inductive step let $s \in \mathrm{L}(\widehat{S}_1/G)$ and $s \in \mathrm{L}(S_1/G)$.

$$\begin{aligned}
s\sigma \in \mathrm{L}(\widehat{S}_1/G) &\Rightarrow s\sigma \in \mathrm{L}(G) \ \wedge \ \sigma \notin \gamma(\widehat{S}_1, \mathrm{p}_1(s)) \\
&\Rightarrow s\sigma \in \mathrm{L}(G) \ \wedge \ \big(\sigma \notin \Sigma_\mathrm{c}(S_1) \vee \mathrm{p}_1(s)\sigma \in \mathrm{p}_1(\mathrm{L}(S_1 \wedge S_2/G))\big) \\
&\Rightarrow \big[\text{because } \mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(S_1/G)\big] \\
&\qquad s\sigma \in \mathrm{L}(G) \ \wedge \ \big(\sigma \notin \Sigma_\mathrm{c}(S_1) \vee \mathrm{p}_1(s)\sigma \in \mathrm{p}_1(\mathrm{L}(S_1/G))\big) \\
&\Rightarrow s\sigma \in \mathrm{L}(G) \ \wedge \ \big(\sigma \notin \Sigma_\mathrm{c}(S_1) \vee \sigma \notin \gamma(S_1, \mathrm{p}_1(s))\big) \\
&\Rightarrow \big[\text{because } \sigma \notin \Sigma_\mathrm{c}(S_1) \Rightarrow \sigma \notin \gamma(S_1, \mathrm{p}_1(s))\big] \\
&\qquad s\sigma \in \mathrm{L}(G) \ \wedge \ \sigma \notin \gamma(S_1, \mathrm{p}_1(s)) \\
&\Rightarrow s\sigma \in \mathrm{L}(S_1/G).
\end{aligned}$$

By symmetry it follows that $\mathrm{L}(\widehat{S}_2/G) \subseteq \mathrm{L}(S_2/G)$. So

$$\mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G) = \mathrm{L}(\widehat{S}_1/G) \cap \mathrm{L}(\widehat{S}_2/G) \subseteq \mathrm{L}(S_1/G) \cap \mathrm{L}(S_2/G) = \mathrm{L}(S_1 \wedge S_2/G).$$

(Point 2, $\mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G)$.) First it will be proven by induction that $\mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(\widehat{S}_1/G)$. The initial step follows from $\varepsilon \in \mathrm{L}(S_1 \wedge S_2/G)$ and $\varepsilon \in \mathrm{L}(\widehat{S}_1/G)$. For the inductive step let $s \in \mathrm{L}(S_1 \wedge S_2/G)$ and $s \in \mathrm{L}(\widehat{S}_1/G)$.

$$\begin{aligned}
&s\sigma \in \mathrm{L}(S_1 \wedge S_2/G) \\
&\Rightarrow \sigma \notin \Sigma_1 \ \vee \ \big(\sigma \in \Sigma_1 \wedge \mathrm{p}_1(s\sigma) = \mathrm{p}_1(s)\sigma \in \mathrm{p}_1(\mathrm{L}(S_1 \wedge S_2/G))\big) \\
&\Rightarrow \big[\text{by construction of } \gamma(\widehat{S}_1, \mathrm{p}_1(s))\big] \\
&\qquad \sigma \notin \Sigma_1 \ \vee \ (\sigma \in \Sigma_1 \wedge \sigma \notin \gamma(\widehat{S}_1, \mathrm{p}_1(s))) \\
&\Rightarrow \big[\text{because } \gamma(\widehat{S}_1, \mathrm{p}_1(s)) \subseteq \Sigma_1\big] \ \ \sigma \notin \gamma(\widehat{S}_1, \mathrm{p}_1(s)) \\
&\Rightarrow \big[\text{because } s \in \mathrm{L}(\widehat{S}_1/G) \text{ and } s\sigma \in \mathrm{L}(G)\big] \ \ s\sigma \in \mathrm{L}(\widehat{S}_1/G).
\end{aligned}$$

By symmetry it follows that $\mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(\widehat{S}_2/G)$. So

$$\mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(\widehat{S}_1/G) \cap \mathrm{L}(\widehat{S}_2/G) = \mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G).$$

(Point 1.) This follows directly from point 2 and Proposition 2.7. □

THEOREM 3.7. *Let* $\Sigma_1 \cap \Sigma_2 = \emptyset$. *Let* $(S_1, S_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$. *Let* $(\widehat{S}_1, \widehat{S}_2) \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ *be defined by Definition* 3.5. *Then* (S_1, S_2) *is maximal if and only if* $(\widehat{S}_1, \widehat{S}_2)$ *is a Nash equilibrium.*

Proof (Maximal $\Rightarrow$ Nash). If (S_1, S_2) is maximal, then by Theorem 3.4 (S_1, S_2) is a strong Nash equilibrium, which by points 1 and 2 of Proposition 3.6 implies that $(\widehat{S}_1, \widehat{S}_2)$ is a Nash equilibrium.

(Nash $\Rightarrow$ Maximal.) Assume $(\widehat{S}_1, \widehat{S}_2)$ is a Nash equilibrium, but (S_1, S_2) is not maximal. Then, by point 2 of Proposition 3.6 $(\widehat{S}_1, \widehat{S}_2)$ is not maximal. There exists a pair $(S_1', S_2') \in \mathcal{C}(\Sigma_1) \times \mathcal{C}(\Sigma_2)$ such that $S_1' \wedge S_2' \sqsubseteq S^{\uparrow}$ and $\mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G) \subsetneqq \mathrm{L}(S_1' \wedge S_2'/G)$. We will first prove that

$$\widehat{S}_1 \wedge S_2' \sqsubseteq S^{\uparrow} \text{ and } S_1' \wedge \widehat{S}_2 \sqsubseteq S^{\uparrow}.$$

It will be proven by induction that $\mathrm{L}(\widehat{S}_1/G) \subseteq \mathrm{L}(S_1'/G)$. The initial step follows from $\varepsilon \in \mathrm{L}(\widehat{S}_1/G)$ and $\varepsilon \in \mathrm{L}(S_1'/G)$. For the inductive step let $s \in \mathrm{L}(\widehat{S}_1/G)$ and $s \in \mathrm{L}(S_1'/G)$.

$$\begin{aligned}
s\sigma \in \mathrm{L}(\widehat{S}_1/G) &\Rightarrow s\sigma \in \mathrm{L}(G) \ \wedge \ \sigma \notin \gamma(\widehat{S}_1, \mathrm{p}_1(s)) \\
&\Rightarrow s\sigma \in \mathrm{L}(G) \ \wedge \ (\sigma \notin \Sigma_{\mathrm{c}}(S_1) \vee \mathrm{p}_1(s)\sigma \in \mathrm{p}_1(\mathrm{L}(S_1 \wedge S_2/G))) \\
&\Rightarrow [\text{because } \mathrm{L}(S_1 \wedge S_2/G) \subseteq \mathrm{L}(S_1' \wedge S_2'/G) \subseteq \mathrm{L}(S_1'/G)] \\
&\quad\ s\sigma \in \mathrm{L}(G) \ \wedge \ (\sigma \notin \Sigma_{\mathrm{c}}(S_1) \vee \mathrm{p}_1(s)\sigma \in \mathrm{p}_1(\mathrm{L}(S_1'/G))) \\
&\Rightarrow s\sigma \in \mathrm{L}(G) \ \wedge \ (\sigma \notin \Sigma_{\mathrm{c}}(S_1) \vee \sigma \notin \gamma(S_1', \mathrm{p}_1(s))) \\
&\Rightarrow [\text{because } \sigma \notin \Sigma_{\mathrm{c}}(S_1) \Rightarrow \sigma \notin \gamma(S_1', \mathrm{p}_1(s))] \\
&\quad\ s\sigma \in \mathrm{L}(G) \ \wedge \ \sigma \notin \gamma(S_1', \mathrm{p}_1(s)) \\
&\Rightarrow s\sigma \in \mathrm{L}(S_1'/G).
\end{aligned}$$

It follows that $\mathrm{L}(\widehat{S}_1/G) \subseteq \mathrm{L}(S_1'/G)$. Now

$$\begin{aligned}
\mathrm{L}(\widehat{S}_1 \wedge S_2'/G) &= \mathrm{L}(\widehat{S}_1/G) \cap \mathrm{L}(S_2'/G) \subseteq \mathrm{L}(S_1'/G) \cap \mathrm{L}(S_2'/G) \\
&= \mathrm{L}(S_1' \wedge S_2'/G) \subseteq \mathrm{L}(S^{\uparrow}/G).
\end{aligned}$$

So, by Proposition 2.7, $\widehat{S}_1 \wedge S_2' \sqsubseteq S^{\uparrow}$. It follows by symmetry that $S_1' \wedge \widehat{S}_2 \sqsubseteq S^{\uparrow}$.

As $\mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G) \subsetneqq \mathrm{L}(S_1' \wedge S_2'/G)$ there exists a trace $s \in \mathrm{L}(S_1' \wedge S_2'/G)$ such that $s \notin \mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G)$. Let $v\sigma$ be the prefix of s such that $\sigma \in \Sigma$, $v \in \mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G)$, and $v\sigma \notin \mathrm{L}(\widehat{S}_1 \wedge \widehat{S}_2/G)$. Assume without loss of generality that $\sigma \in \Sigma_2$. Then, by the assumption that $\Sigma_1 \cap \Sigma_2 = \emptyset$, $\sigma \notin \Sigma_1$. So $\sigma \notin \gamma(\widehat{S}_1, \mathrm{p}_1(v)) \subseteq \Sigma_1$. Thus $v\sigma \in \mathrm{L}(\widehat{S}_1/G)$. As $v\sigma \in \mathrm{L}(S_1' \wedge S_2'/G) \subseteq \mathrm{L}(S_2'/G)$, it follows that $v\sigma \in \mathrm{L}(\widehat{S}_1 \wedge S_2'/G)$. But this contradicts the fact that $(\widehat{S}_1, \widehat{S}_2)$ is a Nash equilibrium. Hence we can conclude that if $(\widehat{S}_1, \widehat{S}_2)$ is a Nash equilibrium, then (S_1, S_2) is maximal. □

A prefix closed and decomposable language $K \subseteq \mathrm{L}(G)$ can be considered maximal if and only if the pair of most restricting supervisors in the control equivalence class corresponding with language K is a Nash equilibrium.

4. Construction of Nash equilibria. Theorems 3.4 and 3.7 give characterizations of the maximal solutions in terms of Nash equilibria. However, they do not state how Nash equilibria can be obtained. For dynamic games in the field of game and team theory, a necessary condition for a Nash equilibrium can be given by the coupled Bellman–Hamilton–Jacobi equations. A solution to these equations is under certain additional conditions also sufficient for a Nash equilibrium. A procedure for the construction of a solution is known [9]. It alternately keeps one of the controllers fixed and tries to optimize the other. At each iteration only one of the controllers is optimized. For dynamic games it is not guaranteed that the procedure converges. And if it converges, it is not guaranteed that it does so in a finite number of steps.

For supervisory control the Bellman–Hamilton–Jacobi equations are not applicable. Yet, the procedure can still be used. At each iteration one of the supervisors is kept fixed and the other is optimized. Only one supervisor is synthesized in each step. This can be seen as a supervisory control problem for a single supervisor. The combination of the fixed supervisor and the uncontrolled system is taken as the uncontrolled system for this control problem. As only one supervisor is synthesized (and all controllable events are observable) a unique optimal solution exists. In the next iteration this optimal supervisor is taken fixed and the other supervisor is optimized. This procedure is repeated until the pair of supervisors remains invariant. Below this procedure is formalized.

Assume without loss of generality that S_1 is the supervisor which is kept fixed. Consider the supervisory control problem with partial observations for the plant S_1/G and with legal language $L(S^{\uparrow})$. Then define

$$K = \sup \left\{ \begin{array}{l} K' \subseteq L(S_1/G) | (1)\ K' \subseteq L(S^{\uparrow}), \\ (2)\ K' \text{ controllable with respect to } (L(S_1/G), \Sigma_c(\Sigma_2)), \\ (3) \text{ and normal with respect to } L(S_1/G) \text{ and } p_2 \end{array} \right\}. \tag{8}$$

The supremal supervisor S_2 with respect to the uncontrolled system S_1/G is defined by

$$\gamma(S_2, s_2) = \{\sigma \in \Sigma_{\rm c}(S_2) : s_2\sigma \notin {\rm p}_2(K)\} \quad \forall s_2 \in {\rm p}_2({\rm L}(G)). \tag{9}$$

If S_2 is kept fixed, then S_1 is computed analogously. The formula for K in this case is obtained from that of (8) by interchanging the indices 1 and 2.

LEMMA 4.1. *Let S_1 be the supervisor which is kept fixed. Let S_2 and K be as defined above. Then* ${\rm L}(S_1 \wedge S_2/G) = K$.

Proof. The proof will be by complete induction. As $\varepsilon \in {\rm L}(S_1 \wedge S_2/G)$ and $\varepsilon \in K$ the initial step is satisfied. For the inductive step let $s \in {\rm L}(S_1 \wedge S_2/G)$ and $s \in K$.

$$\begin{aligned} s\sigma \in {\rm L}(S_1 \wedge S_2/G) &\iff s\sigma \in {\rm L}(S_1/G) \wedge \sigma \notin \gamma(S_2, {\rm p}_2(s)) \\ &\iff s\sigma \in {\rm L}(S_1/G) \ \wedge\ (\sigma \notin \Sigma_{\rm c}(\Sigma_2) \vee {\rm p}_2(s)\sigma \in {\rm p}_2(K)) \\ &\iff [\text{because } s \in K \text{ and } K \text{ is controllable}] \\ &\qquad s\sigma \in {\rm L}(S_1/G) \ \wedge\ (s\sigma \in K \vee {\rm p}_2(s)\sigma \in {\rm p}_2(K)) \\ &\iff [\text{because } K \subseteq {\rm p}_2^{-1}({\rm p}_2(K))] \\ &\qquad s\sigma \in {\rm L}(S_1/G) \wedge s\sigma \in {\rm p}_2^{-1}({\rm p}_2(K)) \\ &\iff \big[\text{because } K \text{ is normal with respect to} {\rm L}(S_1/G)\big] \\ &\qquad s\sigma \in K. \qquad \square \end{aligned}$$

The procedure is described by the following steps.

PROCEDURE 4.2.

1. *Choose a pair of most restrictive supervisors* (S_1^0, S_2^0) *as starting point of the procedure. Take, for instance, the pair of most restrictive supervisors corresponding with the fully decentralized solution. Let* $j = 0$.
2. *If* j *is even, then let* S_2^{j+1} *be the supremal supervisor with respect to uncontrolled system* S_1^j/G *and event set* Σ_2. *Let* $S_1^{j+1} = S_1^j$. *If* j *is odd, then let* S_1^{j+1} *be the supremal supervisor with respect to uncontrolled system* S_2^j/G *and event set* Σ_1. *Let* $S_2^{j+1} = S_2^j$.
3. *If* $(S_1^{j+1}, S_2^{j+1}) \neq (S_1^j, S_2^j)$, *then increment* j *and continue with step* 2.

First it will be shown that all pairs of supervisors (S_1^j, S_2^j) are most restricting.

LEMMA 4.3. *Let* $j \in N$ *and assume that* (S_1^j, S_2^j) *is most restrictive. Then* (S_1^{j+1}, S_2^{j+1}) *obtained in the second step of the procedure is also most restrictive.*

Proof. Assume without loss of generality that j is odd. So $S_2^{j+1} = S_2^j$ and S_1^{j+1} is the supremal supervisor with respect to S_2^j/G. Let $K^j = \mathrm{L}(S_1^j \wedge S_2^j/G)$ and $K^{j+1} = \mathrm{L}(S_1^{j+1} \wedge S_2^{j+1}/G)$. Comparing (9) with Definition 3.5, it is not difficult to see that S_1^{j+1} is most restrictive with respect to K^{j+1}. Supervisor $S_2^{j+1} = S_2^j$ is most restrictive with respect to language K^j. It remains to show that it is most restrictive with respect to K^{j+1}.

$$\begin{aligned} &\sigma \in \gamma(S_2^{j+1}, s_2) = \gamma(S_2^j, s_2) \Rightarrow \sigma \in \Sigma_{\mathrm{c}}(S_2^j) \wedge s_2\sigma \notin \mathrm{L}(S_2^j/G) \\ &\Rightarrow [\text{because } K^{j+1} \subseteq \mathrm{L}(S_2^j/G)]\ \sigma \in \Sigma_{\mathrm{c}}(S_2^j) \wedge s_2\sigma \notin K^{j+1}. \end{aligned}$$

As K^{j+1} is supremal it follows that $K^j \subseteq K^{j+1}$.

$$\begin{aligned} \sigma \notin \gamma(S_2^{j+1}, s_2) = \gamma(S_2^j, s_2) \Rightarrow & \left[\text{because } S_2^j \text{ is most restrictive with respect to} K^j\right] \\ & \sigma \notin \Sigma_{\mathrm{c}}(S_2^j) \vee s_2\sigma \in K^j \\ & \Rightarrow \left[\text{because } K^j \subseteq K^{j+1}\right]\ \sigma \notin \Sigma_{\mathrm{c}}(S_2^j) \vee s_2\sigma \in K^{j+1}. \end{aligned}$$

It follows that S_2^{j+1} is most restrictive with respect to K^{j+1}. And thus (S_1^{j+1}, S_2^{j+1}) is most restrictive. ☐

Next it will be shown that if $(S_1^{j+1}, S_2^{j+1}) = (S_1^j, S_2^j)$, then (S_1^j, S_2^j) forms a Nash equilibrium. So if Σ_1 and Σ_2 are disjoint, then this pair is a maximal solution.

THEOREM 4.4. *Let* $j \in \mathbb{N}$ *and let* $S_1^j, S_2^j, S_1^{j+1}, S_2^{j+1}$ *be constructed by the procedure above. If* $(S_1^{j+1}, S_2^{j+1}) = (S_1^j, S_2^j)$, *then* (S_1^j, S_2^j) *forms a Nash equilibrium.*

Proof. Assume without loss of generality that j is odd. Then, according to the second step of the procedure, $S_2^{j+1} = S_2^j$ and S_1^{j+1} is the supremal supervisor with respect to S_2^j/G. As $S_1^{j+1} = S_1^j$ it follows that S_1^j is optimal if S_2^j is kept fixed. This proves the first part of the Nash equilibrium condition.

From the previous iteration of the procedure it follows that $S_1^j = S_1^{j-1}$ and that S_2^j is the supremal supervisor with respect to S_1^{j-1}/G. In the next iteration supervisor S_2^{j+2} will be synthesized. Supervisor S_2^{j+2} is the optimal solution with respect to $S_1^{j+1}/G = S_1^j/G = S_1^{j-1}/G$. So S_2^{j+2} will be equal to S_2^j. Supervisor S_2^j is optimal if S_1^j is kept fixed. This proves the second part of the Nash equilibrium condition. And thus (S_1^j, S_2^j) is a Nash equilibrium. ☐

Example 4.5. Consider the system described in Example 2.18 and Figure 2. Take the pair of most restrictive supervisors corresponding with the fully decentralized solution as starting point of the procedure. In this case $S_1^0 = \mathrm{proj}(S^{\uparrow}, \Sigma_1)$ and $S_2^0 = \mathrm{proj}(S^{\uparrow}, \Sigma_2)$. Let $\Sigma_1 = \{\mathtt{a}_1, \mathtt{b}_1\}$, $\Sigma_2 = \{\mathtt{a}_2, \mathtt{b}_2\}$, and $\Sigma_{\mathrm{c}} = \{\mathtt{a}_1, \mathtt{a}_2\}$. Note that

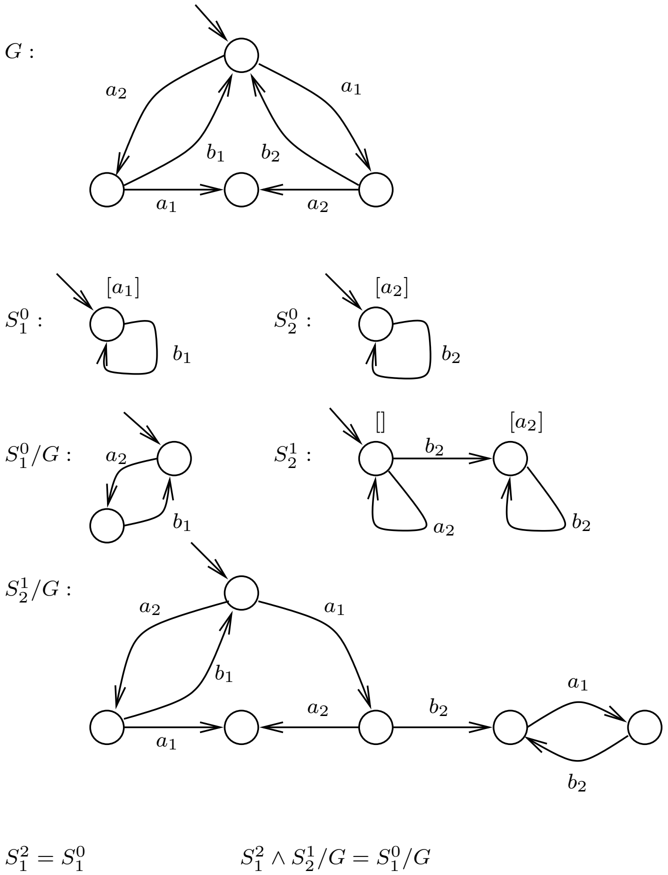


FIG. 4. *Construction of a Nash equilibrium pair of supervisors for Example* 4.5.

the event sets Σ_1 and Σ_2 are disjoint. The construction of the Nash equilibrium is summarized in Figure 4. In this construction use is made of the expression for the control law of (9). First S_1^0 is kept fixed and the optimal supervisor S_2^1 with respect to the uncontrolled system S_1^0/G is derived. Note that $L(S_1^0/G) \subseteq L(S^{\uparrow})$ and thus by (8) $K = L(S_1^0/G)$.

Next, S_2^1 is kept fixed. The language $L(S_2^1/G) \not\subseteq L(S^{\uparrow})$, so K is a proper subset of $L(S_2^1/G)$. The optimal supervisor S_1^2 with respect to the uncontrolled system S_2^1/G is derived. It turns out that $S_1^2 = S_1^0$. In subsequent steps the pair of supervisors remains invariant. The pair (S_1^2, S_2^1) is thus a Nash equilibrium, and therefore, according to Theorem 3.7, a maximal solution. The closed-loop system according to this Nash equilibrium is identical to S_1^0/G.

Example 4.6. Now, consider a slight alteration of the control problem of Example 4.5. Let $\Sigma_c = \Sigma$ and let the rest be unchanged. Take, as before, the pair of most restrictive supervisors corresponding with the fully decentralized solution as a starting point. In this case also the b-events are disabled. The construction of the pair of supervisors is then illustrated in Figure 5.

The procedure will converge to the limit pair (S_1^*, S_2^*). However, this solution will not be obtained in a finite number of steps.

The example shows that a small change in the parameters of the problem may

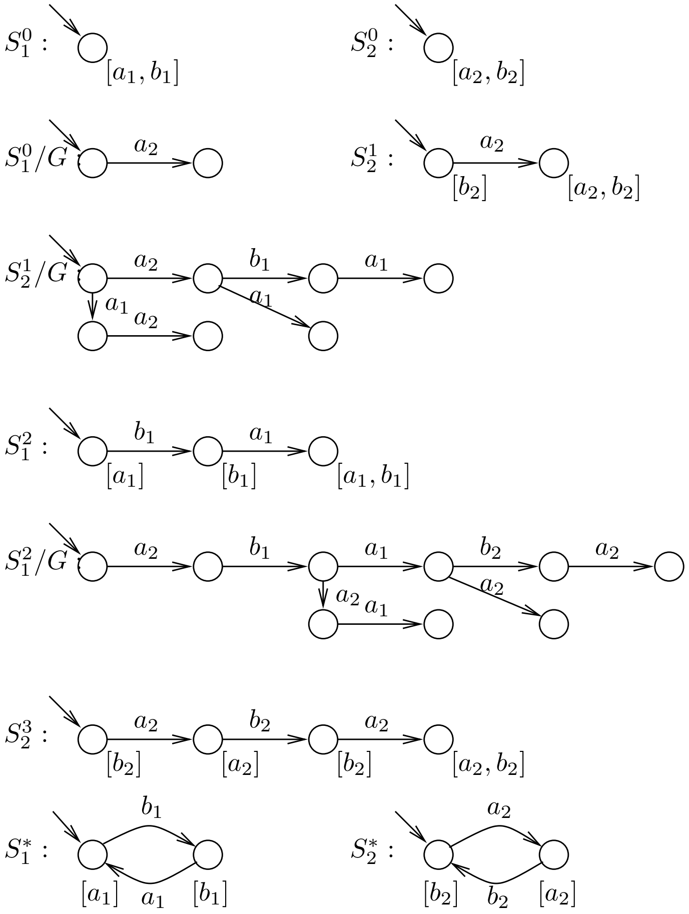


FIG. 5. *Construction of a Nash equilibrium pair of supervisors for Example* 4.6.

lead to a different solution. It may even cause the procedure to become nonhalting.

Up until now these particularities are not fully understood. Further research is needed to adapt the algorithm such that it always converges in a finite number of steps. Also, further research is required to understand the relationship between the initial parameters and the eventual solution. For decentralized control of finite-dimensional linear systems, there is an example for which the procedure does not stop after a finite number of steps and for which a decentralized controller is infinite-dimensional (see [18]). For concrete decentralized supervisory control problems a few steps of the procedure may yield a useful pair of supervisors.

It would be ideal if the procedure could produce a representation of all maximal solutions. It is not certain whether such a representation is finite.

5. Conclusions. For decentralized supervisory control a pair of supervisors is defined to be a maximal solution if there does not exist another such pair with a strictly larger closed-loop language. It has been argued that a maximal solution is of interest to control synthesis of decentralized discrete-event systems. The construction of a maximal solution is handled by analogy with game and team problems. A pair of supervisors is defined to be a Nash equilibrium if, when one supervisor is kept fixed, the other cannot be changed so as to enlarge the closed-loop language and conversely. The main result is then that a pair of supervisors is a strong Nash equilibrium if and only if it is a maximal solution.

A procedure is presented for the construction of a strong Nash equilibrium of a pair of supervisors. The procedure alternatingly keeps one supervisor fixed and solves a supervisory control problem for the other supervisor. The procedure is shown to work on an example. Another example establishes that the procedure may not stop after any finite number of steps.

Major open questions are (1) the classification of all maximal solutions for the decentralized supervisory control problem, and (2) conditions under which Procedure 4.2 stops after a finite number of steps. Experience should be gained with this approach to decentralized supervisory control.

REFERENCES

[1] T. Basar and G. Olsder, *Dynamic Noncooperative Game Theory*, Academic Press, New York, London, 1982.

[2] R. Cieslak, C. Desclaux, A. Fawaz, and P. Varaiya, *Supervisory control of discrete-event processes with partial observations*, IEEE Trans. Automat. Control, 33 (1988), pp. 249–260.

[3] Y. Ho, *Team decision theory and information structures*, Proc. IEEE, 68 (1980), pp. 644–654.

[4] P. Kozak and W. M. Wonham, *Fully Decentralized Solutions of Supervisory Control Problems*, report 9310, Systems and Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Canada, 1993.

[5] R. Kumar and M. A. Shayman, *Centralized and decentralized supervisory control of nondeterministic systems under partial observation*, SIAM J. Control Optim., 35 (1997), pp. 363–383.

[6] F. Lin and W. M. Wonham, *Decentralized control and coordination of discrete-event systems*, in Proceedings of the 27th IEEE Conference on Decision and Control, Austin, TX, IEEE, New York, 1988, pp. 1125–1130.

[7] F. Lin and W. M. Wonham, *Decentralized control and cooperation of discrete event systems with partial observations*, IEEE Trans. Automat. Control, 35 (1990), pp. 1330–1337.

[8] J. Nash, *Non-cooperative games*, Ann. of Math. (2), 54 (1951), pp. 286–295.

[9] G. Olsder, *Comments on a numerical procedure for the solution of differential games*, IEEE Trans. Automat. Control, 20 (1975), pp. 704–705.

[10] A. Overkamp, *Discrete Event Control Motivated by Layered Network Architectures*, Ph.D. thesis, University of Groningen, Groningen, The Netherlands, 1996.

[11] A. Overkamp and J. van Schuppen, *A characterization of maximal solutions for decentralized discrete event control problems*, in Proceedings of the International Workshop on Discrete Event Systems, Edinburgh, Scotland, UK, IEE, London, UK, 1996, pp. 278–283.

[12] R. Radner, *Team decision problems*, Ann. Math. Statist., 33 (1962), pp. 857–881.

[13] P. Ramadge and W. M. Wonham, *The control of discrete event systems*, Proc. IEEE, 77 (1989), pp. 81–98.

[14] K. Rudie, *Decentralized Control of Discrete-Event Systems*, Ph.D. thesis, Department of Electrical Engineering, University of Toronto, Toronto, Canada, 1992.

[15] K. Rudie and W. M. Wonham, *Supervisory control of communicating processes*, in Protocol Specification, Testing and Verification X, L. Logrippo, R. Probert, and H. Ural, eds., North-Holland, Amsterdam, 1990, pp. 243–257.

[16] K. Rudie and W. M. Wonham, *Protocol verification using discrete-event systems*, in Proceedings of the 31st IEEE Conference on Decision and Control, New York, NY, 1992, pp. 3770–3777.

[17] K. Rudie and W. M. Wonham, *Think globally, act locally: Decentralized supervisory control*, IEEE Trans. Automat. Control, 37 (1992), pp. 1692–1708.

[18] N. Sandell Jr., P. Varaiya, M. Athans, and M. Safonov, *Survey of decentralized control methods for large scale systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 108–128.

[19] A. Tanenbaum, *Computer Networks*, Prentice-Hall International, London, UK, 1981.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 512–532

ON THE SYNTHESIS OF OPTIMAL SCHEDULERS IN DISCRETE EVENT CONTROL PROBLEMS WITH MULTIPLE GOALS[*]

HERVÉ MARCHAND[†], OLIVIER BOIVINEAU[‡], AND STÉPHANE LAFORTUNE[‡]

Abstract. This paper deals with a new type of optimal control for discrete event systems. Our control problem extends the theory of [R. Sengupta and S. Lafortune, *SIAM J. Control Optim.*, 36 (1998), pp. 488–541] that is characterized by the presence of uncontrollable events, the notion of occurrence and control costs for events, and a worst-case objective function. A significant difference with [R. Sengupta and S. Lafortune, *SIAM J. Control Optim.*, 36 (1998), pp. 488–541] is that our aim is to make the system evolve through a set of multiple goals, one by one, with no order necessarily prespecified, whereas the previous theory only deals with a single goal. Our solution approach is divided into two steps. In the first step, we use the optimal control theory in [R. Sengupta and S. Lafortune, *SIAM J. Control Optim.*, 36 (1998), pp. 488–541] to synthesize individual controllers for each goal. In the second step, we develop the solution of another optimal control problem, namely, how to modify if necessary and piece together, or schedule, all of the controllers built in the first step in order to visit each of the goals with the least total cost. We solve this problem by defining the notion of a scheduler and then by mapping the problem of finding an optimal scheduler to an instance of the well-known traveling salesman problem (TSP) [E. L. Lawler, J. K. Lenstra, A. H. G. Rinooy Kan, and D. B. Shmoys, *The Traveling Salesman Problem*, John Wiley, 1985]. We finally suggest various strategies to reduce the complexity of the TSP resolution while still preserving global optimality.

Key words. discrete event systems, optimal control, scheduler, traveling salesman problem

AMS subject classifications. 93A99, 49-XX, 90C27, 90B35

PII. S0363012998341964

1. Introduction and motivation.

We are interested in a new class of optimal control problems for discrete event systems (DES). We adopt the formalism of supervisory control theory [16] and model the system as the regular language generated by a finite state machine (FSM). Our control problem follows the theory in [19, 20, 21] and is characterized by the presence of uncontrollable events, the notion of occurrence and control costs for events, and a worst-case objective function. A significant difference with the work in [21] and with the other works dealing with optimal control of DES [6, 11, 14, 24] is that we wish to make the system evolve through a set of marked states (or multiple goals) one by one, with no order necessarily specified a priori; in contrast, the previous theories only deal with a single marked state.

Our problem formulation is motivated by several application domains such as test objective generation in verification and diagnostics, planning in environments with uncertain results of actions, and routing in communication networks.

- In test objective generation, a given system has been designed to meet some specific requirements. However, it may happen that some of these require-

[*] Received by the editors July 10, 1998; accepted for publication (in revised form) May 11, 2000; published electronically September 15, 2000. This research was supported in part by INRIA and by the Department of Defense Research and Engineering (DDR&E) Multidisciplinary University Research Initiative (MURI) on Low Energy Electronics Design for Mobile Platforms and managed by the Army Research Office (ARO) under grant DAAH04-96-1-0377.

http://www.siam.org/journals/sicon/39-2/34196.html

[†] IRISA / INRIA - Rennes, F-35042 RENNES, France (hmarchan@irisa.fr). The work of this author was carried out at the Department of Electrical Engineering and Computer Science, University of Michigan.

[‡] Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122 (oboivine@eecs.umich.edu, stephane@eecs.umich.edu).

ments have been overlooked or neglected. Failures can occur as a consequence of negligence. Test objective generation is a way of (ideally exhaustively) checking for inconsistencies in the behavior of the system [1, 3, 17]. The marked states (the states of interest) would be some particular states in which the behavior of the system to be tested is suspected to be flawed. The method that we develop generates a behavior for the system that allows it to reach all these states in an optimal way, with respect to the given occurrence and control cost functions for the events. Each time a state of interest is reached, a behavioral test can be performed on this particular state in order to check if it meets the requirements and conforms to the designed or expected behavior.

- In artificial intelligence (AI), the behavior of an agent is often sought to be optimized with respect to an optimality criterion [5]. Moreover, dealing with multiple goals is an active area of research in AI [13]. The model and the methods that we develop in this work can easily be applied to an agent evolving in an environment where the results of its actions are not always the ones expected. Under certain restrictions, there is a mapping between partial controllability in DES and the notion of a nondeterministic environment[1] in AI [18]. The notion of an optimal scheduler that we define and construct can be used to do planning with multiple goals.
- Broadcasting and multicasting in a communication network is an instance of a multiagent system. Here, the marked states would represent the nodes of the network to which we would want the information to be sent. The uncontrollability of certain events would be interpreted as the uncertainty regarding the actual route that the information would take, since the entire route is not up to the decision of the single sending agent. The solution that we generate can be used to determine the number of duplicated messages that must be sent in parallel through the network in order for all the desired recipients to receive the piece of information.

Our solution approach consists of two steps. The starting point is an FSM which represents the desired behavior of a given system. From this FSM, we can generate a controller that verifies any property that we would wish to associate to it, from the set of acceptable controllers. The desirable property is often taken to minimize a quantitative performance measure. In our case, we generate a controller which verifies a range of properties. This is what has been called the DP-optimality property of an FSM [21]. DP-optimality stands for dynamic programming optimality. This comes from the fact that we use back-propagation from the goal state to generate the controller, based on event cost functions. The controller is represented as an FSM also. The theory of DP-optimal controllers has been developed in the restricted case of one unique marked state [19, 20, 21]. We use the theory in [21] to synthesize a set of optimal controllers corresponding to the different marked states, each treated individually. This yields a set of FSMs that are generated independently from each other. These controllers are synthesized in a manner that gives them an optimal substructure, consistent with the notion of DP-optimality of [21]. The objective function has a worst-case form. The total worst-case computational complexity of the first step is

[1] The notion of a nondeterministic environment in AI is different from the notion of a nondeterministic FSM in control of DES. In AI, a nondeterministic environment is one where the actions undertaken by the agent might not lead to the expected arrival state of the world, whereas in control of DES, a nondeterministic FSM is one in which there are identically labeled transitions that lead from one state to different states.

cubic in the number of states in the systems. At this point, the notion of a DP-optimal controller is replaced by the notion of a stepwise DP-optimal scheduler. By scheduler, we mean a sequence of behaviors that are modeled by FSMs. We develop the solution of a "higher-level" optimal control problem, where we use all the controllers built in the first step in order to visit each of the marked states with least total cost; we call this problem that of finding a "stepwise DP-optimal scheduler." We solve this problem by defining the notion of a scheduler and then by mapping the problem of finding a stepwise DP-optimal scheduler to an instance of the well-known traveling salesman problem (TSP) [8]. We finally suggest different strategies to reduce the computational complexity of this step while still preserving global optimality by taking advantage of some particular properties of the structure of stepwise DP-optimal schedulers.

One of the differences between DP-optimality and stepwise DP-optimality resides in the controller having an FSM structure, whereas the scheduler is a concatenation of FSMs. All the states appear only once in a controller, whereas states can appear several times in a scheduler, but under different circumstances, i.e., in different submachines. Also, another difference between DP-optimal controllers and a stepwise DP-optimal scheduler for an FSM is the existence of a unique maximal DP-optimal controller which contains all the other DP-optimal controllers as submachines, whereas there is no notion of a unique maximal stepwise DP-optimal scheduler.

This paper is organized as follows. In section 2, the necessary notations are introduced. In section 3, we recall the basic definitions and properties of the optimal control theory of DES of [19, 20, 21]. More precisely, we review the notion of a DP-optimal submachine of an FSM G. This definition is used as a springboard to section 4, where we introduce the enlarged problem in the case of multiple marked states. In section 5, we define the notion of an optimal scheduler; such a scheduler ensures that the system will visit each state in a given set of states at least once while minimizing a given cost function over the trajectories of the system. We then suggest possible simplifications that can be made to reduce the overall complexity of the computation of a stepwise DP-optimal scheduler. Section 6 illustrates this new notion of optimality with an example. Section 7 presents some possible applications of the theory that is developed throughout this paper. A conclusion and discussion on future works are presented in section 8.

2. Preliminaries. In this section, the main concepts and notations are defined (more definitions will be made when necessary in the following sections). The system to be controlled is modeled as an FSM defined by a 5-tuple $G = \langle \Sigma, Q, q_0, Q_m, \delta \rangle$, where Σ is the set of events, Q is the (finite) set of states, q_0 is the initial state, Q_m is the set of marked states, and δ is the partial transition function defined over $\Sigma^* \times Q \to Q$. The notation $\delta_G(\sigma, q)!$ means that $\delta_G(\sigma, q)$ is defined, i.e., there is a transition labeled by event σ out of state q in machine G. Likewise, $\delta_G(s, q)$ denotes the state reached by taking the sequence of events defined by trace s from state q in machine G. The behavior of the system is described by the prefix-closed language $\mathcal{L}(G)$ [2], generated by G. $\mathcal{L}(G)$ is a subset of Σ^*, where Σ^* denotes the Kleene closure of the set Σ [4]. Similarly, the language $\mathcal{L}_m(G)$ corresponds to the marked behavior of the FSM G, i.e., the set of trajectories of the system ending in one of the marked states of G.

Some of the events in Σ are uncontrollable, i.e., their occurrence cannot be prevented by a controller, while the others are controllable. In this regard, Σ is partitioned as $\Sigma = \Sigma_c \cup \Sigma_{uc}$, where Σ_c represents the set of controllable events and Σ_{uc} represents the set of uncontrollable events. In what follows, we will only be interested in *trim* FSMs (i.e., FSMs whose states are all accessible from q_0 and coaccessible to Q_m). For explicit mathematical definitions, the reader may refer to [2]. We say that an FSM $A = \langle \Sigma, Q_A, q_{0A}, Q_{mA}, \delta_A \rangle$ is a submachine of G if $\Sigma_A \subseteq \Sigma$, $Q_A \subseteq Q$, $Q_{m_A} \subseteq Q_m$, and $\forall \sigma \in \Sigma_A, q \in Q_A,\quad \delta_A(\sigma, q)! \Rightarrow (\delta_A(\sigma, q) = \delta(\sigma, q))$. The statement $A \subseteq G$ denotes that A is a submachine of G. We also say that A is a submachine of G at q whenever $q_{0A} = q \in Q$ and $A \subseteq G$. For any $q \in Q$, we will use $\mathcal{M}(G, q, Q_m) = \{A \subseteq G : A \text{ is trim with respect to } Q_{m_A} \text{ and } q_{0_A} = q\}$ to represent the set of trim submachines of G at q with respect to Q_m. This set has a maximal element in the sense that this maximal element contains all other elements as submachines. It is denoted as $M(G, q, Q_m)$. For convenience, we write $\mathcal{M}(G, q)$ and $M(G, q)$ when there is only one marked state, i.e., when $Q_m = \{q_m\}$.

As stated in [21], to take into account the numerical aspect of the optimal control problem, costs are associated with each event of Σ. To this effect, we introduce an occurrence cost function $c_e : \Sigma \to \mathbb{R}^+ \cup \{0\}$ and a control cost function $c_c : \Sigma \to \mathbb{R}^+ \cup \{0, \infty\}$. Occurrence cost functions are used to model the cost incurred in executing an event (energy, time, etc.). Control cost functions are used to represent the fact that disabling a transition possibly incurs a cost. The control cost function is infinity for events of Σ_{uc}. These cost functions are then used to introduce a cost on the trajectories of a submachine A of G. To this effect, we first define a projection p_j that, when applied to a trace of events $s = \sigma_1^s \sigma_2^s \dots \sigma_{\|s\|}^s$, gives the subtrace of s of length j starting from σ_1^s ($p_j(s) = \sigma_1^s \sigma_2^s \dots \sigma_j^s$ if $j \le \|s\|$, and is undefined otherwise). We also introduce $\Sigma_d^G(A, q)$ as the set of disabled events at state q for the system to remain in submachine A of G.

Definition 2.1. *Let A be a submachine of G, and let $\mathcal{L}_m(A)$ be the marked language of A. Then the following are defined.*

- *For any state $q \in Q_A$ and string $s = \sigma_1^s \sigma_2^s \dots \sigma_{\|s\|}^s$ such that $\delta_A^*(s, q)$ exists, the cost of the string s is given by*

$$c^g(q, A, s) = \sum_{j=1}^{\|s\|} c_e(\sigma_j^s) \; + \; \sum_{j=1}^{\|s\|} \sum_{\substack{\sigma \in \Sigma_d^G(A, q') \\ q' = \delta_A(p_j(s), q)}} c_c(\sigma). \tag{2.1}$$

- *The objective function denoted as $c_{sup}^g(.)$ is given by*

$$c_{sup}^g(A) = \sup_{s \in \mathcal{L}_m(A)} c^g(q_{0A}, A, s). \tag{2.2}$$

Basically, the cost of a trajectory is the sum of the occurrence costs of the events belonging to this trajectory to which is added the cost of disabling events on the way to remain in A. If an uncontrollable event is disabled, this renders the cost of a trajectory infinite because the second term of (2.1) becomes infinity. The notation $c_{sup}^g(A)$ represents the worst-case behavior that is possible in submachine A.

3. Review of the DP-optimal problem for one final state. In general, the purpose of optimal control is to study the behavioral properties of a system, to take advantage of a particular structure, and to generate a controller which constrains the system to a desired behavior according to quantitative and qualitative aspects. In the basic setup of supervisory control theory (see [15, 16] and Chapter 3 of [2]), optimality is with respect to set inclusion, and thus all legal behaviors are equally good (zero cost) and illegal behaviors are equally bad (infinite cost). The work in [21] enriches this setup by the addition of quantitative measures in the form of occurrence and control cost functions, to capture the fact that some legal behaviors are better than others. The problem is then to synthesize a controller that is not only legal, but also "good" in the sense of given quantitative measures. Some other studies appear in [6, 11, 14, 24]. In this section, we present some results of [21] that are necessary for developing the solution procedure for optimal schedulers. Our aim here is not to describe in detail all the theory, which can be found in [19, 20, 21], but to present the principal notations and results that we use in what follows.

Definition 3.1. *A submachine A of G is said to be controllable if $\forall\ q \in Q_A$, such that there exists $s \in \Sigma^*$ and $\delta(s, q_{o_A}) = q$, the following is satisfied:*

$$\forall\sigma((\sigma \in \Sigma_{uc}) \wedge (\delta(\sigma, q)!)) \Rightarrow \delta_A(\sigma, q)!$$

We now define the optimization problem for a single marked state q_m.

Definition 3.2. *$\forall q \in Q$, $A_o \in \mathcal{M}(G, q)$ is an optimal submachine if*

$$c^g_{\sup}(A_o) = \min_{A\in\mathcal{M}(G,q)} c^g_{\sup}(A) < \infty.$$

For such a submachine A_o, $c^g_{\sup}(A_o)$ represents the optimal cost (in fact, the worst inevitable cost) necessary to reach q_m from q_0. It means that a submachine with a lower cost could not ensure the accessibility of q_m from q_0. The following lemma (Lemma 2.15 in [19]) is stated to note that optimal solutions lie within the class of controllable submachines.

Lemma 3.3. *Let $A \in \mathcal{M}(G, q, Q_m)$. If $c^g_{sup}(A) < \infty$, then A is controllable.*
Theorem 4.2 of [19] gives necessary and sufficient conditions for the existence of optimal submachines as follows.

Theorem 3.4. *An optimal submachine of G exists if and only if there exists a submachine A of G such that A is trim, controllable and $\forall s \in \mathcal{L}(G)$ and $q \in Q$ such that $\delta(s, q) = q$ we have $c^g(q, A, s) = 0$.*

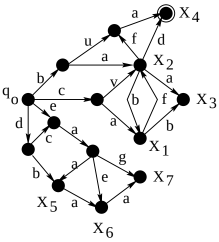
Intuitively, this theorem states that an optimal solution exists when there are controllable submachines of G in which all cycles have a zero cost. The controllability assumption ensures that the positive cost cycles can be broken using controllable events alone. We now introduce the notion of DP-optimal submachines. This kind of submachine will be used intensively in the next sections.

Definition 3.5. *A submachine $A_{DO} \in \mathcal{M}(G, q)$ is DP-optimal if it is optimal and $\forall q' \in Q_{A_{DO}}$, $M(A_{DO}, q')$ is an optimal submachine in $\mathcal{M}(G, q')$.*

If a particular DP-optimal FSM includes all other DP-optimal FSMs as submachines of itself, then we call it the *maximal DP-optimal submachine.* The maximal DP-optimal submachine of a machine G at q with respect to the final marked state q_m will be denoted by $M^o_D(G, q, q_m)$. Note that all DP-optimal submachines are acyclic. The existence of a DP-optimal submachine of G is given by the following theorem (Theorem 4.3 of [19]).

Theorem 3.6. *If an optimal submachine of G exists, then the unique maximal DP-optimal submachine $G^m_{des} = M^o_D(G, q_0, q_m)$ of G with respect to the final state q_m also exists.*

The cyclic DP-optimal algorithm. Consider an FSM $G = \langle\Sigma, Q, q_0, q_m, \delta\rangle$ with a unique initial state q_0 and a unique marked state q_m. Assume that all occurrence costs are strictly positive; then there exists an algorithm [21], named **DP-Opt**, with a worst-case complexity $\mathcal{O}(|Q|^2|\Sigma|\log(|\Sigma|)+|Q|^3|\Sigma|)$ (Theorem 6.10 of [21]), that constructs the desired maximal DP-optimal submachine of the FSM G with respect to q_0 and q_m, that we denote as G^m_{des}. The algorithm also returns the worst inevitable



Event	c_e	Remarks
a,f	1	Controllable
b,c,d,e	2	Controllable
g	4	Controllable
u	3	Uncontrollable
v	2	Uncontrollable

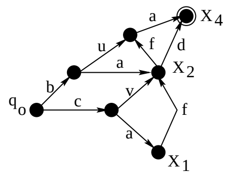


FIG. 3.1. *The initial system G, the event cost function, and the maximal DP-optimal submachine G^4_{des}.*

cost $c^g_{sup}(G^m_{des})$. Moreover, during the computation of the algorithm, we can recover the submachines $M^o_D(G, q, q_m)$ associated with $c^g_{sup}(M^o_D(G, q, q_m))$ for each state visited during the computation. A simplified version of this algorithm can be found in [10] (when the control cost function is reduced to the null function for controllable events).

Example of the DP-optimal problem. We conclude this section by illustrating the DP-optimal problem through an example that is reused in section 6. Let G be an FSM and $\Sigma = \{a, b, c, d, e, f, g, u, v\}$ such that a, b, c, d, e, f, and g are controllable; u and v are uncontrollable. G and the event cost function defined on Σ are as in Figure 3.1. We assume $c_c \in \{0, \infty\}$. Finally, the initial state is q_0 and the final state is X_4.

Using the **DP-Opt** program, we obtain the maximal DP-optimal submachine of G, denoted G^4_{des}, for which the worst inevitable cost is equal to $c^g_{sup}(G^4_{des}) = 6$.

We can observe all the properties of the generated submachine. First, it is controllable, since from any state, there exists a path that leads surely to the goal X_4. Also, it is optimal, since all the paths leading to X_4 have a finite and minimized worst-case cost (notably, no uncontrollable event at state X_4 needs to be disabled). Finally, the DP-optimality property can be observed. From every state q of G^4_{des}, the path from q to X_4 which has the highest cost contains an uncontrollable event u that cannot be disabled.

We have reviewed the optimal control problem and the notion of DP-optimal submachines when only one marked state is present in the system. We now turn our attention to the case of multiple marked states and present our results for this new problem. This will require the introduction of a new, more comprehensive, optimality criterion.

4. The optimal control problem with multiple marked states. In the previous section, we were interested in finding a DP-optimal submachine of G that makes the system evolve from an initial state q_0 to a final state q_m by minimizing a cost function along the various trajectories of the system. Here, our goal is different. We consider an FSM G with a set of multiple marked (or final) states $\mathcal{X} = (X_i)_{i\in[1,\ldots,n]}$. Our aim is now to have the system reach each and every one of the states of $\mathcal{X}$. To account for the fact that it may not be possible to find such a path, we assume in the following the possibility of *resetting* the system to its initial state q_0, when the system has evolved in one of the states of $\mathcal{X}$. The *Reset* event that is added in this section is much more than an artifact for developing the theory. Indeed, many interpretations can be associated with it. First, there are physical systems that can actually be reset to their initial state (like a World Wide Web browser, for example). Second, the *Reset* event can be seen as an event whose occurrence signals the impossibility of visiting all the states of $\mathcal{X}$ without visiting the initial state q_0 more than once. This apparent impossibility can be alleviated by having multiple systems perform in parallel. For example, in the case of a communication network, a message that is sent cannot be brought back to the initial state. However, it can be regenerated, and then the number of *Reset* events can be regarded as an indicator of the number of copies of the message that must be generated and sent in parallel in a broadcast or a multicast (See section 7).

4.1. Stepwise DP-optimality definition. Due to the *Reset* event, the system is now represented by the following FSM $G = \langle \Sigma \cup \{Reset\}, Q, q_0, \mathcal{X}, \delta\rangle$, with $\delta(Reset, X_i) = q_0 \ \forall X_i \in \mathcal{X}$. As in the previous section, we introduce cost functions that take into account the particular *Reset* event: the occurrence cost function $c_e : \Sigma \cup \{Reset\} \to \mathbb{R}^+ \cup \{0\}$ such that $\forall \sigma \in \Sigma, c_e(\sigma) \geq 0$ and $c_e(Reset) = 0$, and the control cost function $c_c : \Sigma \cup \{Reset\} \to \mathbb{R}^+ \cup \{0, \infty\}$ such that $\forall \sigma \in \Sigma, c_c(\sigma) \geq 0$ and $c_c(Reset) = 0$.

DEFINITION 4.1. *Let $s \in \mathcal{L}_m(G)$. The trajectory s is said to be* valid *if there exists at least n prefixes of s, $(s_i)_{i\in[1,\ldots,n]}$, such that $\delta(q_0, s_i) = X_i \in \mathcal{X}$.*

In other words, a trajectory is valid if it makes the system evolve into each of the marked states in $\mathcal{X}$. Note that the definition does not require that the trajectory visit each marked state exactly once. Besides, due to the *Reset* event, the system has the possibility of coming back in its initial state along the trajectory. The set of valid trajectories of the FSM G will be denoted as $\mathcal{S}$.

Given that our primary interest is in the states of $\mathcal{X}$, we introduce the notion of a *valid state trajectory.*

DEFINITION 4.2. *Let s be a valid trajectory in $\mathcal{S}$, such that $s = t_1^s \ldots t_l^s$, with $l > n$ and $\delta(q_0, t_1^s \ldots t_k^s) = X_k^s \in \mathcal{X} \cup \{q_0\}$. We define the function D from $\mathcal{S}$ into $\{q_0\}(\mathcal{X}^*\{q_0\})^*$, such that $D(s) = (X_k^s)_{k\in[1,\ldots,l]}$.*[2] *Such a trajectory is called a* valid state trajectory *with respect to $\mathcal{X}$. We denote as $\mathcal{D}$ the set of valid state trajectories in G, with respect to the set of valid trajectories $\mathcal{S}$: $\mathcal{D} = D(\mathcal{S})$.*

A valid state trajectory $d \in \mathcal{D}$ corresponds to a trajectory in $\{q_0\}(\mathcal{X}^*\{q_0\})^*$ that contains all the states of $\mathcal{X}$ (with possible repetitions).

Since we must deal with a set of marked states rather than with a single marked state, we need to introduce a model that comprises all the states of the set $\mathcal{X}$ and that accounts for the global behavior of the system. It is not possible to use a classical merge operation ($\oplus$, Definition 6.2 in [21]), because states might appear in different submachines in different contexts, i.e., with different partial transition functions associated with them. Therefore, instead of using a merge, we introduce the notion of a scheduler. A scheduler can be thought of as a concatenation of (DP-optimal, in our case) submachines. The role of the scheduler is then to make the system evolve according to one submachine at a time and to account for switching between them at appropriate instants. In what follows, the symbol "$\circ$" will denote the concatenation of two submachines A and A' of G. It is defined in terms of languages. Let $\mathcal{L}_m(A)$ and $\mathcal{L}_m(A')$ be the marked languages of A and A'. Then $\mathcal{L}_m(A \circ A') = \{st \ : \ s \in \mathcal{L}_m(A), \ t \in \mathcal{L}_m(A')\}$. Note that $\mathcal{L}_m(A \circ A') \subseteq \mathcal{L}_m(G)$ if and only if $Q_{m_A} = \{q_{0_{A'}}\}$ and $Q_{m_{A'}} \subseteq Q_{m_G} = \mathcal{X}$. Also note that, due to possible cycles in the FSM G, $A \circ A'$ is in general no longer a submachine of G since some state q of G may be shared by the two submachines A and A' but without the same transitions.

[2] This function allows the "extraction" of the state trajectory in G from the valid trajectory s.

Definition 4.3. *Let $d = (X^d_{k'})_{k' \in [0,\dots,l]} \in \mathcal{D}$ be a valid state trajectory of $\mathcal{X} \cup \{q_0\}$, and let $(A_k)_{k \in [1,\dots,l]}$ such that $l \geq n$ and $A_k \in \mathcal{M}(G, X^d_{k-1}, X^d_k)$ $\forall k \in [1,\dots,l]$; then the structure $A = A_1 \circ A_2 \circ \cdots \circ A_l$ is called a* scheduler *with respect to G and $\mathcal{X}$. The set of schedulers with respect to G and $\mathcal{X}$ is denoted as $\mathcal{M}^{sc}(G, \mathcal{X})$.*

In this particular case, for each submachine of the scheduler, there is only one initial state and one final state. Hence, for two consecutive submachines A_i and A_{i+1}, we have $q_{m_{A_i}} = q_{0_{A_{i+1}}}$. Note that for a scheduler $A = A_1 \circ A_2 \circ \cdots \circ A_l$, some A_k may be simply reduced to the simple FSM $(X^d_k \xrightarrow{Reset} q_0)$. This FSM is clearly a DP-optimal submachine from X^d_k to q_0. Besides, in some cases, $\mathcal{M}^{sc}(G, \mathcal{X})$ can be reduced to $\emptyset$. The cost associated with a scheduler $A = A_1 \circ A_2 \circ \cdots \circ A_l$, denoted as $C^{sc}_{sup}(A)$, is given by

$$C^{sc}_{sup}(A) = \sum_{i=1}^{l} c^g_{sup}(A_i). \tag{4.1}$$

The following definition extends the notion of DP-optimality to the notion of stepwise DP-optimality.

Definition 4.4. *Let $A \in \mathcal{M}^{sc}(G, \mathcal{X})$ be a scheduler, such that A makes the system evolve through a valid state trajectory $d = (X^d_{k'})_{k' \in [0,\dots,l]}$ of $\mathcal{D}$. $A = A_1 \circ A_2 \circ \cdots \circ A_l$ is said to be* stepwise DP-optimal *if each of the submachines $A_k \in \mathcal{M}(G, X^d_{k-1}, X^d_k)$ is DP-optimal with respect to its initial state X^d_{k-1} and final state X^d_k, and if the following condition is satisfied:*

$$C^{sc}_{sup}(A_o) = \min_{A \in \mathcal{M}^{sc}(G,\mathcal{X})} C^{sc}_{sup}(A) < \infty.$$

We wish to draw attention to the following assumption.

Assumption 4.1. From now on, we assume that the DP-optimal submachines under consideration, with the exception of $(X^d_k \xrightarrow{Reset} q_0)$, are maximal. This is done for two main reasons. First, the algorithm **DP-Opt** (see Appendix A of [10]) outputs exactly the maximal DP-optimal submachines. Second, taking the maximal DP-optimal submachines allows the system greater freedom. Indeed, it contains all the other DP-optimal submachines; therefore, it has more possible paths from the initial state to the final marked state. In most applications, it is desirable to lower the probability of taking the worst-case cost path, which is the intent of taking the maximal DP-optimal submachine for $(G^i_{des})_{i \in [1,\dots,n]}$. The more possible paths there are, the less likely it is for the system to take the worst-case cost path. Note that the *Reset* machine $(X^d_k \xrightarrow{Reset} q_0)$ need not be maximal (this can only happen if occurrence costs cannot be equal to zero); in this case, however, given our interpretation of the *Reset* event, we will include the single transition $(X^d_k \xrightarrow{Reset} q_0)$ in the scheduler.

Under this assumption, the following property is a direct consequence of Definition 4.4.

Property 4.2. *Let G be an FSM, and let $\mathcal{X}$ be the set of marked states of G. Let A be a stepwise DP-optimal scheduler, such that $A = A_1 \circ A_2 \circ \cdots \circ A_l$. Let $d = (X^d_k)_{k \in [0,\dots,l]}$ of $\mathcal{D}$ be the associated valid state trajectory. Then $\forall k \in [1,\dots,l]$, $A_k = M^o_D(G, X^d_{k-1}, X^d_k)$. Furthermore, the global cost of the scheduler is*

$$C^{sc}_{sup}(A) = \sum_{k=1}^{l} c^g_{sup}(M^o_D(G, X^d_{k-1}, X^d_k)) < \infty. \tag{4.2}$$

This property states that if a stepwise DP-optimal scheduler exists, then all the submachines constituting this scheduler are the respective $M_D^o(G, X_{k-1}, X_k)$. Moreover the cost of the scheduler is then simply equal to the sum of the costs of these DP-optimal submachines. We will refer to this important result as the *additivity property* of the stepwise DP-optimal scheduler. In what follows, the set of all schedulers A such that all the submachines of A are of the form $M_D^o(G, X_i, Xj)$ for $X_i, X_j \in \mathcal{X} \cup \{q_0\}$, is denoted $\mathcal{M}_D^{sc}(G, \mathcal{X})$.

Now that we have defined the notion of a stepwise DP-optimal scheduler and given some of its properties, we need to give necessary and sufficient conditions for its existence. The next subsection gives these conditions and also proves desirable properties of such a scheduler.

4.2. Existence of a stepwise DP-optimal scheduler. Theorem (4.7) presented below gives necessary and sufficient conditions for the existence of a stepwise DP-optimal scheduler. First we prove the following lemma.

LEMMA 4.5. *If the DP-optimal submachines $M_D^o(G, X_i, X_j)$ and $M_D^o(G, X_j, X_k)$ of G exist, then there exists a DP-optimal submachine $M_D^o(G, X_i, X_k)$. Moreover, we have the following* triangular inequality:

$$c^g_{sup}(M_D^o(G, X_i, X_k)) \le c^g_{sup}(M_D^o(G, X_i, X_j)) + c^g_{sup}(M_D^o(G, X_j, X_k)). \tag{4.3}$$

Proof. Assume the existence of $M_D^o(G, X_i, X_j) = \langle \Sigma_{ij}, Q_{ij}, X_i, X_j, \delta_{ij} \rangle$ and of $M_D^o(G, X_j, X_k) = \langle \Sigma_{jk}, Q_{jk}, X_j, \{X_k\}, \delta_{jk} \rangle$. Consider the intersection of the states of these two submachines as being $Q_{ij} \cap Q_{jk} = \{X_j, q_1, \ldots, q_n\}$. Note that this intersection might be reduced to $\{X_j\}$. We construct a new submachine $G_{ik} = \langle \Sigma_{ik}, Q_{ik}, q_{0_{ik}}, Q_{m_{ik}}, \delta_{ik} \rangle$ from these submachines:

$$G_{ik} = \begin{cases} \Sigma_{ik} & = \quad \Sigma_{ij} \cup \Sigma_{jk}, \;\; Q_{ik} \;\; = \;\; Q_{ij} \cup Q_{jk}, \\ q_{0_{ik}} & = \quad X_i, \qquad\quad Q_{m_{ik}} \;\; = \;\; \{X_k\}, \\ \delta_{ik}(\sigma, q) & = \quad \begin{cases} \delta_{jk}(\sigma, q) \text{ if it exists and } q \in Q_{jk}, \\ \delta_{ij}(\sigma, q) \text{ if it exists and } q \in Q_{ij} - \{X_j, q_1, \ldots, q_n\}, \\ \text{undefined otherwise.} \end{cases} \end{cases}$$

This submachine G_{ik} is well defined. Any possible ambiguity has been eliminated by separately dealing with the states $\{X_j, q_1, \ldots, q_n\}$ in the definition of δ_{ik}. G_{ik} is obtained by always following the partial transition function of $M_D^o(G, X_j, X_k)$ as a default behavior, and following the partial transition function of $M_D^o(G, X_i, X_j)$ otherwise whenever possible. First, the machines $M_D^o(G, X_i, X_j)$ and $M_D^o(G, X_j, X_k)$ are trim. Second, G_{ik} is constructed by forward propagation; therefore, all the states of G_{ik} are accessible with respect to the initial state X_i and are coaccessible with respect to the marked state X_k. Therefore, G_{ik} is trim.

Moreover, G_{ik} is controllable. Indeed, the partial transition function δ_{ik} says that as long as the system has not reached a state of the set $\{X_j, q_1, \ldots, q_n\}$, it follows the partial transition function of δ_{ij}. Due to the DP-optimality of $M_D^o(G, X_i, X_j)$, the system will always reach a state of the set $\{X_j, q_1, \ldots, q_n\}$ with a finite cost. Indeed, if the system never visits a state in $\{q_1, \ldots, q_n\}$, it will eventually reach X_j. Let us call q the first state of the set $\{X_j, q_1, \ldots, q_n\}$ that is visited by the system as it evolves. At this point, the default partial transition function becomes δ_{jk}; therefore, the system will eventually reach the marked state X_k with a finite cost since the submachine $M_D^o(G, X_i, X_j)$ is DP-optimal. Since the cost of reaching X_k from q is finite, the overall cost of reaching X_k is necessarily finite. From Lemma 3.3, G_{ik} is controllable.

Finally, G_{ik} has no positive cost cycles. $M_D^o(G, X_i, X_j)$ and $M_D^o(G, X_i, X_k)$ do not have positive cost cycles (by definition of DP-optimality). As we have described previously, before the system reaches a state of $\{X_j, q_1, \ldots, q_n\}$ for the first time, it will not complete a positive cost cycle (from the DP-optimal nature of $M_D^o(G, X_i, X_j)$). After the system reaches a state of $\{X_j, q_1, \ldots, q_n\}$ for the first time, it will not complete a positive cost cycle either (from the DP-optimal nature of $M_D^o(G, X_j, X_k)$). Therefore, no new cycles have been introduced. The only cycles that may exist in G_{ik} are those of $M_D^o(G, X_i, X_j)$ and $M_D^o(G, X_j, X_k)$.

Given that G_{ik} is trim, controllable, and contains no cycles of positive cost in G, FSM G_{ik} satisfies the preconditions of Theorem (3.4), and there exists an optimal submachine of G_{ik}. Following Theorem 3.6, there also exists a DP-optimal submachine $M_D^o(G, X_i, X_k)$ of G_{ik}.

The proof of the *triangular inequality* relies on what we have said previously. The cost of reaching a state of the set $\{X_j, q_1, \ldots, q_n\}$, from the initial state X_i, is less than $c^g_{sup}(M_D^o(G, X_i, X_j))$ (equality is possible but not necessary when X_j is reached). Once one of the states $\{X_j, q_1, \ldots, q_n\}$ has been reached, the cost for the system to reach the marked state X_k is less than $c^g_{sup}(M_D^o(G, X_j, X_k))$ (equality is possible but not necessary when the system visits X_j) because the corresponding machine is DP-optimal. More formally, let us take a trace s of events that leads from the initial state X_i to the final state X_k, i.e., such that $\delta_{ik}(s, X_i) = X_k$. As seen earlier, s visits at least one state of the set $\{X_j, q_1, \ldots, q_n\}$. Let us call it q again. We can now subdivide s into s_1 and s_2 such that $s = s_1 s_2$, $\delta_{ik}(s_1, X_i) = q$, and $\delta_{ik}(s_2, q) = X_k$. From the DP-optimality of the two submachines $M_D^o(G, X_i, X_j)$ and $M_D^o(G, X_j, X_k) \forall s$ such that $s = s_1 s_2$, $\delta_{ik}(s_1, X_i) = q$, $\delta_{ik}(s_2, q) = X_k$, we can compare

$$\begin{cases} c^g(X_i, M_D^o(G, X_i, X_j), s_1) \leq c^g_{sup}(M_D^o(G, X_i, X_j)), \\ c^g(q, M_D^o(G, X_j, X_k), s_2) \leq c^g_{sup}(M_D^o(G, X_j, X_k)). \end{cases}$$

Since this is true for all traces leading from X_i to X_j, we can deduce the triangular inequality. □

The following corollary uses the construction in the proof of Lemma 4.5 to introduce a necessary condition for the existence of a stepwise DP-optimal scheduler. The proof is straightforward and can be found in [10].

COROLLARY 4.6. *If G^k_{des} does not exist, then there exists no subscheduler that makes the system evolve from q_0 to X_k, should it be indirectly via states of $\mathcal{X}$.*

As a consequence of these results, we can ensure that a state X_k is accessible in an optimal way if and only if G^k_{des} exists. We are now able to give the necessary and sufficient conditions of the existence of a stepwise DP-optimal scheduler. This is stated by Theorem 4.7.

THEOREM 4.7. *There exists a corresponding stepwise DP-optimal scheduler $A \in \mathcal{M}_D^{sc}(G, \mathcal{X})$ if and only if the n DP-optimal submachines G^i_{des} of G exist $\forall X_i \in \mathcal{X}, i \in [1, \ldots, n]$.*

Proof. The necessary condition is given by Corollary 4.6, which states that if there is a state X_i of $\mathcal{X}$ such that there does not exist a DP-optimal submachine G^i_{des}, then there is no way to reach this state with a finite cost (thus in a DP-optimal way) and the goal cannot be achieved. All the states of $\mathcal{X}$ cannot be visited, since one of them cannot be visited. The condition is sufficient since FSM A, such that

$A = G^1_{des} \circ (X_1 \xrightarrow{Reset} q_0) \circ G^2_{des} \circ (X_2 \xrightarrow{Reset} q_0) \circ \cdots \circ G^n_{des} \circ (X_n \xrightarrow{Reset} q_0)$, visits all the states of $\mathcal{X}$. A is then a possible scheduler allowing the achievement of the goal. $\square$

This theorem implies that the stepwise DP-optimal problem has a solution when there exists a DP-optimal submachine for each of the X_i. Besides, if a stepwise DP-optimal solution exists, it need not be unique in general. There is no notion of a maximal stepwise DP-optimal scheduler, as in the DP-optimal problem [21]. The problem of finding one of the optimal schedulers is now explored.

5. Determination of a stepwise DP-optimal scheduler. In this section, we need to assume that the occurrence costs are strictly positive: $\forall \sigma \in \Sigma,\ c_e(\sigma) > 0$. This assumption is necessary when we use the **DP-Opt** algorithm in order to ensure polynomial complexity. We also assume that a DP-optimal submachine exists for all the states $X_i \in \mathcal{X}$. From here on, G^i_{des} will denote the maximal DP-optimal submachine of the particular FSM $G_i = \langle \Sigma, Q, q_0, X_i, \delta \rangle$ output by the **DP-Opt** algorithm. We take advantage of the DP-optimal structure of each of the G^i_{des}. We explore the possibility of starting the system at q_0, reaching a state X_i, and instead of doing a *Reset*, continuing the graph to a state X_j. To do so, we convert the problem to a path-cost minimization problem on a graph equivalent to a TSP.

5.1. Modeling of the problem. In order to convert the stepwise DP-optimal problem into a path-cost minimization problem, we use the **DP-Opt** algorithm. This algorithm computes for each $X_i \in \mathcal{X}$ the DP-optimal submachine G^i_{des}. During this computation, a state X_j belonging to $\mathcal{X}$ may be reached. Due to the DP-optimality definition, the algorithm also gives the DP-optimal submachine between X_j and X_i. The worst inevitable case cost between these two states can be collected as well and placed in a matrix $C \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$ that has the following form (see [10] for the algorithm and further details):

$$\text{(5.1)} \quad \begin{array}{ll} \bullet\ C[i,i] = \infty, & \bullet\ C[i,0] = 0, i \neq 0, \\ \bullet\ C[0,i] = c^g_{sup}(G^i_{des}), i \neq 0, & \bullet\ C[k,i] = \begin{cases} c^g_{sup}(M^o_D(G, X_k, X_i)) \text{ if it exists,} \\ \infty \text{ otherwise.} \end{cases} \end{array}$$

From additivity Property 4.2, the cost of a scheduler $A = A_1 \circ A_2 \circ \cdots \circ A_l$ of $\mathcal{M}^{sc}_D$ is equal to

$$\text{(5.2)}\ C^{sc}_{sup}(A) = \sum_{k=1}^{l} c^g_{sup}(A_k) \ = \ \sum_{k=1}^{l} c^g_{sup}(M^o_D(G, X_{d_{k-1}}, X_{d_k})) \ = \ \sum_{k=1}^{l} C[d_{k-1}, d_k].$$

Considering (5.2), the new optimization problem is now reduced to finding a path with a minimal cost in the directed graph associated with the matrix C. This closely resembles the TSP with the slight difference that multiple visits to states of $\mathcal{X}$ are possible. In this new problem, the "cities" are represented by the set of nodes $\mathcal{X}$, and the "streets" are represented by machines $(G^i_{des})_{i \in [1,\ldots,n]}$ and $M^o_D(G, X_i, X_j)$ when these are available. The costs of these paths are given by the maximum costs for each machine, i.e., the $(c^g_{sup}(G^i_{des}))_{i \in [1,\ldots,n]}$ and the $(c^g_{sup}(M^o_D(G, X_i, X_j)))_{i,j \in [1,\ldots,n]}$. Figure 5.1 illustrates this conversion from the graph of the FSM to the reachability graph.

Note that some elements of C might be equal to ∞ after all G^i_{des} have been computed, which does not mean that the corresponding DP-optimal submachines do not exist. This means that they have not been computed in the algorithm **DP-Opt**.

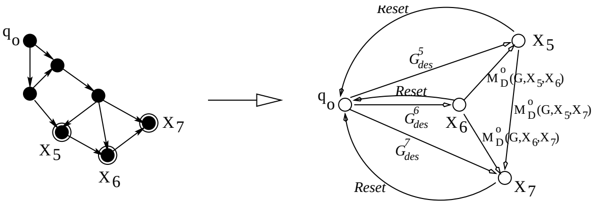


FIG. 5.1. *Conversion from the FSM to a reachability graph on the marked states.*

Indeed, let us suppose that some $M^o_D(G, X_i, X_j)$ has not been computed (and that therefore $C[i, j] = \infty$). It means that it is less costly to perform a *Reset* from state X_i to state q_0, and to reach X_j through G^j_{des}. Another way of seeing this is to look at what the **DP-Opt** algorithm does. It backtracks from the marked state, say X_j. If it reaches X_i before q_0, this means that the cost from X_i to X_j is less than the cost from q_0 to X_j, in which case it is less costly to go directly from X_i to X_j than to reset the system. On the other hand, if state X_i is not reached when the algorithm reaches q_0 during its backtracking, it means that the cost to go from X_i to X_j is greater than the cost of resetting the system (0 in our case) and taking the DP-optimal submachine G^j_{des}. This explains why these paths are not taken into account as possible paths and are directly replaced in the matrix C by an infinite cost.

5.2. Generation of the stepwise DP-optimal scheduler. The problem of finding a stepwise DP-optimal scheduler A_0 has been brought down to solving an instance of the TSP, a classic combinatorial optimization problem. Many methods exist to solve the problem in an acceptable amount of time [8]. We specify once more the conditions in which we solve the TSP. The costs of the paths are all nonnegative. The nodes of $\mathcal{X}$ must be visited *at least once.* One requirement of the TSP is that the salesman come back to the city he started from. This condition does not change anything in our problem since this maps to a *Reset*, which has null cost in our model. Finally, note that the cost matrix C, given in section 5.1, is not necessarily symmetric.

The first step is to transform our modified version of the TSP, where we can visit a node more than once (but at least once), into an ordinary TSP where we must visit each node exactly once. This is typically done by transforming the matrix C into a matrix C', called the all-pairs shortest-paths matrix [8]. Many techniques exist to perform such a computation. Among them is the Floyd–Warshall algorithm [8], which runs in a worst case of $\mathcal{O}(n^3)$, where n represents the number of vertices. Once the all-pairs shortest-paths matrix C' is obtained, we can feed it to a TSP solver.

C and C' have the same dimension but represent different features of the graph. C contains, as noninfinite elements, the costs of the links that actually exist in the graph of the TSP. C' contains the minimum costs necessary to go from one marked state to another, along DP-optimal submachines. C' is a reachability matrix, whereas C is a connectivity matrix. Notably, C' shows if states can be reached by using the *Reset* event. Concretely, to obtain C' from C, one only needs to replace any infinite value in C by the value in the same column in the first line (the cost of the G^i_{des} associated with column i).

Resolution of the TSP. The actual solving of the TSP from matrix C' can be done by using several methods. The most common method is the branch and bound method (see Chapters 9 and 10 of [9] and [12, 7]). An outline of the method can be found in Appendix B of [10]. The worst-case complexity for solving the TSP is

$(n+1)!$. However, the branch and bound method is expected to give a solution to the TSP in a tolerable amount of time. (To give a feel of the time complexity of this method, a 1,000 node fully-connected TSP can be solved in about 20 minutes on a standard workstation.)

The principle of the branch and bound method is quite natural. A branching strategy and a bounding strategy are used alternatively. The branching strategy consists of forcing a supplementary constraint to the system, usually by forcing a set of subpaths in the graph. This allows us to find a solution that is suboptimal in general but that is sometimes optimal. The bounding strategy focuses on finding a lower bound on the cost of the optimal solution by relaxing one of the constraints of the problem (usually by relaxing the constraint that the solution must be a tour). The branching yields a search tree, and the bounding yields a way of quickly finding a suboptimal solution which is close to the optimal solution of the problem. The final solution is optimal.

5.3. Restitution of the stepwise DP-optimal scheduler. From a solution of the TSP, we now build a corresponding stepwise DP-optimal scheduler. The resolution of the TSP provides an optimal solution that gives the ordering in which the states should be visited so as to minimize the worst-case cost. A solution is under the form of a set of $n+1$ pairs (there are $n+1 = |\mathcal{X} \cup \{q_0\}|$ states), in which each state appears exactly once as an initial state and exactly once as a final state of a pair. For pairs (X_i, X_j) that represent a physically existing submachine $M_D^o(G, X_i, X_j)$, i.e., for which $C[i,j] < \infty$, it is sufficient to map these pairs to their associated submachine. As for the pairs (X_i, X_j) that do not map to an existing DP-optimal submachine, i.e., those for which $C[i,j] = \infty$ and $C'[i,j] < \infty$, they are divided into two pairs, namely, (X_i, q_0) and (q_0, X_j). The first is mapped to a *Reset* to the initial state, and the second is mapped to the DP-optimal submachine G_{des}^j.

THEOREM 5.1. *Given a solution of the TSP, by adopting the previous mapping, the obtained scheduler is stepwise DP-optimal.*

Proof. The initial solution of the TSP with respect to the matrix C' is given by a tour of the form $\{(q_0, X_{i_1}); (X_{i_1}, X_{i_2}); \dots (X_{i_j}, X_{i_{j+1}}); \dots; (X_{i_n}, q_0)\}$ with a corresponding cost $TSP(C') = C'[0, i_1] + C'[i_1, i_2] + \cdots + C'[i_j, i_{j+1}] + \cdots + C'[i_n, 0]$. Consider now the transformation previously adopted. If the pair (X_i, X_j) originally exists, i.e., $C[i,j] < \infty$, then the path is admissible in the original problem and we replace the pair by the submachine $M_D^o(G, X_i, X_j)$, where the corresponding cost $c_{sup}^g(M_D^o(G, X_i, X_j))$ is equal to $C[i,j]$. If the pair (X_i, X_j) does not map to an existing DP-optimal submachine, i.e., $C[i,j] = \infty$, then we need to *Reset* the system before directly going to X_j through G_{des}^j. The *triangular inequality* of Lemma (4.5) ensures that in this case, $C'[i,j] = c_{sup}^g(G_{des}^j)$. The pair is then replaced by the subscheduler $(X_i \xrightarrow{Reset} q_0) \circ G_{des}^j$, with the corresponding cost equal to $c_{sup}^g(G_{des}^j)$.

We then obtain a new sequence of pairs, with a cost equal to $TSP(C')$ but for which all the submachines actually exist in the original problem. The DP-optimality of each submachine of the scheduler is given by construction, since we only consider submachines of the form $M_D^o(G, X_i, X_j)$ or G_{des}^j. The minimal cost of the scheduler is ensured by the optimality of the TSP solution and by the fact that the mapping does not add new costs. □

An interesting property is given next. It states that all the submachines that constitute a stepwise DP-optimal scheduler are directly derived from all the DP-optimal submachines built during the computation of the matrix C (see section 5.1 and (5.1)).

PROPOSITION 5.2. *A stepwise DP-optimal scheduler A_o obtained by the TSP solution is composed of exactly n different DP-optimal submachines (not counting the possible Resets of the system). Moreover, all these submachines are obtained from the DP-optimal submachines $(G^i_{des})_{i\in[1,\dots,n]}$ computed during the matrix generation step (see* (5.1)*).*

Proof. The general solution of the TSP for the matrix C' is a tour of the form $\{(q_0, X_{i_1}); (X_{i_1}, X_{i_2}); \dots; (X_{i_j}, X_{i_j}); \dots; (X_{i_n}, q_0)\}$. Note that there are exactly $n+1$ pairs in this tour (but the last pair is a trivial one, i.e., a *Reset*). If a pair (X_i, X_j) originally exists, i.e., if $C[i,j] < \infty$, then the path is in the original problem and it is replaced by the submachine $M^o_D(G, X_i, X_j)$. If not, i.e., if $C[i,j] = \infty$, then the system is reset before directly going to X_j through G^j_{des}. The pair is then replaced by the subscheduler $(X_i \xrightarrow{Reset} q_0) \circ G^j_{des}$. The n pairs are then replaced by either the DP-optimal submachine $M^o_D(G, X_i, X_j) = Trim(G^j_{des}, X_i, X_j)$, or by the subscheduler $(X_i \xrightarrow{Reset} q_0) \circ G^j_{des}$. The final solution of our problem has then exactly n nontrivial submachines that can be obtained from the n DP-optimal submachine $(G^i_{des})_{i=[1,\dots,n]}$ of G, by a trim operation. □

COROLLARY 5.3. *In a stepwise DP-optimal scheduler obtained by the TSP solution, the states of $\mathcal{X}$ are visited exactly once by the stepwise DP-optimal scheduler.*[3]

We wish to draw attention to the following fact. The stepwise DP-optimal scheduler visits each marked state exactly once when it is obtained from the TSP solution. However, the system itself, through its evolution described by the FSM G, may visit a marked state of G more than once. This comes from the fact that the scheduler is constructed on C', whereas the behavior of the system modeled by the FSM G should be observed at a less abstract level, namely at the level of the FSM, G.

Remark 5.1. In [10], we also presented the resolution of the stepwise DP-optimal problem in the case of a nonzero occurrence cost for the *Reset* event (see section 5.3 of [10] for further details).

5.4. Some simplifications of the TSP resolution. In order to solve the stepwise DP-optimal problem, we have to solve the corresponding TSP for the matrix C. The TSP is an NP-complete problem. It is then greatly advantageous to find some simplification methods, taking advantage of the special structure of a stepwise DP-optimal scheduler, in order to reduce the computational complexity of the corresponding TSP without loss of global optimality. Proofs and algorithms are omitted in this section due to lack of space. They can be found in the companion paper [10].

5.4.1. Divide and conquer. In some cases, it is possible to divide the matrix C into several smaller ones. In such cases, it suffices to solve the TSP on each of these submatrices. The following proposition states the necessary and sufficient conditions for this simplification.

PROPOSITION 5.4. *Assume there exists a partition of $\mathcal{X} = \cup_{k\in[1,\dots,l]}(\mathcal{X}_k)$ such that $\forall k_1, k_2 \in [1,\dots,l]$, $\forall X_i \in \mathcal{X}_{k_1}$, and $\forall X_j \in \mathcal{X}_{k_2}$, the submachine $M^o_D(G, X_i, X_j)$ is not defined.*

If A_o is a stepwise DP-optimal scheduler with respect to $\mathcal{X}$, then it is possible to find a set of schedulers $A_{\mathcal{X}_k}$, where each $A_{\mathcal{X}_k}$ is stepwise DP-optimal with respect to $\mathcal{X}_k$ with an optimal cost $c^{sc}_{sup}(A_{\mathcal{X}_k})$ to visit of all the states of $\mathcal{X}_k$, and such that $A_o = \circ^l_{k=1} A_{\mathcal{X}_k}$ with $c^{sc}_{sup}(A_o) = \sum^l_{k=1} c^{sc}_{sup}(A_{\mathcal{X}_k})$.

[3] The proof is omitted and can be found in [10].

In view of Proposition 5.4, the global problem can be solved on each submatrix C_k, $k \in [1,\ldots,l]$, corresponding to the particular set of states $\mathcal{X}_k \cup \{q_0\}$. The necessary computation to find the connected components before applying Proposition 5.4 can be performed in $\mathcal{O}(n+E)$, where E is the number of vertices of the directed graph associated to matrix C (see [10] for details).

5.4.2. Terminal path simplification. We address here a property of the scheduler that can lead to a simplification on the matrix C. This property states that if there exists a kind of "dead-end" in the graph of the matrix, then it is always better to follow this path until the end than to perform a *Reset* and come back to visit the end of this path later.

Proposition 5.5. *Assume that there exists a subset $\mathcal{X}_i = (X_{i_k})_{k\in[1,\ldots,m]}$ of $\mathcal{X}$, with $m < n$ and such that*

1. *$\forall k \in [1,\ldots,m-1]$, $M^o_D(G, X_{i_k}, X_{i_{k+j}})$ exists for $j \in [1,\ldots,m-k]$,*
2. *$\forall k \in [2,\ldots,m]$, $M^o_D(G, X_{i_k}, X_{i_{k-j}})$ does not exist for $j \in [1,\ldots,k-1]$,*
3. *$\forall k \in [1,\ldots,m]$ and $\forall X_l \in \mathcal{X} - \mathcal{X}_i$, $M^o_D(G, X_{i_k}, X_l)$ is not defined.*

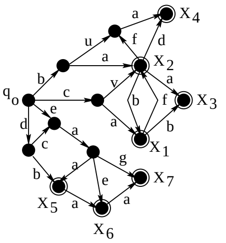
Under these assumptions, the submachines $(G^{i_k}_{des})_{k\in[2,\ldots,m]}$ do not belong to the stepwise DP-optimal scheduler A_o.

Proposition 5.5 deals with situations where a dead-end occurs. By dead-end, we mean a set of states $\{q_1,\ldots,q_n\}$ in which $\forall i \in [1,\ldots,n]$, $(q_j)_{j>i}$ are the only states coaccessible from q_i. If there exists a dead-end in the graph of marked states, the system will never enter that dead-end directly through one of the G^i_{des} but will only enter indirectly from the initial state q_0. This means that no direct submachine of the type G^i_{des} will be used by a scheduler to enter a dead-end. Any visit to a state of the dead-end is done via a visit to a state that does not belong to the dead-end.

An algorithm that performs this simplification on the matrix C according to the three assumptions in Proposition 5.5 is presented in [10]. Its complexity is linear in the number of states of $\mathcal{X}$. With this simplification, the paths of the form $q_0 \to X_{i_k}$ do not constitute valid paths any longer and, consequently, will not be taken into account as possible solutions in the corresponding TSP solution. This terminal path simplification can narrow down the search space when solving the TSP.

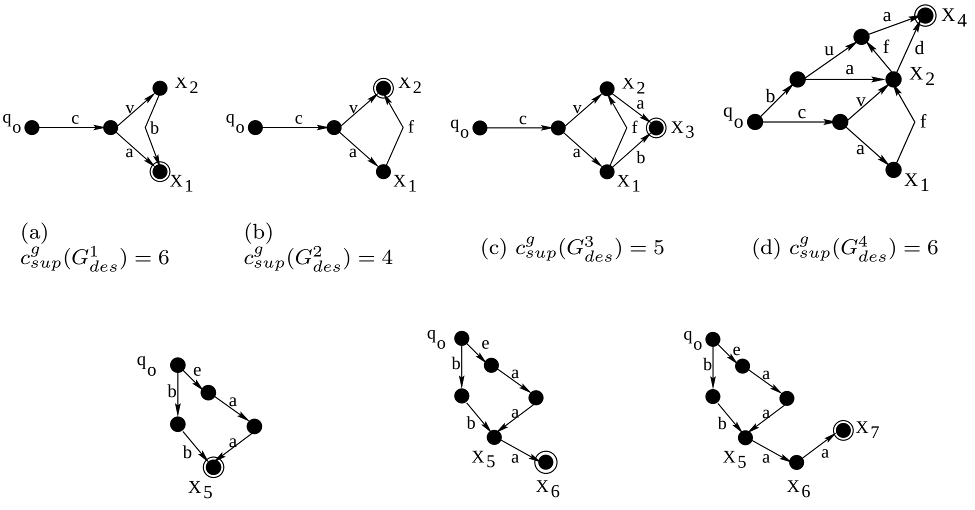
5.4.3. Predefined partial order for the visit of $\mathcal{X}$. Throughout section 5, we have assumed that we had no prespecified order in which to visit the marked states in $\mathcal{X}$. This may be the case in several applications. However, in other applications, such as test-generation, we may be interested in the path taken by a system more than in the final state it reaches. The designer may want to enforce the system to follow a given path. The path would be characterized by the states it traverses, which would be marked. This would yield an ordering, not on the marked states, but on the subpaths themselves. A possible extension of this predefined partial order assumption would be to consider our problem in a hierarchical setting in the same spirit as in [23, 22].

Let us consider a simple example. Assume that we have the marked states $\{X_1,\ldots,X_{10}\}$ to visit in an optimal way. If we do not prespecify the order in which they should be visited, the TSP will be solved on a 10 × 10 matrix. The designer may want to observe the behavior of the system when it visits states X_1 through X_3 in that order, X_4 through X_7 in that order, and X_7 through X_{10} in that order. This would reduce the TSP to a 3 × 3 matrix, abstracting away from the ten states to four macrostates: $\{q_0\}, \{X_1, X_2.X_3\}, \{X_4, X_5, X_6, X_7\}$, and $\{X_8, X_9, X_10\}$. The solution thus obtained will not be stepwise DP-optimal per se. It will be optimal given the additional constraints imposed by the designer.



Event	c_e	Remarks
a,f	1	Controllable
b,c,d,e	2	Controllable
g	4	Controllable
Reset	0	Controllable
u	3	Uncontrollable
v	2	Uncontrollable

FIG. 6.1. *The initial system G and the event cost function.*



(a) $c^g_{sup}(G^1_{des}) = 6$

(b) $c^g_{sup}(G^2_{des}) = 4$

(c) $c^g_{sup}(G^3_{des}) = 5$

(d) $c^g_{sup}(G^4_{des}) = 6$

(e) $c^g_{sup}(G^5_{des}) = 4$

(f) $c^g_{sup}(G^6_{des}) = 5$

(g) $c^g_{sup}(G^7_{des}) = 6$

FIG. 6.2. *The DP-optimal FSMs for the different state of $\mathcal{X}$.*

6. Example. The following example is constructed to illustrate the essential stages of the optimal control problem for multiple marked states to visit. For the sake of simplicity, we have assumed that all control costs have zero cost for controllable events and infinite costs for uncontrollable events. We here consider a system modeled by the FSM G, which represents its legal behavior. In this example, there are seven states denoted $(X_i)_{i\in[1,\dots,7]} = \mathcal{X}$ to visit in no particular order. Some costs are allocated to each of the events of the FSM G. The event costs and their status (controllable or not) are as depicted in Figure 6.1.

Note that in Figure 6.1, the *Reset* events are not represented but exist between each of the $(X_i)_{i\in[1,\dots,7]}$ and the initial state q_0. The first phase of the algorithm consists of computing the various DP-optimal submachines $(G^i_{des})_{i\in[1,\dots,7]}$ for each of the final states of $\mathcal{X}$. This part is performed using the **DP-Opt** algorithm (see Appendix A of [10]). The seven figures given next (Figure 6.2) correspond to the DP-optimal submachines for each of the final states $(X_i)_{i\in[1,\dots,7]}$. We also give the worst inevitable cost for each submachine $(G^i_{des})_{i\in[1,\dots,7]}$.

According to (5.1) in section 5.1, we obtain the matrix C, encoding the worst inevitable cost between two states X_i and X_j.

	q_0	X_1	X_2	X_3	X_4	X_5	X_6	X_7
q_0	∞	**6**	**4**	**5**	**5**	**4**	**5**	**6**
X_2	0	**2**	∞	**1**	**2**	∞	∞	∞
X_3	0	∞	∞	∞	∞	∞	∞	∞
X_4	0	∞	∞	∞	∞	∞	∞	∞
X_5	0	∞	∞	∞	∞	∞	**1**	**2**
X_6	0	∞	∞	∞	∞	∞	∞	**1**
X_7	0	∞	∞	∞	∞	∞	∞	∞

Following Proposition 5.4, we can see that $\mathcal{X}_1 = \{X_1, X_2, X_3, X_4\}$ and $\mathcal{X}_2 = \{X_5, X_6, X_7\}$ form a partition of the set of states $\mathcal{X} = (X_i)_{i\in[1,\dots,7]}$. Moreover, the states of $\mathcal{X}_2$ satisfy Proposition 5.5. Thus, in order to solve the TSP, we can now consider the two following matrices. Note that in the second one, we have replaced $C[0, 6]$ and $C[0, 7]$ by ∞ as stated by Proposition 5.5.

	q_0	X_1	X_2	X_3	X_4
q_0	∞	**6**	**4**	**5**	**5**
X_1	0	∞	**1**	**2**	**3**
X_2	0	**2**	∞	**1**	**2**
X_3	0	∞	∞	∞	∞
X_4	0	∞	∞	∞	∞

	q_0	X_5	X_6	X_7
q_0	∞	**4**	∞	∞
X_5	0	∞	**1**	**2**
X_6	0	∞	∞	**1**
X_7	0	∞	∞	∞

One solution (there are several) of the TSP for each submatrix is

$$\{(q_0, X_4); (X_4, X_1); (X_1, X_2); (X_2, X_3); (X_3, q_0)\},$$
$$\{(q_0, X_5); (X_5, X_6); (X_6, X_7); (X_7, q_0)\}.$$

This is the output of the TSP resolution method run on each of the subproblems. The optimal worst-case costs are 13 and 6, respectively. From Theorem (5.1), there exist two stepwise DP-optimal schedulers A_{0_1} and A_{0_2}. We need to retrieve them from the output of the resolution of the TSP and build a global stepwise DP-optimal scheduler A_0.

We look at each one of the pairs and see if they correspond to a DP-optimal submachine. In this case $(q_0, X_4), (X_1, X_2), (X_2, X_3)$, and (X_3, q_0) correspond to G^4_{des}, $M^o_D(G, X_1, X_2)$, $M^o_D(G, X_2, X_3)$, and $X_3 \overset{Reset}{\longrightarrow} q_0$, respectively. (X_4, X_1) does not have any associated DP-optimal submachine. We decompose it: (X_4, X_1) becomes $(X_4 \overset{Reset}{\longrightarrow} q_0)$ concatenated with G^1_{des}. An optimal DP-optimal scheduler A_{o_1} is the following:

$$A_{o_1} = G^4_{des} \circ (X_4 \overset{Reset}{\longrightarrow} q_0) \circ G^1_{des} \circ M^o_D(G, X_1, X_2) \circ M^o_D(G, X_2, X_3) \circ (X_3 \overset{Reset}{\longrightarrow} q_0).$$

For the second subproblem of the divide and conquer method, we also map the pairs to original DP-optimal submachines, yielding scheduler A_{o_2}.

$$A_{o_2} = G^5_{des} \circ M^o_D(G, X_5, X_6) \circ M^o_D(G, X_6, X_7) \circ (X_7 \overset{Reset}{\longrightarrow} q_0).$$

Note that we use exactly four DP-optimal submachines for the first subproblem and three for the second, as expected from Proposition 5.3. We finally generate the global stepwise DP-optimal scheduler A_o as the concatenation of the two subschedulers A_{o_1} and A_{o_2}, yielding the following scheduler, A_o.

$$\begin{aligned} A_o &= A_{o_1} \circ A_{o_2} \\ &= G^4_{des} \circ (X_4 \xrightarrow{Reset} q_0) \circ G^1_{des} \circ M^o_D(G, X_1, X_2) \circ M^o_D(G, X_2, X_3) \circ (X_3 \xrightarrow{Reset} q_0) \\ &\quad \circ G^5_{des} \circ M^o_D(G, X_5, X_6) \circ M^o_D(G, X_6, X_7) \circ (X_7 \xrightarrow{Reset} q_0). \end{aligned}$$

In fact, this scheduler is actually composed of three different nontrivial subschedulers. The stepwise DP-optimal A_o can be rewritten as

$$\begin{aligned} A_o &= G^4_{des} \circ (X_4 \xrightarrow{Reset} q_0) \circ A_1 \circ (X_3 \xrightarrow{Reset} q_0) \circ A_2 \circ (X_7 \xrightarrow{Reset} q_0). \\ & where \left\{ \begin{array}{lcl} A_1 &=& M^o_D(G, X_1, X_2) \circ M^o_D(G, X_2, X_3), \\ A_2 &=& G^5_{des} \circ M^o_D(G, X_5, X_6) \circ M^o_D(G, X_6, X_7). \end{array} \right. \end{aligned} \tag{6.1}$$

This last expression shows the minimum number of *Reset*s that are necessary to visit all the X_i in an optimal way (three, in this case).

7. Potential applications of the theory. Applications of the theory that we have elaborated cover various fields of engineering. One application that can be developed from the theory is test objective generation. In test objective generation, the goal is to check whether a particular system meets the expectations or the requirements that are associated with it. In this framework, the states of interest may be states in which the system is suspected to behave incoherently or incorrectly, or states in which misbehavior could be dramatic or dangerous. These would be the states that would be marked. The theory that we developed allows us to visit all these states and to test the behavior of the system in each one. Once the system has reached one of the marked states, all the known events can be disabled to check if the system stops or enters a forbidden state. A timeout can be set, for example. If the system has not behaved incoherently after that timeout, we can decide to pursue the visit of the marked states. Other more involved strategies can be applied to determine whether a state is faulty or not. For each state, either the behavior of the system is acceptable, or it is not. In the first case where the state is flawless, the next submachine of the scheduler is activated in order to make the system evolve in the next state of interest to be tested. In the case where a failure has been detected in the state, either we stop since the system is faulty and does not correspond to the awaited specifications, or we proceed to determine other possible faults. To do so, we *reset* the system to its initial state q_0, and go directly to the next state, say X_i, through its direct DP-optimal submachine, namely G^i_{des}, and the process continues.

Another application area is planning in the case of multiple goals in AI. Several search algorithms exist when one unique goal is sought (see part II of [18]). Planning in the case of multiple goals remains challenging and interesting. The framework in which we have developed the theory allows goals to be independent or related. Once again, the *Reset* event has an interesting interpretation in AI. It represents the impossibility to meet all the goals without returning to the initial state. It may represent the possibility of using several agents to achieve the goals, each running in parallel. The number of *Reset* events gives the necessary and sufficient number of agents that are needed to perform the goal of reaching all the subgoals in parallel, without any conflicts. These applications constitute interesting further work.

We give a last potential application example of our theory: routing in a communication network. In the same way that several agents can perform in parallel to achieve different tasks, in a communication network a message can be broadcast by generating multiple copies of it and sending these copies in parallel, along the stepwise DP-optimal paths. These paths are actually the stepwise DP-optimal schedulers seen as stepwise DP-optimal subnetworks. The marked states represent the agents to whom the messages are destined. The costs may be the energy consumed for each transmission between nodes. The uncontrollability of certain events may reflect the possibility of other agents changing the terminal path to certain nodes, based on their own view of the network.

Example 7.1. Consider the example of section 6 as a routing problem in a communication problem. Relation (6.1) (i.e., the solution of the stepwise DP-optimal problem) highlights the manner in which the information would need to be sent through the communication network. Given that there are exactly three *Reset* events, the sender should generate exactly three messages and send them in parallel. For each message, the sender can specify (in each header, for example) the desired route that each message should take, according to the sender's view of the network and calculations. (This routing is actually given by the corresponding stepwise DP-optimal subscheduler.) However, the uncontrollability is represented by the fact that other intermediate routing nodes may have a view of the network that is different from that of the sender, in which case the former might decide on a new route.

8. Conclusion. In this paper, we have introduced a new type of optimal control for DES. Previous work in optimal control deals with numerical performances in supervisory control theory when the goal to achieve is a unique state of interest. In contrast, our goal was to make the system evolve through a set of goals one by one, with no order necessarily specified a priori. The order in which the states are visited was part of the optimization problem since it had an influence on the cost of visiting all the goal states.

The system to be controlled is represented by an FSM with a set of multiple marked states $\mathcal{X} = (X_i)_{i\in[1,\dots,n]}$ representing the states of interest. Our aim was to have the system reach each and every one of the $(X_i)_{i\in[1,\dots,n]}$. To do so, we have introduced the notion of a scheduler. A scheduler can be thought of as a concatenation of submachines. The role of the scheduler is to make the system evolve according to one submachine at a time and account for switching between them at appropriate instants, i.e., when one of the states of interest has been reached. We have then introduced the notion of a stepwise DP-optimal scheduler of an FSM G with respect to the set $\mathcal{X}$. This particular type of scheduler is custom made given the system on which the optimization is to be run. It has the particularity of being composed of DP-optimal submachines which allow optimality from state of interest to state of interest (stepwise). Moreover, the ordering of these DP-optimal submachines allows global optimality in the sense that the total worst-case cost of visiting all the states of $\mathcal{X}$ is minimized.

We gave a necessary and sufficient condition for the existence of a stepwise DP-optimal scheduler, namely, the existence of n DP-optimal submachines between the initial state q_0 and each state of the n states of $\mathcal{X}$. This condition is not very restrictive, since if it does not hold, that means that one of the states is not reachable in a controllable manner, i.e., not surely reachable from the initial state q_0. In such a case, it is obvious that the state in question will never be reachable with a surely finite cost.

From a computational point of view, we showed that our optimal problem could be brought down to an instance of the TSP. The solution of this particular TSP gives a direct access to both the structure of a stepwise DP-optimal scheduler and the worst-case cost for visiting all the states of interest. Considering the high computational complexity of this step, we also gave ways of taking advantage of some particular properties of the structure of a stepwise DP-optimal scheduler, leading to the reduction of the computational complexity of the corresponding TSP without loss of global optimality.

Finally, besides the possible applications briefly presented in section 7, future work will most probably extend the theory to the case of a system where the events are partially observable.

Acknowledgment. The authors wish to thank the reviewers for their relevant comments.

REFERENCES

[1] E. Brinskma, *A theory for the derivation of tests*, Protocol Specification, Testing and verification, 7 (1988), pp. 63–74.

[2] C. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[3] J. Fernandez, C. Jard, T. Jéron, and C. Viho, *Experiment in automatic generation of test suites for protocols with verification technology*, Sci. Comput. Programming, 29 (1997), pp. 123–146.

[4] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, MA, 1979.

[5] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, *Planning and acting in partially observable stochastic domains*, Artificial Intelligence, 101 (1998), pp. 99–134.

[6] R. Kumar and V. Garg, *Optimal control of discrete event dynamical systems using network flow techniques*, in Proceedings of the 29th Allerton Conference on Communication, Control, and Computing, Champaign, IL, 1991, pp. 705–714.

[7] V. Kumar and L. N. Kanal, *A general branch and bound formulation for and/or graph and game tree search*, in Search in Artificial Intelligence, Springer-Verlag, New York, Berlin, 1988, pp. 91–130.

[8] E. L. Lawler, J. K. Lenstra, A. H. G. Rinooy Kan, and D. B. Shmoys, *The Traveling Salesman Problem*, John Wiley, New York, 1985.

[9] E. L. Lawler and D. E. Wood, *Branch-and-bound methods: A survey*, Oper. Res., 14 (1966), pp. 699–719.

[10] H. Marchand, O. Boivineau, and S. Lafortune, *On the Synthesis of Optimal Schedulers in Discrete Event Control Problems with Multiple Goals*, Tech. Report n° CGR-98-10, Control Group Reports, College of Engineering, University of Michigan, Ann Arbor, MI, 1998. Available via ftp. from ftp://ftp.eecs.umich.edu/techreports/systems/control_group/lafortune/.

[11] H. Marchand and M. Le Borgne, *On the optimal control of polynomial dynamical systems over z/pz*, in Proceedings of the 4th IEE International Workshop on Discrete Event Systems, Cagliari, Italy, 1998, pp. 385–390.

[12] K. Murty, *Operations Research: Deterministic Optimization Models*, Prentice-Hall, Upper Saddle River, N.J., 1995.

[13] D. J. Musliner, E. H. Durfee, and K. G. Shin, {*Circa*}*:* {*A*} *cooperative intelligent real time control architecture*, IEEE Trans. Systems, Man, and Cybernetics, 23 (1993), pp. 1561–1574.

[14] K. Passino and P. Antsaklis, *On the optimal control of discrete event systems*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 2713–2718.

[15] P. J. Ramadge and W. M. Wonham, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.

[16] P. J. G. Ramadge and W. M. Wonham, *The control of discrete event systems*, Proceedings of the IEEE, 77 (1989), pp. 81–98.

[17] A. Rouger and M. Phalippou, *Test cases generation from formal specifications*, in Proceedings of the ISS'92, Yokohama, Japan, 1992, p. C10.2.

[18] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice-Hall, Upper Saddle River, NJ, 1995.

[19] R. Sengupta and S. Lafortune, *A Deterministic Optimal Control Theory for Discrete Event Systems: Computational Results*, Tech. Report n° CGR-93-16, Control Group Reports, College of Engineering, University of Michigan, Ann Arbor, MI, 1993. Available via ftp. from ftp://ftp.eecs.umich.edu/techreports/systems/control_group/lafortune/.

[20] R. Sengupta and S. Lafortune *A Deterministic Optimal Control Theory for Discrete Event Systems: Formulation and Existence Theory*, Tech. Report n° CGR-93-7, Control Group Reports, College of Engineering, University of Michigan, Ann Arbor, MI, 1993. Available via ftp. from ftp://ftp.eecs.umich.edu/techreports/systems/control_group/lafortune/.

[21] R. Sengupta and S. Lafortune, *An optimal control theory for discrete event systems*, SIAM J. Control Optim., 36 (1998), pp. 488–541.

[22] G. Shen and P. Caines, *Control consistency and hierarchically accelerated dynamic programming*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 1686–1691.

[23] G. Shen, P. Caines, and P. Hubbard, *Control Consistency and Hierarchically Accelerated Dynamic Programming*, Tech. report, Department of Electrical Engieering, McGill University, Montreal, Quebec, Canada, 1997.

[24] E. Tronci, *Optimal state supervisory control*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 2237–2242.

SIAM J. Control Optim.
Vol. 39, No. 2, pp. 533–570

LAW OF THE ITERATED LOGARITHM FOR A CONSTANT-GAIN LINEAR STOCHASTIC GRADIENT ALGORITHM[*]

J. A. JOSLIN[†] and A. J. HEUNIS[‡]

Abstract. We study almost-sure limiting properties, taken as $\varepsilon \searrow 0$, of the finite horizon sequence of random estimates $\{\theta_0^\varepsilon, \theta_1^\varepsilon, \theta_2^\varepsilon, \dots, \theta_{\lfloor T/\varepsilon \rfloor}^\varepsilon\}$ for the linear stochastic gradient algorithm

$$\theta_{n+1}^\varepsilon = \theta_n^\varepsilon + \varepsilon \left[a_{n+1} - (\theta_n^\varepsilon)' X_{n+1} \right] X_{n+1}, \qquad \theta_0^\varepsilon \overset{\triangle}{=} \theta_* \quad \text{nonrandom},$$

where $T \in (0,\infty)$ is an arbitrary constant, $\varepsilon \in (0,1]$ is a (small) adaptation gain, and $\{a_n\}$ and $\{X_n\}$ are data sequences which drive the algorithm. These limiting properties are expressed in the form of a functional law of the iterated logarithm.

Key words. stochastic gradient algorithm, L-mixing processes, functional law of the iterated logarithm

AMS subject classifications. 60F15, 60F17, 93E10

PII. S0363012997331007

1. Introduction.

A commonly used stochastic gradient algorithm has the following structure:

$$\theta_{n+1}^\varepsilon = \theta_n^\varepsilon + \varepsilon \left[a_{n+1} - (\theta_n^\varepsilon)' X_{n+1} \right] X_{n+1}, \quad \theta_0^\varepsilon \overset{\triangle}{=} \theta_*, \quad \forall\, n = 0, 1, \dots \tag{1.1}$$

Here $\varepsilon \in (0,1]$ is a (small) constant called the *adaptation gain*; $\theta_* \in \Re^d$ is a fixed nonrandom *initial condition*; and $\{a_n,\ n = 1,2,\dots\}$ and $\{X_n,\ n = 1,2,\dots\}$ are, respectively, $\Re$ and $\Re^d$-valued *data processes* which "drive" the algorithm and in turn give rise to the $\Re^d$-valued process of *estimates* $\{\theta_n^\varepsilon,\ n = 0,1,2,\dots\}$.

Within this context one often wants to characterize asymptotic properties of the random sequence $\{\theta_0^\varepsilon, \theta_1^\varepsilon, \theta_2^\varepsilon, \dots, \theta_{\lfloor T\varepsilon^{-1} \rfloor}^\varepsilon\}$ as $\varepsilon \to 0$, where $T \in (0,\infty)$ is a fixed but arbitrary constant called the *horizon.* Such asymptotic properties deal with the *finite horizon* characteristics of (1.1). Perhaps the most basic of these asymptotic properties is given by the so-called *ODE method* which, in the present context, essentially says the following. Under reasonably general conditions on the processes $\{a_n\}$ and $\{X_n\}$, one can relate the sequence of estimates $\{\theta_n^\varepsilon\}$ to the solution $\{\theta^0(\tau),\ \tau \in [0,\infty)\}$ of the ODE

$$\dot{\theta^0}(\tau) = \bar{b} - \bar{R}\theta^0(\tau), \quad \theta^0(0) = \theta_*, \tag{1.2}$$

(where the d-vector $\bar{b}$ and $d \times d$-matrix $\bar{R}$ in (1.2) are typically given by conditions such as (C2) in section 2) by the following *finite horizon weak law of large numbers.* For each $\delta \in (0,\infty)$ and $T \in (0,\infty)$ we have

$$\lim_{\varepsilon \to 0} P \left[\max_{0 \le \tau \le T} |\theta_{\lfloor \tau/\varepsilon \rfloor}^\varepsilon - \theta^0(\tau)| > \delta \right] = 0 \tag{1.3}$$

[*] Received by the editors December 2, 1997; accepted for publication (in revised form) May 11, 2000; published electronically September 15, 2000. Research for this work was supported by NSERC of Canada.

http://www.siam.org/journals/sicon/39-2/33100.html

[†] Siemens Canada, Kanata, Ontario K2K 2M5, Canada (jjoslin@magma.ca).

[‡] Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (heunis@kingcong.uwaterloo.ca).

(see Kushner and Shwartz [19, Theorem 1, p. 178]), where $\lfloor x \rfloor$ denotes the integer part of $x \in [0, \infty)$. The ODE method is comprehensively covered in, for example, the books of Benveniste, Métivier, and Priouret [1], Kumar and Varaiya [17], and Kushner and Yin [18].

In classical probability, the law of large numbers is complemented by Donsker's functional central limit theorem (CLT), which has the following form. Suppose that $\{\xi_k,\ k = 1, 2, \ldots\}$ is an independent and identically distributed (i.i.d.) sequence of random variables with $E\xi_k = 0$, $E\xi_k^2 = 1$. Define the partial sums

$$(1.4) \qquad S_n \overset{\triangle}{=} \sum_{k=1}^{n} \xi_k \qquad \forall\, n = 1, 2, \ldots$$

and, for each $m = 1, 2, \ldots$, let $\{\Xi^m(\tau),\ \tau \in [0, \infty)\}$ be the continuous piecewise-linear process given by

$$(1.5) \qquad \Xi^m(\tau) \overset{\triangle}{=} \begin{cases} 0 & \text{if } \tau = 0, \\ m^{-\frac{1}{2}} S_n & \text{if } \tau = n/m, \quad \forall\, n = 1, 2, \ldots, \\ \text{linear interpolation,} & \text{otherwise.} \end{cases}$$

Then, for each $T \in (0, \infty)$, the process $\{\Xi^m(\tau),\ \tau \in [0, T]\}$ converges weakly to a standard Wiener process $\{W(\tau), \tau \in [0, T]\}$ as $m \to \infty$ (see Theorem 10.1 of Billingsley [2]). An analogous *finite horizon functional CLT* can be established for the random sequence $\{\theta_0^\varepsilon, \theta_1^\varepsilon, \theta_2^\varepsilon, \ldots, \theta_{\lfloor T\varepsilon^{-1} \rfloor}^\varepsilon\}$ obtained from (1.1). For each $\varepsilon \in (0, 1]$ define the $\Re^d$-valued continuous piecewise-linear process $\{\Theta^\varepsilon(\tau),\ \tau \in [0, \infty)\}$ by

$$(1.6) \qquad \Theta^\varepsilon(\tau) \overset{\triangle}{=} \begin{cases} \varepsilon^{-\frac{1}{2}} \left(\theta_{\tau/\varepsilon}^\varepsilon - \theta^0(\tau) \right), & \tau = k\varepsilon, \quad \forall\, k = 0, 1, 2, \ldots, \\ \text{linear interpolation,} & \text{otherwise.} \end{cases}$$

Then, subject to certain regularity conditions on the data sequences $\{a_n\}$ and $\{X_n\}$, for each $T \in (0, \infty)$ the process $\{\Theta^\varepsilon(\tau),\ \tau \in [0, T]\}$ converges weakly in $C[0, T]$ (the space of continuous functions from $[0, T]$ into $\Re^d$) to a limiting Gauss–Markov process $\{\hat{\Theta}(\tau),\ \tau \in [0, T]\}$ as $\varepsilon \to 0$. A precise formulation of this result, pertaining to a general class of algorithms which includes (1.1) as a special case and providing a complete characterization of the Gauss–Markov limit, may be found in [1, Theorem 1, p. 107] and [3, Theorem 2, p. 969]. The Gauss–Markov limit is also discussed further in Remark 3.4.

In the context of a sum of i.i.d. random variables $\{\xi_k,\ k = 1, 2, \ldots\}$ with $E\xi_k = 0$ and $E\xi_k^2 = 1$, Donsker's functional CLT is complemented by another basic result, namely, Strassen's *functional law of the iterated logarithm* (see Theorem 3 of [25]). To see the form of this result, fix some $T \in (0, \infty)$ and put

$$(1.7)\ K_\xi^T \overset{\triangle}{=} \left\{ \phi : [0, T] \to \Re\ : \phi(0) = 0,\ \phi(\cdot) \text{ abs. continuous, } \frac{1}{2} \int_0^T |\dot{\phi}(s)|^2\, ds \le 1 \right\}.$$

It is well known that K_ξ^T is a compact set of continuous functions (with the supremum norm of uniform convergence over $[0, T]$), and Strassen's functional law of the iterated logarithm (LIL) says the following. For P-almost all ω the sequence of continuous functions $\{\Xi^m(\tau, \omega)/\sqrt{2 \log \log m},\ \tau \in [0, T]\}$, indexed by $m = 1, 2, \ldots$, converges towards K_ξ^T as $m \to \infty$, and the set of its accumulation points coincides exactly with K_ξ^T, the sense of convergence being that of uniform convergence over $[0, T]$. This result—called by Williams [26, page 208] a "staggering generalization" of Kolmogorov's classical LIL—can be used as a tool for deriving many subtle fine-structure properties of the sample-paths of the partial-sum sequence $\{S_n\}$. For example, it can be used to show that, with probability one, the quantity

$$\frac{1}{N}\text{cardinality}\left(\left\{1 \le n \le N :\ S_n > \frac{1}{2}\sqrt{2n\log\log n}\right\}\right), \qquad N = 1, 2, \ldots,$$

exceeds 0.99999 for infinitely many values of N, whereas (again with probability one) the same quantity exceeds the slightly larger number 0.999999 for only finitely many N (see Strassen [25] for this as well as other examples of how one can use the functional LIL to analyze the partial sum process).

The preceding discussion suggests the problem of establishing a functional LIL for the random sequence $\{\theta_0^\varepsilon, \theta_1^\varepsilon, \theta_2^\varepsilon, \ldots, \theta_{\lfloor T\varepsilon^{-1}\rfloor}^\varepsilon\}$ arising from (1.1). This should have the same relation to the finite horizon functional CLT indicated previously as Strassen's functional LIL has to Donsker's functional CLT, and therefore should be of the following general form. For each fixed $T \in (0, \infty)$ there is a compact set $K_\Theta^T \subset C[0, T]$ (with a characterization analogous to that of K_ξ^T in (1.7)) and, for P-almost all ω, the family of continuous functions $\{\Theta^\varepsilon(\tau, \omega)/\sqrt{2\log\log\varepsilon^{-1}},\ \tau \in [0, T]\}$, indexed by $\varepsilon \in (0, 1]$, converges towards K_Θ^T (as $\varepsilon \to 0$) and the set of its $C[0, T]$-accumulation points coincides exactly with K_Θ^T. This is a *finite horizon functional LIL* for the sequence of estimates $\{\theta_0^\varepsilon, \theta_1^\varepsilon, \theta_2^\varepsilon, \ldots, \theta_{\lfloor T\varepsilon^{-1}\rfloor}^\varepsilon\}$, and our goal is to establish a result of this kind (see Theorem 3.3 to follow) subject to certain conditions on the data sequences $\{a_n\}$ and $\{X_n\}$ which are set forth in section 2. We choose to concentrate attention on the algorithm (1.1) because there seem to be major technical obstacles to getting this result for the more general classes of fixed-gain algorithms proposed, for example, in [1] and [3], whereas the linear structure of (1.1) simplifies matters considerably.

The usual method for establishing a functional LIL, pioneered by Strassen [25], is to first prove a so-called *strong invariance principle* for the partial-sum sequence $\{S_n\}$. Essentially, this says that one can always construct a Wiener process $\{W_t\}$ on the same probability space on which the S_n of (1.4) are defined (or perhaps on some extension of this space) such that

$$S_{\lfloor t\rfloor} - W_t = o(\sqrt{t\log\log t}) \quad (t \to \infty) \quad \text{almost surely (a.s.).} \tag{1.8}$$

Then one uses (1.8), together with known sample-path properties of the Wiener process, as the basis for establishing Strassen's functional LIL. In the context of stochastic algorithms with *decreasing gain* one can follow a similar approach (see [12] and [24]) but for *constant-gain* algorithms, such as (1.1), it is not at all obvious how to formulate and prove an analogue of (1.8). Accordingly, we shall adopt a different approach, in the spirit of a method pioneered by Chover [5], who showed how to establish Strassen's functional LIL using a CLT with *rate of convergence* in place of the strong invariance principle (1.8). This general approach, which was made to work in [15] for the stochastic averaging principle, will be extended here to work for the algorithm (1.1).

The organization of the paper is as follows: in section 2 we state and discuss conditions on the sequences $\{a_n\}$ and $\{X_n\}$ which drive the algorithm (1.1). In section 3 we establish the main result, namely Theorem 3.3. The proof in section 3 relies on two key technical results, namely an auxiliary LIL (see Theorem 3.5) and an a.s. approximation theorem (see Theorem 3.6). These are established in sections 4 and 5, respectively. Another essential technical result, needed for the proofs in section 4, is a functional CLT with rate of convergence (see Theorem 4.4), and this is proved in section 6. In section 7 we develop a miscellany of subsidiary technical lemmas which are used in the previous sections. In section 8 we restate, in a form best suited to our needs, some results from probability theory which are used in section 3 to section 7. Finally, we define the notation at the beginning of each section where it is first needed.

2. Conditions.

Notation 2.1. We use the following notation: $\Re^d$, $\Re^{d\times r}$ denote the usual vector spaces of real d-dimensional column vectors and real d by r matrices, respectively, with vector norm $|x| \stackrel{\triangle}{=} (\sum_{i=1}^d x_i^2)^{1/2}$ for all $x \in \Re^d$, and matrix operator norm $|A| \stackrel{\triangle}{=} \max_{x\in\Re^r,\ |x|=1} |Ax|$ for all $A \in \Re^{d\times r}$. Write $(B)'$ for the transpose of a matrix B. For an $\Re^d$ or $\Re^{d\times r}$-valued random element X and $p \in [1,\infty)$, put $\|X\|_p \stackrel{\triangle}{=} (E[|X|^p])^{1/p}$. For $x \in [0,\infty)$, $\lfloor x \rfloor$ is the largest integer n such that $n \le x$.

The data sequences $\{a_n,\ n = 1,2,\ldots\}$ and $\{X_n,\ n = 1,2,\ldots\}$ driving the algorithm (1.1) will always be special instances of the class of L-mixing processes introduced by Gerencsér [9], and formulated in a discrete-parameter setting as follows.

Definition 2.2. *Suppose that* $\{\mathcal{F}_n,\ n = 1,2,\ldots\}$ *and* $\{\mathcal{F}_n^+,\ n = 1,2,\ldots\}$ *are sequences of sub-σ-algebras in the probability triple* $(\Omega,\mathcal{F},P)$*, increasing and decreasing, respectively, with* $\mathcal{F}_n$ *and* $\mathcal{F}_n^+$ *independent for each* $n = 1,2,\ldots.$ *An* $\Re^{d\times r}$*-valued random process* $\{z_n,\ n = 1,2,\ldots\}$ *on* $(\Omega,\mathcal{F},P)$ *is L-mixing with respect to the system* $(\mathcal{F}_n,\mathcal{F}_n^+)$ *when* (i) $\{z_n\}$ *is* $\{\mathcal{F}_n\}$*-adapted,* (ii) *for each* $p \in [1,\infty)$ *we have* $\sup_n \|z_n\|_p < \infty$*, and, for*

$$\gamma_p(s) \stackrel{\triangle}{=} \sup_{n>s} \|z_n - E\left[z_n \mid \mathcal{F}_{n-s}^+\right]\|_p \quad \forall\, s = 1,2,\ldots, \quad \textit{we have} \quad \sum_{1\le s<\infty} \gamma_p(s) < \infty.$$

We shall require the following strengthened notion of L-mixing.

Definition 2.3. *Suppose that* $\{\mathcal{F}_n,\ n = 1,2,\ldots\}$ *and* $\{\mathcal{F}_n^+,\ n = 1,2,\ldots\}$ *are sequences of sub-σ-algebras in the probability triple* $(\Omega,\mathcal{F},P)$*, as in Definition* 2.2. *An* $\Re^{d\times r}$*-valued random process* $\{z_n,\ n = 1,2,\ldots\}$ *on* $(\Omega,\mathcal{F},P)$ *is geometrically L-mixing with respect to the system* $(\mathcal{F}_n,\mathcal{F}_n^+)$*, when* (i) $\{z_n\}$ *is* $\{\mathcal{F}_n\}$*-adapted,* (ii) $\sup_n \|z_n\|_p < \infty$ *for each* $p \in [0,\infty)$*,* (iii) *there is a constant* $\lambda \in (0,1)$ *and, for each* $p \in [1,\infty)$*, a constant* $C_p \in [0,\infty)$ *such that*

$$\sup_{n>s} \|z_n - E\left[z_n \mid \mathcal{F}_{n-s}^n\right]\|_p \le C_p\, \lambda^s \qquad \forall\, s = 1,2,\ldots,$$

where

$$\mathcal{F}_m^n \stackrel{\triangle}{=} \mathcal{F}_n \cap \mathcal{F}_m^+, \qquad \textit{when} \qquad 1 \le m < n. \tag{2.1}$$

(Thus, $\mathcal{F}_m^n$ *is the collection of all events which are members of both* $\mathcal{F}_n$ *and* $\mathcal{F}_m^+$*.) The constant* $\lambda \in (0,1)$ *is called a rate of the geometrically L-mixing process.*

Remark 2.4. Notice that conditioning is on the σ-algebra $\mathcal{F}_{n-s}^+$ in Definition 2.2, whereas it is on the smaller σ-algebra $\mathcal{F}_{n-s}^n$ in Definition 2.3 (see Remark 2.5 for more discussion on this). Using the elementary inequality $\|z - E[z \mid \mathcal{H}]\|_p \le 2\|z - E[z \mid \mathcal{G}]\|_p$, which holds for $z \in L_p(\Omega,\mathcal{F},P)$, $p \in [1,\infty)$, and sub-σ-algebras $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$, one sees immediately that a geometrically L-mixing process with respect to a given system $(\mathcal{F}_n, \mathcal{F}_n^+)$ is also L-mixing with respect to the system $(\mathcal{F}_n, \mathcal{F}_n^+)$. Besides being motivated by L-mixing, the notion of geometric L-mixing in Definition 2.3 is also closely related to that of *exponential stability* introduced by Ljung [20] and Ljung and Caines [21] (see also Caines [4, p. 488] and Davis and Vinter [6, p. 217]). An important instance of a geometrically L-mixing process is the output of a stable finite-dimensional linear system with the form

$$x_{n+1} = Ax_n + Be_n, \qquad z_n = Cx_n + De_n, \qquad n = 1, 2, \ldots,$$

where $\{e_1, e_2, e_3, \ldots\}$ is a "driving" sequence of independent random vectors, and x_1 and $\sigma\{e_1, e_2, \ldots\}$ are independent, with $\sup_n \|e_n\|_p < \infty$ and $\|x_1\|_p < \infty$ for each $p \in [1, \infty)$. Here one defines

$$\text{(2.2)} \quad \mathcal{F}_n \overset{\triangle}{=} \sigma\{x_1, e_1, e_2, \ldots, e_n\}, \qquad \mathcal{F}_n^+ \overset{\triangle}{=} \sigma\{e_{n+1}, e_{n+2}, \ldots\} \qquad \forall\, n = 1, 2, \ldots,$$

and, by the argument in [4, pages 488-489] it is easily shown that $\{z_n\}$ is geometrically L-mixing with respect to the system $(\mathcal{F}_n, \mathcal{F}_n^+)$. It is this close link to stable linear systems that makes L-mixing a very appropriate model for signals in data communication problems. The L^p-bounds established for L-mixing in [9] and [10] also render this class of processes extremely tractable, and will be used frequently in the arguments that follow.

Remark 2.5. Suppose that $\{\mathcal{F}_n,\ n = 1, 2, \ldots\}$ and $\{\mathcal{F}_n^+,\ n = 1, 2, \ldots\}$ are sequences of sub-σ-algebras in the probability triple $(\Omega, \mathcal{F}, P)$, as in Definitions 2.2 and 2.3, and $\{z_n,\ n = 1, 2, \ldots\}$ is some $\Re^{d\times r}$-valued and geometrically L-mixing process with respect to the system $(\mathcal{F}_n, \mathcal{F}_n^+)$. For each $s = 1, 2, \ldots$, define the process $\{z_n[s], n = 1, 2, \ldots\}$ by

$$\text{(2.3)} \qquad z_n[s] \overset{\triangle}{=} E\left[z_n \mid \mathcal{F}_{n-s}^n\right] \qquad \forall\, n = 1, 2, \ldots,$$

where we put $\mathcal{F}_{n-s}^n \overset{\triangle}{=} \mathcal{F}_1^n$ when $n \leq s$. Observe that, for each $s = 1, 2, \ldots$, the process $\{z_n[s],\ n = 1, 2, \ldots\}$ is s-dependent, since, from (2.3), one has $\sigma\{z_1[s], z_2[s], \ldots, z_m[s]\} \subset \mathcal{F}_m$, while $\sigma\{z_n[s], z_{n+1}[s], \ldots, z_{n+k}[s]\} \subset \mathcal{F}_{n-s}^+$, and the σ-algebras $\mathcal{F}_m$ and $\mathcal{F}_{n-s}^+$ are clearly independent when $n - m > s$. Thus, we see from Definition 2.3 that a geometrically L-mixing process $\{z_n\}$ can be nicely approximated by the s-dependent process $\{z_n[s]\}$. This approximation property was used by Ljung and Caines [21] to establish asymptotic normality in off-line system identification, and will likewise be necessary for proving the main results of this work.

From now on we shall always suppose that the data processes $\{a_n,\ n = 1, 2, \ldots\}$ and $\{X_n,\ n = 1, 2, \ldots\}$, which drive the recursion (1.1), are $\Re$ and $\Re^d$-valued, respectively, defined on a common probability triple $(\Omega, \mathcal{F}, P)$, and subject to the following conditions (C1) to (C4).

(C1) There are sequences $\{\mathcal{F}_n,\ n = 1, 2, \ldots\}$ and $\{\mathcal{F}_n^+,\ n = 1, 2, \ldots\}$ of sub-σ-algebras, as in Definition 2.2, such that $\{a_n\}$ and $\{X_n\}$ are geometrically L-mixing with respect to the system $(\mathcal{F}_n, \mathcal{F}_n^+)$.

(C2) There exists a d-vector $\bar{b}$ and a $d\times d$ matrix $\bar{R}$ such that the limits

$$\text{(2.4)} \quad \bar{b} \overset{\triangle}{=} \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=n_0}^{N+n_0} E[a_n X_n]\,, \quad \bar{R} \overset{\triangle}{=} \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=n_0}^{N+n_0} E[X_n (X_n)']$$

exist uniformly with respect to $n_0 = 1, 2, \ldots$.

To state the remaining conditions we need the following notation:

$$(2.5)\qquad \begin{cases} b_n \stackrel{\triangle}{=} a_n X_n, & \bar{b}_n \stackrel{\triangle}{=} E[b_n]\,, \quad \tilde{b}_n \stackrel{\triangle}{=} b_n - \bar{b}_n, \quad \hat{b}_n \stackrel{\triangle}{=} \bar{b}_n - \bar{b}, \\ R_n \stackrel{\triangle}{=} X_n(X_n)', & \bar{R}_n \stackrel{\triangle}{=} E[R_n]\,, \quad \tilde{R}_n \stackrel{\triangle}{=} R_n - \bar{R}_n, \quad \hat{R}_n \stackrel{\triangle}{=} \bar{R}_n - \bar{R}. \end{cases}$$

Also, recall that $\{\theta^0(\tau),\ \tau \in [0,\infty)\}$ is the (unique) solution of (1.2). The remaining conditions are as follows.

(C3a) For each $T \in (0,\infty)$ there exist constants $C_1(T) \in [0,\infty)$ and $\varepsilon(T) \in (0,1]$ such that

$$\max_{0 \le k \le 1+\lfloor T\varepsilon^{-1}\rfloor} \left| \sum_{j=0}^{k-1} (\hat{b}_{j+1} - \hat{R}_{j+1}\theta^0(\varepsilon j)) \right| \le C_1(T)$$

for each $\varepsilon \in (0, \varepsilon(T)]$.

(C3b) There exist constants $\varepsilon_0 \in (0,1]$, $\alpha \in [0,3/4)$, and $C_2 \in [0,\infty)$ such that

$$\left| \sum_{j=k+1}^{k+N} \hat{R}_{j+1}(I - \varepsilon\bar{R})^j \right| \le C_2 N^\alpha$$

for all $\varepsilon \in (0,\varepsilon_0]$ and all $k, N = 1,2,\ldots$.

(C4) There is a constant $C_3 \in [0,\infty)$ and a function $A : \Re^d \to \Re^{d\times d}$ such that $A(\theta)$ is symmetric positive definite for each $\theta \in \Re^d$, and

$$(2.6)\qquad \left| \frac{1}{N+1}\operatorname{cov}\left(\sum_{n=n_0}^{N+n_0} \tilde{H}_n(\theta) \right) - A(\theta) \right| \le \frac{C_3[1+|\theta|^2]}{N+1}$$

for all $\theta \in \Re^d$ and $N, n_0 = 1,2,3,\ldots$, where

$$(2.7)\qquad \tilde{H}_n(\theta) \stackrel{\triangle}{=} \tilde{b}_n - \tilde{R}_n\theta \quad \forall\, \theta \in \Re^d, \quad \forall\, n = 1,2,\ldots.$$

Remark 2.6. Condition (C2) defines the d-vector $\bar{b}$ and the $d \times d$-matrix $\bar{R}$, which then gives the right side of the ODE (1.2). Conditions (C2), (C3a), and (C3b) control the "amount of nonstationarity" in the data processes $\{a_n\}$ and $\{X_n\}$. (C3a) is the same as the first of the conditions appearing in (3.4) of Khas'minskii [13] but is just rewritten in the context of algorithm (1.1) and plays a role similar to that of its counterpart in [13], while (C3b) is a mild condition which limits fluctuations of $\bar{R}_n \stackrel{\triangle}{=} E[X_n(X_n)']$ about $\bar{R}$ defined in (C2).

Remark 2.7. Suppose the $\Re^{d\times r}$-valued process $\{z_n,\ n = 1,2,\ldots\}$ is geometrically L-mixing with respect to a system $(\mathcal{F}_n, \mathcal{F}_n^+)$. Then it follows at once that the centralized process $\{z_n - Ez_n,\ n = 1,2,\ldots\}$ is geometrically L-mixing with respect to the system $(\mathcal{F}_n, \mathcal{F}_n^+)$. Moreover, if the $\Re^{r\times q}$-valued process $\{y_n,\ n = 1,2,\ldots\}$ is also geometrically L-mixing with respect to the system $(\mathcal{F}_n, \mathcal{F}_n^+)$, then it is easily verified using Definition 2.3 that the $\Re^{d\times q}$-valued process $\{z_n y_n,\ n = 1,2,\ldots\}$ is geometrically L-mixing with respect to system $(\mathcal{F}_n, \mathcal{F}_n^+)$. In view of these observations and condition (C1), it follows that the processes $\{\tilde{b}_n,\ n = 1,2,\ldots\}$ and $\{\tilde{R}_n,\ n = 1,2,\ldots\}$ given by (2.5) are zero-mean and geometrically L-mixing with respect to the system $(\mathcal{F}_n, \mathcal{F}_n^+)$, and $\{\tilde{H}_n(\theta),\ n = 1,2,\ldots\}$ given by (2.7) is zero-mean and geometrically L-mixing for each $\theta \in \Re^d$.

Remark 2.8. Condition (C4) gives the function $A(\cdot)$ which will be used to formulate the limiting set K_Θ^T in the main result (see Theorem 3.3). In some special cases it is possible to give explicit formulae for $A(\theta)$. For example, suppose that the driving data $\{a_n, X_n\}$ is weakly stationary so that, for each θ, we have $E[\tilde{H}_1(\theta)] = E[\tilde{H}_{1+r}(\theta)]$ and $E[\tilde{H}_1(\theta)\tilde{H}_n'(\theta)] = E[\tilde{H}_{1+r}(\theta)\tilde{H}_{n+r}'(\theta)]$ for all $r, n = 1, 2, \ldots$. Then it is easily shown that (C4) follows from (C1) with $A(\theta)$ given by

$$A(\theta) \overset{\triangle}{=} E[\tilde{H}_1(\theta)\tilde{H}_1'(\theta)] + \sum_{n=2}^{\infty} E[\tilde{H}_1(\theta)\tilde{H}_n'(\theta)] + \sum_{n=2}^{\infty} E[\tilde{H}_n(\theta)\tilde{H}_1'(\theta)].$$

For another example, suppose that the driving data is cyclostationary. This implies that there is a positive integer constant P such that, for each θ, we have $E[\tilde{H}_n(\theta)] = E[\tilde{H}_{n+P}(\theta)]$ and, for $\Lambda(\theta, m, n) \overset{\triangle}{=} E[\tilde{H}_m(\theta)\tilde{H}_n'(\theta)]$, we have $\Lambda(\theta, m, n) = \Lambda(\theta, m + P, n + P)$ for all $m, n = 1, 2, \ldots$. With the help of the periodicity relation we can extend $\Lambda(\theta, m, n)$ to all integers $-\infty < m, n < \infty$. By a straightforward adaptation of the argument on page 222 of [13] or page 76 of [14] it may be shown that (C4) follows from (C1) with $A(\theta)$ given by

$$A(\theta) = \frac{1}{P} \sum_{m=1}^{P} \sum_{n=-\infty}^{\infty} \Lambda(\theta, m, n).$$

A property of the mapping $A(\theta)$ in (C4) that will soon be needed is the following lemma.

Lemma 2.9 (proved in section 7). *Suppose conditions* (C1) *and* (C4). *Then the function* $A(\cdot)$ *given by* (2.6) *is locally Lipschitz continuous over* $\Re^d$.

3. The main result.

Notation 3.1. For the results of this and later sections we need the following additional notation. $C[0,T]$ indicates the space of continuous functions $f : [0,T] \to \Re^d$, for some $T \in (0, \infty)$, with norm $\|\cdot\|_C$ defined by $\|f\|_C \overset{\triangle}{=} \sup_{0 \le \tau \le T} |f(\tau)|$, and $AC_0[0,T] \overset{\triangle}{=} \{\psi \in C[0,T] : \ \psi(0) = 0, \ \psi(\cdot) \text{ abs. continuous}\}$. Also, for $x \in C[0,T]$ and $K \subset C[0,T]$, put $\|x - K\|_C \overset{\triangle}{=} \inf_{y \in K} \|x - y\|_C$. If $\{x_m, \ m = 1, 2 \ldots\}$ is a sequence in $C[0,T]$, then $\mathrm{acc}\{x_m\}$ denotes the set of its accumulation points in $C[0,T]$ (if any), and the notation $\{x_m(\tau), \ \tau \in [0,T]\} \twoheadrightarrow K$ for some set $K \subset C[0,T]$ means that (i) $\lim_{m\to\infty} \|x_m - K\|_C = 0$ and (ii) $\mathrm{acc}\{x_m\} = K$. The $\twoheadrightarrow$ symbol (due to Kuelbs [16]) provides a succinct language for expressing Strassen's functional LIL formulated in section 1, namely, for each $T \in (0, T)$ and processes $\{\Xi^m(\tau), \ \tau \in [0,T]\}$ given by (1.5), we have

$$\left\{ \frac{\Xi^m(\tau)}{\sqrt{2 \log\log m}}, \ \tau \in [0,T] \right\} \twoheadrightarrow K_\xi^T \qquad \text{a.s.} \tag{3.1}$$

For the main result of this paper, we must slightly extend this symbolism. If $x_\varepsilon \in C[0,T]$ for all $\varepsilon \in (0,1]$, then $\mathrm{acc}\{x_\varepsilon\}$ denotes the set of accumulation points in $C[0,T]$ (if any) as $\varepsilon \searrow 0$, and the notation $\{x_\varepsilon(\tau), \ \tau \in [0,T]\} \twoheadrightarrow K$ means that (i) $\lim_{\varepsilon \to 0} \|x_\varepsilon - K\|_C = 0$ and (ii) $\mathrm{acc}\{x_\varepsilon\} = K$.

Remark 3.2. From now on we fix an arbitrary finite horizon $T \in (0, \infty)$. Using the $\twoheadrightarrow$ notation of the preceding paragraph, our main result is the following LIL for algorithm (1.1), which characterizes a.s. limiting properties of $\{\Theta^\varepsilon(\tau), \ \tau \in [0,T]\}$, the restriction to $[0,T]$ of the process Θ^ε defined by (1.6).

Theorem 3.3. *Suppose conditions* (C1)–(C4) *of section* 2*, fix some finite horizon* $T \in (0,\infty)$*, and let* $\bar{R}$*,* $A(\cdot)$*,* $\{\theta^0(\tau),\ \tau \in [0,T]\}$ *and* $\{\Theta^\varepsilon(\tau),\ \tau \in [0,T]\}$*, be defined by* (2.4), (2.6), (1.2), and (1.6)*, respectively. Define the mapping* $I_\Theta^T : AC_0[0,T] \to [0,\infty]$ *by*

$$I_\Theta^T(\phi) \overset{\triangle}{=} \frac{1}{2}\int_0^T (\dot{\phi}(s) + \bar{R}\phi(s))' A^{-1}(\theta^0(s))(\dot{\phi}(s) + \bar{R}\phi(s))\, ds, \tag{3.2}$$

and let

$$K_\Theta^T \overset{\triangle}{=} \{\phi \in AC_0[0,T] \colon I_\Theta^T(\phi) \le 1\}. \tag{3.3}$$

Then K_Θ^T *is a compact subset of* $C[0,T]$*, and we have*

$$\left\{ \frac{\Theta^\varepsilon(\tau)}{\sqrt{2\log\log\varepsilon^{-1}}},\ \tau \in [0,T] \right\} \longrightarrow\!\!\!\!\rightarrow K_\Theta^T \quad a.s. \tag{3.4}$$

Remark 3.4. To motivate the proof of Theorem 3.3 we briefly recall how one can establish the functional CLT giving weak convergence of $\{\Theta^\varepsilon(\tau),\ \tau \in [0,T]\}$ to some limiting Gauss–Markov process $\{\hat{\Theta}(\tau),\ \tau \in [0,T]\}$ (see section 1). For each $\varepsilon \in (0,1]$ define the piecewise-linear continuous process $\{W^\varepsilon(\tau),\ \tau \in [0,\infty)\}$ by $W^\varepsilon(0) \overset{\triangle}{=} 0$, and

$$W^\varepsilon(\tau) \overset{\triangle}{=} \begin{cases} \varepsilon^{\frac{1}{2}} \sum_{j=1}^k \tilde{b}_j - \tilde{R}_j\theta^0((j-1)\varepsilon) & \forall\, \tau = k\varepsilon,\ \ k = 1,2,\dots, \\ \text{linear interpolation,} & \text{otherwise.} \end{cases} \tag{3.5}$$

Fix some arbitrary $T \in (0,\infty)$ as in Remark 3.2. For $\varepsilon \in (0,1]$, define the mapping $G^\varepsilon : C[0,T] \to C[0,T]$ as follows. For every $w \in C[0,T]$, let $G^\varepsilon(w)$ be given by the solution $v \in C[0,T]$ of the recursion

$$v(\tau) = \begin{cases} w(0), & \text{if } \tau = 0, \\ w(\tau) - \varepsilon \sum_{j=0}^{k-1} \bar{R}\, v(\varepsilon j), & \text{if } \tau = k\varepsilon,\ \ k = 1,2,\dots, \\ \text{linear interpolation,} & \text{otherwise.} \end{cases} \tag{3.6}$$

By a detailed analysis of the recursion (1.1) (which is not given here) we can show

$$\Theta^\varepsilon(\cdot) \approx G^\varepsilon(W^\varepsilon)(\cdot), \tag{3.7}$$

in the sense that $\{\Theta^\varepsilon(\tau),\ \tau \in [0,T]\}$ and $\{G^\varepsilon(W^\varepsilon)(\tau),\ \tau \in [0,T]\}$ have approximately the same distributions in $C[0,T]$ for small $\varepsilon \in (0,1]$. This suggests that we can get a weak limit for $\{\Theta^\varepsilon(\tau), \tau \in [0,T]\}$ when we establish a weak limit for $\{W^\varepsilon(\tau),\ \tau \in [0,T]\}$ and show that G^ε converges suitably to some limiting mapping $G : C[0,T] \to C[0,T]$ as $\varepsilon \to 0$. Indeed, using the fact that $\{\tilde{b}_j\}$ and $\{\tilde{R}_j\}$ are geometrically L-mixing (see Remark 2.7) and trivially modifying the arguments of [1, pp. 105–106], it can be proved that $\{W^\varepsilon(\tau),\ \tau \in [0,T]\}$ converges weakly in $C[0,T]$ to the restriction to $[0,T]$ of the Gauss–Markov process $\{\hat{W}^0(\tau),\ \tau \in [0,\infty)\}$ given by

$$\hat{W}^0(\tau) \overset{\triangle}{=} \int_0^\tau A^{\frac{1}{2}}(\theta^0(s))\, d\hat{B}(s) \qquad \forall\, \tau \in [0,\infty), \tag{3.8}$$

for some standard $\Re^d$-valued Brownian motion $\hat{B}(\cdot)$ on a probability triple $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$. In addition, it is easily seen from elementary analysis that $G^\varepsilon(\cdot)$ converges uniformly on compact subsets of $C[0,T]$ to the mapping $G : C[0,T] \to C[0,T]$ defined as follows.

For every $w \in C[0,T]$, let $G(w)$ be given by the solution $v \in C[0,T]$ of the linear integral equation

$$v(\tau) = w(\tau) - \int_0^\tau \bar{R}\, v(s)\, ds \quad \forall\, \tau \in [0,T]. \tag{3.9}$$

These facts suggest that $\{G^\varepsilon(W^\varepsilon)(\tau),\ \tau \in [0,T]\}$ converges weakly to the limit $\{G(\hat{W}^0)(\tau),\ \tau \in [0,T]\}$ and thus, in view of (3.7), that the weak limit of $\{\Theta^\varepsilon(\tau),\ \tau \in [0,T]\}$ is likewise the process $\{G(\hat{W}^0)(\tau),\ \tau \in [0,T]\}$ or, equivalently, the process given by the solution of the stochastic differential equation

$$d\hat{\Theta}(\tau) = d\hat{W}^0(\tau) - \bar{R}\hat{\Theta}(\tau)\, d\tau, \quad \hat{\Theta}(0) = 0, \quad \forall\, \tau \in [0,T]. \tag{3.10}$$

In short, one first proves a CLT giving weak convergence of the simpler process $\{W^\varepsilon(\tau),\ \tau \in [0,T]\}$ to the limit $\{\hat{W}^0(\tau),\ \tau \in [0,T]\}$ defined by (3.8), then uses the approximation (3.7) to transfer this result into a CLT giving weak convergence of $\{\Theta^\varepsilon(\tau),\ \tau \in [0,T]\}$ to the limit $\{\hat{\Theta}(\tau),\ \tau \in [0,T]\}$ defined by (3.10). Our strategy for establishing Theorem 3.3 will be based on a very analogous method. First establish a functional LIL for the process $\{W^\varepsilon(\tau),\ \tau \in [0,T]\}$ as follows.

THEOREM 3.5 (proved in section 4). *Suppose conditions* (C1)–(C4) *of section* 2, *and fix some finite horizon* $T \in (0,\infty)$. *Let* $\{\theta^0(\tau),\ \tau \in [0,T]\}$, $\{W^\varepsilon(\tau),\ \tau \in [0,T]\}$, *and* $A(\cdot)$ *be defined by* (1.2), (3.5), and (2.6), *respectively, and write*

$$K_W^T \triangleq \left\{ \phi \in AC_0[0,T] :\ \frac{1}{2}\int_0^T (\dot{\phi}(s))' A^{-1}(\theta^0(s))\dot{\phi}(s)\, ds \le 1 \right\}. \tag{3.11}$$

Then K_W^T *is a compact subset of* $C[0,T]$, *and*

$$\left\{ \frac{W^\varepsilon(\tau)}{\sqrt{2\log\log\varepsilon^{-1}}},\ \tau \in [0,T] \right\} \twoheadrightarrow K_W^T \quad a.s. \tag{3.12}$$

It remains to transfer the LIL of Theorem 3.5 for the process $\{W^\varepsilon(\tau),\ \tau \in [0,T]\}$ into the one given by Theorem 3.3 for the process $\{\Theta^\varepsilon(\tau),\ \tau \in [0,T]\}$. To this end, the following almost-sure version of (3.7) is essential.

THEOREM 3.6 (proved in section 5). *Suppose conditions* (C1)–(C3) *of section* 2, *and fix some finite horizon* $T \in (0,\infty)$. *Then, for the process* $\{W^\varepsilon(\tau),\ \tau \in [0,T]\}$ *and mapping* $G^\varepsilon : C[0,T] \to G^\varepsilon$ *defined by* (3.5) *and* (3.6), *respectively, we have*

$$\lim_{\varepsilon \searrow 0} \|G^\varepsilon(W^\varepsilon) - \Theta^\varepsilon\|_C = 0 \ \ a.s. \tag{3.13}$$

With Theorems 3.5 and 3.6 available, the proof of the main result is easy.

Proof of Theorem 3.3. One sees from (3.6) and (3.9) that $G^\varepsilon(\cdot)$ and $G(\cdot)$ are linear and continuous on $C[0,T]$. Using the Arzela–Ascoli theorem, it is easily seen that $G^\varepsilon(\cdot) \to G(\cdot)$ uniformly on compact subsets of $C[0,T]$ as $\varepsilon \searrow 0$. In view of this fact, if $Y^\varepsilon \in C[0,T]$ for all $\varepsilon \in (0,1]$ is such that $\{Y^\varepsilon(\tau),\ \tau \in [0,T]\} \twoheadrightarrow K$ for some compact $K \subset C[0,T]$, then it follows by easy analysis that $\{G^\varepsilon(Y^\varepsilon),\ \tau \in [0,T]\} \twoheadrightarrow G(K)$. Thus, identifying $Y^\varepsilon(\tau)$ with $W^\varepsilon(\tau)/\sqrt{2\log\log\varepsilon^{-1}}$ and using Theorem 3.5, we get

$$\left\{ \frac{G^\varepsilon(W^\varepsilon)(\tau)}{\sqrt{2\log\log\varepsilon^{-1}}},\ \tau \in [0,T] \right\} \twoheadrightarrow G(K_W^T), \quad \text{a.s.} \tag{3.14}$$

Next, from the definitions of K_Θ^T and K_W^T (see (3.3), (3.11)), and the definition of $G(\cdot)$ given by (3.9), we see that

$$G(K_W^T) = K_\Theta^T. \tag{3.15}$$

Now K_W^T is compact (by Theorem 3.5), thus K_Θ^T is compact, and (3.4) follows from (3.14), (3.15), and Theorem 3.6. □

Remark 3.7. One immediate consequence of Theorem 3.3 is that, subject to conditions (C1)–(C4), we have

$$\max_{0\le\tau\le T} |\theta^\varepsilon_{\lfloor\tau/\varepsilon\rfloor} - \theta^0(\tau)| = O(\varepsilon^{\frac{1}{2}}\sqrt{2\log\log\varepsilon^{-1}}) \quad \text{a.s.} \tag{3.16}$$

for each $T \in (0,\infty)$. We thus complement the finite horizon weak law of large numbers (1.3) with a strong law of large numbers together with an a.s. rate of convergence, which, by an argument identical to that in ([12], page 120), may be seen to be the best possible rate of convergence.

Remark 3.8. The general methodology used for establishing Theorem 3.3 is suggested by an approach developed in [15] for proving a functional LIL for random ODEs. The methods of [15] depend in an essential way on rather restrictive boundedness hypotheses which, when carried over directly into the context of (1.1), entail uniform boundedness of the driving data sequence $\{X_n\}$ with respect to $n = 1, 2, \ldots$ and $\omega \in \Omega$. Although this boundedness may be reasonably acceptable for differential equations, it is not realistic for algorithms, and it is necessary to significantly redesign the overall approach used in [15] for differential equations to suit the system (1.1). In the following sections we shall extensively use (i) nice properties of geometric L-mixing processes, in particular their approximability (see Remark 2.5) by s-dependent processes, (ii) L^p-bounds for sums of L-mixing processes over "triangular" domains (see Theorem 5.2), and (iii) the linear structure of (1.1), in order to deal with the problems caused by unboundedness of the data sequences in (1.1).

4. Proof of Theorem 3.5.

Notation 4.1. For this section we need the following additional notation: If (S, ρ) is a metric space then $\mathcal{B}(S)$ denotes its Borel σ-algebra, and if Y is a $\mathcal{F}/\mathcal{B}(S)$-measurable mapping from a triple $(\Omega, \mathcal{F}, P)$ into (S, ρ), then $\mathcal{L}(Y)$ is the probability measure on $\mathcal{B}(S)$ defined by $\mathcal{L}(Y)(A) \overset{\triangle}{=} P\{\omega : \ Y(\omega) \in A\}$ for all $A \in \mathcal{B}(S)$. For probability measures P_1 and P_2 on the metric space $C[0,T]$ with metric given by the norm $\|\cdot\|_C$ (see Notation 3.1), let $\Pi_C(P_1, P_2)$ denote the Prohorov distance between P_1 and P_2 (see section 8 for a general definition of Prohorov distance). If $\{\alpha_n\}$ and $\{\beta_n\}$ are sequences of real numbers, then the notation $\alpha_n \ll \beta_n$ indicates the existence of a constant $C \in [0,\infty)$ such that $|\alpha_n| \le C|\beta_n|$ for all $n = 1, 2, \ldots$ If α_ε and β_ε are real numbers for each $\varepsilon \in (0,1]$ then $\alpha_\varepsilon \ll \beta_\varepsilon$ indicates the existence of numbers $\varepsilon_0 \in (0,1]$ and $C \in [0,\infty)$ such that $|\alpha_\varepsilon| \le C|\beta_\varepsilon|$ for all $\varepsilon \in (0, \varepsilon_0]$.

The proof of Theorem 3.5 relies on the following result which is an immediate consequence of combining Lemma 2.1(iv) and Theorem 4.3 of Kuelbs [16].

THEOREM 4.2. *Fix some $T \in (0,\infty)$, and suppose that $M : [0,T] \to \Re^{d\times d}$ is continuous with $M(s)$ being positive-definite symmetric for each $s \in [0,T]$. Then the set*

$$K \overset{\triangle}{=} \left\{ \phi \in AC_0[0,T] \colon \frac{1}{2}\int_0^T (\dot{\phi}(s))' M^{-1}(s)\dot{\phi}(s)\, ds \le 1 \right\} \tag{4.1}$$

is a compact subset of $C[0,T]$*. Define a Gaussian process* $\{\hat{Y}(\tau),\ \tau \in [0,T]\}$ *by*

$$\hat{Y}(\tau) \overset{\triangle}{=} \int_0^\tau M^{\frac{1}{2}}(s)\, d\hat{B}(s) \qquad \forall\, \tau \in [0,T], \tag{4.2}$$

where $\hat{B}(\cdot)$ *is a standard* $\Re^d$*-valued Brownian motion on a triple* $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$*, and let* $\{Y_r,\ r = 1,2,\ldots\}$ *be a sequence of* $C[0,T]$*-valued random variables defined on a triple* $(\Omega, \mathcal{F}, P)$*. Then the following hold.* (i) *If*

$$\sum_{r=1}^{\infty} \Pi_C(\mathcal{L}(Y_r), \mathcal{L}(\hat{Y})) < \infty, \tag{4.3}$$

then

$$\lim_{r\to\infty} \left\| \frac{Y_r}{\sqrt{2\log r}} - K \right\|_C = 0, \quad a.s. \tag{4.4}$$

(ii) *If, in addition to* (4.3)*, the sequence* $\{Y_r\}$ *is independent, then*

$$\operatorname{acc}\left\{ \frac{Y_r(\omega)}{\sqrt{2\log r}} \right\} = K, \qquad a.s. \tag{4.5}$$

Remark 4.3. In Theorem 4.2 the set K is defined in terms of the covariance function $M(\cdot)$ of a Gaussian process $\{\hat{Y}(\tau)\}$. The theorem says that if a sequence of $C[0,T]$-valued random functions $\{Y_r\}$ converges fast enough in distribution to the Gaussian limit $\hat{Y}$ (see (4.3)), then all accumulation points of the sequence $\{Y_r(\omega)/\sqrt{2\log r},\ r = 1,2,\ldots\}$ are included *within* the set K for P-almost all ω (see (4.4)), regardless of any dependence which may exist in the sequence $\{Y_r\}$. This is true, in particular, when the Y_r are Gaussian with common distribution equal to that of $\hat{Y}$, for then (4.3) is automatically satisfied. If, in addition, the sequence $\{Y_r\}$ is *independent*, then the set of accumulation points of $\{Y_r(\omega)/\sqrt{2\log r},\ r = 1,2,\ldots\}$ is not only included within K but actually coincides *exactly* with K for P-almost all ω (see (4.5)). Taken together, (4.4) and (4.5) constitute a law of the *noniterated* logarithm. In the present section we shall use Theorem 4.4 (to follow) as a tool for verifying (4.3), and then establish Theorem 3.5 on the basis of the law of the noniterated logarithm resulting from Theorem 4.2.

THEOREM 4.4 (proved in section 6). *Suppose conditions* (C1)–(C4) *of section* 2 *hold and fix some finite horizon* $T \in (0,\infty)$*. Let* $\{\hat{W}^0(\tau),\ \tau \in [0,T]\}$ *be defined by* (3.8)*, and let* $\{W^\varepsilon(\tau),\ \tau \in [0,T]\}$ *be defined by* (3.5)*. Then there is a constant* $\eta \in (0,\infty)$ *such that*

$$\Pi_C(\mathcal{L}(W^\varepsilon), \mathcal{L}(\hat{W}^0)) \ll \varepsilon^\eta.$$

Remark 4.5. Fix an arbitrary finite horizon $T \in (0,\infty)$. For the proofs in this and in later sections we shall need the $\Re^d$-valued process $\{\xi_n^\varepsilon,\ n = 1,2,\ldots\}$ defined for each $\varepsilon \in (0,1]$ by

$$\xi_n^\varepsilon \overset{\triangle}{=} \tilde{H}_n(\theta^0(((n-1)\varepsilon) \wedge T)) \equiv \tilde{b}_n - \tilde{R}_n\theta^0(((n-1)\varepsilon) \wedge T) \quad \forall\, n = 1,2,\ldots \tag{4.6}$$

(recall (2.7) and (2.5)). To simplify the notation, we do not indicate dependence of ξ_n^ε upon the horizon T, which is fixed throughout this work. It follows at once from

(3.5) that we have

$$W^\varepsilon(\tau) = \varepsilon^{\frac{1}{2}} \left(\sum_{\nu=1}^{\lfloor \tau/\varepsilon \rfloor} \xi_\nu^\varepsilon + (\tau/\varepsilon - \lfloor \tau/\varepsilon \rfloor) \xi_{\lfloor \tau/\varepsilon \rfloor + 1}^\varepsilon \right) \qquad \forall\, \tau \in [0, T]. \tag{4.7}$$

Since $\{\tilde{b}_n\}$ and $\{\tilde{R}_n\}$ are zero-mean and geometrically L-mixing (see Remark 2.7), and $\theta^0(((n-1)\varepsilon) \wedge T)$ is uniformly bounded with respect to $\varepsilon \in (0,1]$, $n = 1, 2, \ldots$, and nonrandom, it follows at once that $\{\xi_n^\varepsilon,\ n = 1, 2, \ldots\}$ is zero-mean and geometrically L-mixing. In fact, if

$$\xi_n^\varepsilon[s] \overset{\triangle}{=} E\left[\xi_n^\varepsilon \mid \mathcal{F}_{n-s}^n\right] = \tilde{b}_n[s] - \tilde{R}_n[s]\theta^0((n-1)\varepsilon \wedge T) \qquad \forall\, s, n = 1, 2, \ldots, \tag{4.8}$$

then there is a constant $\lambda \in (0,1)$, and, for each $p \in [1, \infty)$, a constant $C_p \in [0, \infty)$, such that

$$\sup_{n > s} \|\xi_n^\varepsilon - \xi_n^\varepsilon[s]\|_p \le C_p\, \lambda^s \qquad \forall\, s = 1, 2, \ldots, \qquad \forall\, \varepsilon \in (0, 1], \tag{4.9}$$

and it follows from Remark 2.5 that $\{\xi_n^\varepsilon[s],\ n = 1, 2, 3, \ldots\}$ is a zero-mean s-dependent process for each $s = 1, 2, \ldots$ and $\varepsilon \in (0, 1]$. Notice that λ and C_p in (4.9) are *uniform* with respect to $\varepsilon \in (0, 1]$.

Remark 4.6. We can apply Theorem 8.5 separately to the geometrically L-mixing processes $\{\tilde{b}_n\}$ and $\{\tilde{R}_n\}$ in (4.6), and use the uniform boundedness in (ε, n) of $\theta^0(((n-1)\varepsilon) \wedge T)$ to get the following. For each $p \in [2, \infty)$ there is a constant $C_p^1 \in [0, \infty)$ such that for any nonrandom sequence $\{A_n\}$ of $d \times d$-matrices we have

$$\left\| \sum_{n=1}^N A_n \xi_n^\varepsilon \right\|_p \le C_p^1 \left(\sum_{n=1}^N |A_n|^2 \right)^{\frac{1}{2}} \qquad \forall\, N = 1, 2, \ldots, \qquad \forall\, \varepsilon \in (0, 1].$$

Notice that the dependencies of the constants in Theorem 8.5, and uniformity in $\varepsilon \in (0, 1]$ of C_p and λ in (4.9), entail that C_p^1 does not depend on ε, N, or the sequence $\{A_n\}$.

Proof of Theorem 3.5. From Lemma 2.9 we know that $A(\theta^0(s))$ is continuous in $s \in [0, T]$. Upon comparing (3.11) and (4.1), we see from Theorem 4.2 that K_W^T is compact. We now use Theorem 4.2 to show that a.s.

$$\text{(i)}\ \ \lim_{\varepsilon \searrow 0} \left\| \frac{W^\varepsilon}{\sqrt{2 \log\log \varepsilon^{-1}}} - K_W^T \right\|_C = 0, \quad \text{(ii)}\ \ K_W^T = \operatorname{acc}\left\{ \frac{W^\varepsilon}{\sqrt{2 \log\log \varepsilon^{-1}}} \right\}, \tag{4.10}$$

which gives Theorem 3.5.

Proof of (4.10)(i). Without loss of generality take $T = 1$. Fix $\sigma \in [\frac{9}{10}, 1)$ and put

$$\varepsilon_r \overset{\triangle}{=} \exp(-r^\sigma) \qquad \forall\, r = 1, 2, 3, \ldots. \tag{4.11}$$

Then, for each $\varepsilon \in [\varepsilon_{r+1}, \varepsilon_r]$, we have

$$\left\| \frac{W^{\varepsilon}(\cdot)}{\sqrt{2\log\log\varepsilon^{-1}}} - \frac{K_W^T}{\sqrt{\sigma}} \right\|_C \leq \left\| \frac{W^{\varepsilon}(\cdot)}{\sqrt{2\log\log\varepsilon^{-1}}} - \frac{W^{\varepsilon_r}(\cdot)}{\sqrt{2\log\log\varepsilon^{-1}}} \right\|_C$$
$$+ \left\| \frac{W^{\varepsilon_r}(\cdot)}{\sqrt{2\log\log\varepsilon^{-1}}} - \frac{W^{\varepsilon_r}(\cdot)}{\sqrt{2\log\log\varepsilon_r^{-1}}} \right\|_C + \left\| \frac{W^{\varepsilon_r}(\cdot)}{\sqrt{2\log\log\varepsilon_r^{-1}}} - \frac{K_W^T}{\sqrt{\sigma}} \right\|_C ,$$

and thus

$$\text{(4.12)} \qquad \sup_{\varepsilon_{r+1}\leq\varepsilon\leq\varepsilon_r} \left\| \frac{W^{\varepsilon}}{\sqrt{2\log\log\varepsilon^{-1}}} - \frac{K_W^T}{\sqrt{\sigma}} \right\|_C \leq \sup_{\varepsilon_{r+1}\leq\varepsilon\leq\varepsilon_r} \frac{\|W^{\varepsilon} - W^{\varepsilon_r}\|_C}{\sqrt{2\log\log\varepsilon_r^{-1}}}$$
$$+ \left\| \frac{W^{\varepsilon_r}}{\sqrt{2\log\log\varepsilon_r^{-1}}} \right\|_C \left(1 - \sqrt{\frac{\log\log\varepsilon_r^{-1}}{\log\log\varepsilon_{r+1}^{-1}}} \right) + \left\| \frac{W^{\varepsilon_r}}{\sqrt{2\log\log\varepsilon_r^{-1}}} - \frac{K_W^T}{\sqrt{\sigma}} \right\|_C .$$

Remark 4.7. We will show that the three terms on the right of (4.12) go to zero a.s. when $r \to \infty$ for each $\sigma \in [\frac{9}{10}, 1)$. Since we can choose σ arbitrarily close to 1, this gives (4.10)(i). Our choice of the sequence $\{\varepsilon_r\}$ is determined by the following considerations. From (4.7) we see that the number of terms in the sum for $W^{\varepsilon}(\tau)$ increases reciprocally with decreasing ε. Thus, to control the supremum appearing in the first term on the right side of (4.12), we want the difference $(\varepsilon_{r+1}^{-1} - \varepsilon_r^{-1})$ to not be too large, which means that $\{\varepsilon_r\}$ must go to zero quite slowly. On the other hand, in the course of the following proof, we shall need Theorem 4.2(i) to deal with the last term on the right of (4.12), using Theorem 4.4 to verify a bound of the form (4.3) (with $Y_r \triangleq W^{\varepsilon_r}$, $\hat{Y} \triangleq \hat{W}^0$), and for this it is important that $\{\varepsilon_r\}$ not go to zero *too slowly.* Our choice of the sequence $\{\varepsilon_r\}$ turns out to be the right compromise, meeting both of these requirements.

We now deal with the first term on the right of (4.12). For each $\gamma \in [0, \infty)$, define piecewise-constant process $\{S_\gamma(\tau),\ \tau \in [0,1]\}$ by

$$\text{(4.13)} \qquad S_\gamma(\tau) \triangleq \begin{cases} \sum_{j=1}^{\lfloor \tau\gamma \rfloor} \left[\tilde{b}_j - \tilde{R}_j \theta^0((j-1)/\gamma) \right], & \text{when } \ \gamma > 0, \\ 0, & \text{when } \ \gamma = 0, \end{cases}$$

and observe, from (4.7) and (4.13), that

$$\text{(4.14)} \qquad W^{\varepsilon}(\tau) = \varepsilon^{\frac{1}{2}} \{ S_{\varepsilon^{-1}}(\tau) + (\tau/\varepsilon - \lfloor \tau/\varepsilon \rfloor)\, \xi^{\varepsilon}_{\lfloor \tau/\varepsilon \rfloor + 1} \} \qquad \forall\, \tau \in [0,1].$$

Put $N_\varepsilon \triangleq \lfloor \varepsilon^{-1} \rfloor$ for all $\epsilon \in (0,1]$, and observe from (4.14) that

$$\text{(4.15)} \ \|W^{\varepsilon} - W^{\varepsilon_r}\|_C \leq \varepsilon^{\frac{1}{2}} \max_{k=1,\ldots,N_\varepsilon} |\xi_k^{\varepsilon}| + \|\varepsilon^{\frac{1}{2}} S_{\varepsilon^{-1}} - \varepsilon_r^{\frac{1}{2}} S_{\varepsilon_r^{-1}}\|_C + \varepsilon_r^{\frac{1}{2}} \max_{k=1,\ldots,N_{\varepsilon_r}} |\xi_k^{\varepsilon_r}|$$

for all $\epsilon \in [\varepsilon_{r+1}, \varepsilon_r]$. Now we need the following result.

LEMMA 4.8 (proved in section 7). *Suppose that $\{\gamma_n\}$ is a sequence of random variables (either $\Re^d$- or $\Re^{d\times r}$-valued for all n) on a probability space $(\Omega, \mathcal{F}, P)$, such that* $\sup_n \|\gamma_n\|_8 < \infty$. *Then, for almost always (a.a.) ω, there exists some constant $C(\omega) \in [0, \infty)$ such that*

$$\max_{n=1,2,\ldots,N} |\gamma_i(\omega)| \leq C(\omega) N^{\frac{1}{7}} \qquad \forall\, N = 1, 2, \ldots .$$

By (4.6) and Lemma 4.8, for P-a.a. ω there are constants $c_1(\omega), c_2(\omega) \in [0,\infty)$ such that

$$(4.16)\ \varepsilon^{\frac{1}{2}} \max_{k=1,\dots,N_\varepsilon} |\xi_k^\varepsilon(\omega)| \le \varepsilon^{\frac{1}{2}} \left\{ \max_{k=1,2,\dots,N_\varepsilon} |\tilde{b}_k(\omega)| + \max_{k=1,\dots,N_\varepsilon} |\tilde{R}_k(\omega)||\theta^0((k-1)\varepsilon)| \right\}$$

$$\le \varepsilon^{\frac{1}{2}} c_1(\omega) \left(N_\varepsilon^{\frac{1}{7}} + N_\varepsilon^{\frac{1}{7}} \max_{0\le\tau\le 1} |\theta^0(\tau)| \right) \le c_2(\omega)\varepsilon_r^{\frac{5}{14}} \quad \forall\, \varepsilon \in [\varepsilon_{r+1}, \varepsilon_r], \quad \forall\, r = 1,2,3,\dots.$$

For the second term on the right of (4.15), fix some integer $p \ge 3$ and observe that

$$(4.17) \qquad E\left[\sup_{\varepsilon_{r+1}\le\varepsilon\le\varepsilon_r} \|\varepsilon^{\frac{1}{2}} S_{\varepsilon^{-1}} - \varepsilon_r^{\frac{1}{2}} S_{\varepsilon_r^{-1}}\|_C^{2p} \right]$$

$$\le 2^{2p} \left\{ E\left[\sup_{\varepsilon_{r+1}\le\varepsilon\le\varepsilon_r} (|\varepsilon_r^{\frac{1}{2}} - \varepsilon^{\frac{1}{2}}| \|S_{\varepsilon^{-1}}\|_C)^{2p} \right] + E\left[\sup_{\varepsilon_{r+1}\le\varepsilon\le\varepsilon_r} (\varepsilon_r^{\frac{1}{2}} \|S_{\varepsilon^{-1}} - S_{\varepsilon_r^{-1}}\|_C)^{2p} \right] \right\}.$$

To deal with the expectations on the right side of (4.17) we need the next result, which is a slight variant of Lemma 3.7 in [15], and is proved in exactly the same way.

LEMMA 4.9. *Suppose conditions* (C1)–(C2) *of section* 2 *hold, and define* $S_\gamma(\tau)$, $\tau \in [0,1]$, *as in* (4.13) *for each* $\gamma \in [0,\infty)$. *Then, corresponding to each integer* $p \ge 3$, *there exists a constant* $\alpha_p \in (0,\infty)$ *such that*

$$E\left[\sup_{\gamma\le u\le\eta} \|S_u - S_\gamma\|_C^{2p} \right] \le \alpha_p \eta(\eta-\gamma)^{p-1} \qquad \forall\, 0 \le \gamma < \eta < \infty.$$

Since $\sigma < 1$ in (4.11), we see from the mean-value theorem applied to $r \to r^\sigma$ that $\varepsilon_r/\varepsilon_{r+1} \le e^\sigma$. Thus, from Lemma 4.9,

$$(4.18) \quad E\left[\sup_{\varepsilon_{r+1}\le\varepsilon\le\varepsilon_r} (\varepsilon_r^{\frac{1}{2}} \|S_{\varepsilon^{-1}} - S_{\varepsilon_r^{-1}}\|_C)^{2p} \right] = \varepsilon_r^p E\left[\sup_{\varepsilon_r^{-1}\le u\le\varepsilon_{r+1}^{-1}} \|S_u - S_{\varepsilon_r^{-1}}\|_C^{2p} \right]$$

$$\le \alpha_p \varepsilon_r^p \varepsilon_{r+1}^{-1} \left(\varepsilon_{r+1}^{-1} - \varepsilon_r^{-1}\right)^{p-1} \le e^\sigma \alpha_p \{(\varepsilon_r - \varepsilon_{r+1})(\varepsilon_{r+1}^{-1})\}^{p-1} \triangleq e^\sigma \alpha_p B_r^p \qquad \forall\, r = 1,2,\dots.$$

In the same way, one shows that the first expectation on the right side of (4.17) has an identical upper bound, and thus the quantity on the left side of (4.17) is $O(B_r^p)$, where the constant implied by O depends only on p. By applying the mean value theorem to $r \to \exp(-r^\sigma)$, we easily see that $B_r^p \le \{\sigma e^\sigma r^{(\sigma-1)}\}^{(p-1)}$. Then, fixing integer $p > \frac{2-\sigma}{1-\sigma}$, it follows that $r^{(\sigma-1)(p-1)} = O(r^{-\beta-1})$ for some $\beta > 0$, and thus the sequence $\{B_r^p,\ r = 1,2,\dots\}$ is summable. Hence, from (4.17) and the monotone convergence theorem,

$$E\left[\sum_{r\ge 1} \sup_{\varepsilon_{r+1}\le\varepsilon\le\varepsilon_r} \|\varepsilon^{\frac{1}{2}} S_{\varepsilon^{-1}} - \varepsilon_r^{\frac{1}{2}} S_{\varepsilon_r^{-1}}\|_C^{2p} \right] < \infty,$$

which implies

$$(4.19) \qquad \lim_{r\to\infty} \left[\sup_{\varepsilon_{r+1}\le\varepsilon\le\varepsilon_r} \|\varepsilon^{\frac{1}{2}} S_{\varepsilon^{-1}} - \varepsilon_r^{\frac{1}{2}} S_{\varepsilon_r^{-1}}\|_C \right] = 0 \text{ a.s.}$$

Similarly to (4.16), one sees that the third term on the right side of (4.15) is $O(\varepsilon_r^{\frac{5}{14}})$ a.s., and hence using (4.15), (4.16), and (4.19), we get

$$\lim_{r\to\infty}\left\{\sup_{\varepsilon_{r+1}\le\varepsilon\le\varepsilon_r}\|W^\varepsilon - W^{\varepsilon_r}\|_C\right\} = 0 \text{ a.s.} \tag{4.20}$$

This controls the first term on the right of (4.12). As for the third term, use Theorem 4.4 to find some constants $c_1 > 0$, $\eta > 0$ and $r_0 > 0$ such that $\Pi_C(\mathcal{L}(W^{\varepsilon_r}), \mathcal{L}(\hat{W}^0)) \le c_1 \exp(-\eta r^\sigma)$ for all $r \ge r_0$, thus $\sum_r \Pi_C(\mathcal{L}(W^{\varepsilon_r}), \mathcal{L}(\hat{W}^0)) < \infty$. Hence Theorem 4.2(i) (with $M(\tau) \triangleq A(\theta^0(\tau))$, $Y_r(\cdot) \triangleq W^{\varepsilon_r}(\cdot)$, $\hat{Y}(\cdot) \triangleq \hat{W}^0(\cdot)$, and $K \triangleq K_W^T$) establishes the convergence $\lim_{r\to\infty} \|\frac{W^{\varepsilon_r}(\cdot)}{\sqrt{2\log r}} - K_W^T\|_C = 0$ *a.s.* Now $\sigma \log r = \log\log \varepsilon_r^{-1}$ (see (4.11)); therefore,

$$\lim_{r\to\infty}\left\|\frac{W^{\varepsilon_r}}{\sqrt{2\log\log\varepsilon_r^{-1}}} - \frac{K_W^T}{\sqrt{\sigma}}\right\|_C = 0 \quad \text{a.s.} \tag{4.21}$$

As for the second term on the right side of (4.12), clearly $\lim_r \frac{\log\log\varepsilon_r^{-1}}{\log\log\varepsilon_{r+1}^{-1}} = 1$ (by L'Hôpital formula), and thus (4.21) and boundedness of K_W^T in $C[0,1]$ shows that this term goes a.s. to zero with $r \to \infty$. It follows that the quantity on the left side of (4.12) tends a.s. to zero as $r \to \infty$.

Proof of (4.10)(ii). Fix $\sigma \in (1, 3/2)$, and define ε_r as in (4.11).

Remark 4.10. In contrast to the situation of Remark 4.7, where $\sigma < 1$ so that ε_r^{-1} increases quite slowly with r, here we take $\sigma > 1$ in (4.11) so that ε_r^{-1} increases *extremely rapidly* with r. In view of (4.7), this ensures that the sum for $W^{\varepsilon_{r+1}}(\tau)$ involves many more terms than does the sum for $W^{\varepsilon_r}(\tau)$, or, equivalently, the sums for $W^{\varepsilon_{r+1}}(\tau)$ and $W^{\varepsilon_r}(\tau)$ have few terms in common. The geometric L-mixing of $\{\xi_\nu^\varepsilon\}$ then suggests that the sequence $\{W^{\varepsilon_r}(\cdot)\}$ of $C[0,1]$-valued random variables should be approximately independent, and we can then expect to use Theorem 4.2(ii) to obtain (4.10)(ii). Indeed, in the course of the following proof we shall make this intuition rigorous, using the geometric L-mixing of $\{\xi_\nu^\varepsilon\}$ and rapid increase of $\{\varepsilon_r\}$ to construct a sequence $\{W_2^r(\cdot)\}$ of independent $C[0,1]$-valued random variables to which we can apply Theorem 4.2(ii), and which approximates the sequence $\{W^{\varepsilon_r}(\cdot)\}$ in a strong sense (see (4.31)). From this it is easy to deduce (4.10)(ii).

Recalling $\lambda \in (0,1)$ in (4.9), define

$$q(r) \triangleq \left\lfloor \frac{r^2}{2\ln\lambda^{-1}} \right\rfloor, \quad \tau_r \triangleq \varepsilon_r\left(1 + q(r) + \frac{1}{\varepsilon_{r-1}}\right) \ \forall\, r = 1, 2, \ldots, \tag{4.22}$$

$$\zeta_u^\varepsilon \triangleq \xi_{\lfloor u\rfloor+1}^\varepsilon, \quad \zeta_u^\varepsilon[q] \triangleq \xi_{\lfloor u\rfloor+1}^\varepsilon[q] \quad \forall \varepsilon \in (0,1], \quad \forall u \in [0,\infty), \quad \forall q = 1, 2 \ldots \lfloor u\rfloor\,. \tag{4.23}$$

For each $r = 1, 2, \ldots$, define $\{W_1^r(\tau),\ \tau \in [0,1]\}$ and $\{W_2^r(\tau),\ \tau \in [0,1]\}$ by

$$W_1^r(\tau) \triangleq \begin{cases} 0, & \text{if } 0 \le \tau \le \tau_r, \\ \varepsilon_r^{\frac{1}{2}} \displaystyle\int_{\tau_r/\varepsilon_r}^{\tau/\varepsilon_r} \zeta_u^{\varepsilon_r}\, du, & \text{if } \tau_r < \tau \le 1, \end{cases} \tag{4.24}$$

$$W_2^r(\tau) \triangleq \begin{cases} 0, & \text{if } 0 \le \tau \le \tau_r, \\ \varepsilon_r^{\frac{1}{2}} \displaystyle\int_{\tau_r/\varepsilon_r}^{\tau/\varepsilon_r} \zeta_u^{\varepsilon_r}[q(r)]\, du, & \text{if } \tau_r < \tau \le 1. \end{cases} \tag{4.25}$$

We first show that $\{W_2^r(\cdot)\}$ is a sequence of $C[0,1]$-valued *independent* random variables. To this end, observe from (4.22), (4.23), and (4.25) that

$$\mathcal{A}_r \overset{\triangle}{=} \sigma\{W_2^r(\tau),\ 0 \leq \tau \leq 1\} \subset \sigma\{\xi^{\varepsilon_r}_{2+q(r)+\lfloor \varepsilon_{r-1}^{-1} \rfloor}[q(r)], \ldots, \xi^{\varepsilon_r}_{1+\lfloor \varepsilon_r^{-1} \rfloor}[q(r)]\}. \tag{4.26}$$

Now $\xi_j^{\varepsilon_r}[q]$ is $\mathcal{F}_{j-q}^j$-measurable for all $q, j = 1, 2, \ldots$ (see (4.8)), and thus from (4.26) we find

$$\mathcal{A}_r \subset \mathcal{F}_{2+\lfloor \varepsilon_{r-1}^{-1} \rfloor}^{1+\lfloor \varepsilon_r^{-1} \rfloor}, \tag{4.27}$$

where $\mathcal{F}_m^n$ is given by (2.1). Now one sees from Definition 2.3 that the finite collection of σ-algebras $\{\mathcal{F}_{m_1}^{n_1}, \mathcal{F}_{m_2}^{n_2}, \ldots, \mathcal{F}_{m_r}^{n_r}\}$ is independent when $m_1 < n_1 < m_2 < n_2 < \cdots < m_r < n_r$; hence (4.27) shows that $\{W_2^r(\cdot),\ r = 1, 2, \ldots\}$ is a sequence of $C[0,1]$-valued independent random variables. In order to use Theorem 4.2(ii) on this sequence, we next show that

$$\sum_{r=1}^{\infty} \Pi_C(\mathcal{L}(W_2^r), \mathcal{L}(\hat{W}^0)) < \infty, \tag{4.28}$$

where $\{\hat{W}^0(\tau),\ \tau \in [0,1\}$ is defined by (3.8). By the triangle inequality we have

$$\begin{aligned} \Pi_C(\mathcal{L}(W_2^r), \mathcal{L}(\hat{W}^0)) \leq \Pi_C(\mathcal{L}(W_2^r), \mathcal{L}(W_1^r)) &+ \Pi_C(\mathcal{L}(W_1^r), \mathcal{L}(W^{\varepsilon_r})) \\ &+ \Pi_C(\mathcal{L}(W^{\varepsilon_r}), \mathcal{L}(\hat{W}^0)), \end{aligned} \tag{4.29}$$

and it remains to upper-bound the three terms on the right side of (4.29). To this end we require the following lemma.

Lemma 4.11 (proved in section 7). *Suppose the hypotheses of Theorem* 3.5. *Then we have*

$$E\left[\max_{0 \leq \tau \leq 1} |W_1^r(\tau) - W_2^r(\tau)|^4\right] \ll \varepsilon_r.$$

Then, from Lemma 8.6(ii) (with $c \overset{\triangle}{=} 4$) we get $\Pi_C(\mathcal{L}(W_1^r), \mathcal{L}(W_2^r)) \ll (\varepsilon_r^{\frac{1}{4}})^{\frac{4}{5}} = \varepsilon_r^{\frac{1}{5}}$. Next, we need the following lemma.

Lemma 4.12 (proved in section 7). *Suppose the hypotheses of Theorem* 3.5. *Then we have*

$$E\left[\max_{0 \leq \tau \leq 1} |W^{\varepsilon_r}(\tau) - W_1^r(\tau)|^4\right] \ll r^{-10}.$$

By Lemmas 4.12 and 8.6(ii) (with $c \overset{\triangle}{=} 4$) we get $\Pi_C(\mathcal{L}(W_1^r), \mathcal{L}(W^{\varepsilon_r})) \ll (r^{-\frac{5}{2}})^{\frac{4}{5}} = r^{-2}$. By Theorem 4.4 there exists $r_0, \eta \in (0, \infty)$ such that $\Pi_C(\mathcal{L}(W^{\varepsilon_r}), \mathcal{L}(\hat{W}^0)) \ll \varepsilon_r^{\eta}$ for all $r \geq r_0$, and hence by (4.29), we get $\Pi_C(\mathcal{L}(W_2^r), \mathcal{L}(\hat{W}^0)) \ll \varepsilon_r^{\frac{1}{5}} + \varepsilon_r^{\eta} + r^{-2}$, and (4.28) follows. Thus, by independence of the sequence $\{W_2^r(\cdot),\ r = 1, 2, \ldots,\}$ and Theorem 4.2(ii) (with $M(\tau) \overset{\triangle}{=} A(\theta^0(\tau))$, $Y_r(\cdot) \overset{\triangle}{=} W_2^r(\cdot)$, $\hat{Y}(\cdot) \overset{\triangle}{=} \hat{W}^0(\cdot)$, and $K \overset{\triangle}{=} K_W^T$), we get $\mathrm{acc}\{\frac{W_2^r}{\sqrt{2 \log r}}\} = K_W^T$ a.s. Now $\sigma \log r = \log \log \varepsilon_r^{-1}$ (see (4.11)), and hence

$$\mathrm{acc}\left\{\frac{W_2^r}{\sqrt{2 \log \log \varepsilon_r^{-1}}}\right\} = \frac{K_W^T}{\sqrt{\sigma}}. \tag{4.30}$$

From Lemma 4.11 and Borel–Cantelli, we have $\lim_{r\to\infty}\|W_1^r - W_2^r\|_C = 0$ a.s. Similarly, by Lemma 4.12, $\lim_{r\to\infty}\|W_1^r - W^{\varepsilon_r}\|_C = 0$ a.s. Thus, by the triangle inequality,

$$\lim_{r\to\infty}\|W_2^r - W^{\varepsilon_r}\|_C = 0 \text{ a.s.} \tag{4.31}$$

From (4.30), (4.31), and the fact that $\{W^{\varepsilon_r}\}$ is a subnet of $\{W^\varepsilon\}$,

$$\frac{K_W^T}{\sqrt{\sigma}} = \operatorname{acc}\left\{\frac{W^{\varepsilon_r}}{\sqrt{2\log\log\varepsilon_r^{-1}}}\right\} \subset \operatorname{acc}\left\{\frac{W^{\varepsilon}}{\sqrt{2\log\log\varepsilon^{-1}}}\right\} \quad \text{a.s.} \tag{4.32}$$

Since (4.32) holds for σ arbitrarily near 1 with $\sigma > 1$, we have

$$K_W^T \subset \operatorname{acc}\left\{\frac{W^{\varepsilon}}{\sqrt{2\log\log\varepsilon^{-1}}}\right\} \quad \text{a.s.} \tag{4.33}$$

Now (4.10)(i) and (4.33) yield (4.10)(ii) as required. □

5. Proof of Theorem 3.6.

Remark 5.1. For ease of notation put

$$Z^\varepsilon(\tau) \overset{\triangle}{=} G^\varepsilon(W^\varepsilon)(\tau) \quad \forall\, \tau \in [0,T]. \tag{5.1}$$

From (3.6) and (4.7) we see that, for each $\varepsilon \in (0,1]$, the sequence $\{Z^\varepsilon(\varepsilon k)\}$ is given by

$$Z^\varepsilon(\varepsilon(k+1)) = (I - \varepsilon\bar{R})Z^\varepsilon(\varepsilon k) + \varepsilon^{\frac{1}{2}}\xi_{k+1}^\varepsilon, \quad Z^\varepsilon(0) = 0, \tag{5.2}$$

which is a linear system driven by the geometrically L-mixing process $\{\xi_k^\varepsilon\}$. Theorem 3.6 effectively relates $\{\Theta^\varepsilon(\varepsilon k)\}$ to the output of this system.

Without loss of generality we shall take $T = 1$. Put $N_\varepsilon \overset{\triangle}{=} \lfloor\varepsilon^{-1}\rfloor$ for all $\varepsilon \in (0,1]$. From (1.1), (1.2), and (2.5),

$$\theta_k^\varepsilon = \theta_* + \varepsilon\sum_{j=0}^{k-1}\left(b_{j+1} - R_{j+1}\theta_j^\varepsilon\right), \qquad \theta^0(\varepsilon k) = \theta_* + \int_0^{\varepsilon k}\left(\bar{b} - \bar{R}\theta^0(s)\right) ds. \tag{5.3}$$

In view of (1.6) we have $\Theta^\varepsilon(\varepsilon k) \overset{\triangle}{=} \varepsilon^{-\frac{1}{2}}(\theta_k^\varepsilon - \theta^0(\varepsilon k))$, and thus, using (5.3), (3.5), and (2.5),

$$\begin{aligned}
\Theta^\varepsilon(\varepsilon k) &= \varepsilon^{\frac{1}{2}}\sum_{j=0}^{k-1} b_{j+1} - \varepsilon^{\frac{1}{2}}\sum_{j=0}^{k-1} R_{j+1}\theta_j^\varepsilon + \varepsilon^{\frac{1}{2}}\sum_{j=0}^{k-1}(\tilde{b}_{j+1} - \tilde{R}_{j+1}\theta^0(\varepsilon j)) \\
&\quad - \varepsilon^{\frac{1}{2}}\sum_{j=0}^{k-1}((b_{j+1} - \bar{b}_{j+1}) - (R_{j+1} - \bar{R}_{j+1})\theta^0(\varepsilon j)) - k\varepsilon^{\frac{1}{2}}\bar{b} + \varepsilon^{-\frac{1}{2}}\int_0^{\varepsilon k}\bar{R}\theta^0(s)\, ds \\
&= W^\varepsilon(\varepsilon k) - \varepsilon^{\frac{1}{2}}\sum_{j=0}^{k-1} R_{j+1}(\theta_j^\varepsilon - \theta^0(\varepsilon j)) + \varepsilon^{\frac{1}{2}}\sum_{j=0}^{k-1}\left((\bar{b}_{j+1} - \bar{b})\right. \\
&\quad \left. - (\bar{R}_{j+1} - \bar{R})\theta^0(\varepsilon j)\right) + \varepsilon^{-\frac{1}{2}}\left(\int_0^{\varepsilon k}\bar{R}\theta^0(s)\, ds - \varepsilon\sum_{j=0}^{k-1}\bar{R}\theta^0(\varepsilon j)\right),
\end{aligned}$$

and hence, again using (2.5),

$$\Theta^\varepsilon(\varepsilon k) = W^\varepsilon(\varepsilon k) - \varepsilon \sum_{j=0}^{k-1} R_{j+1}\Theta^\varepsilon(\varepsilon j) + \varepsilon^{\frac{1}{2}} \sum_{j=0}^{k-1}(\hat{b}_{j+1} - \hat{R}_{j+1}\theta^0(\varepsilon j)) + \varepsilon^{-\frac{1}{2}}\left(\int_0^{\varepsilon k} \bar{R}\theta^0(s)\,ds - \varepsilon\sum_{j=0}^{k-1}\bar{R}\theta^0(\varepsilon j)\right) \tag{5.4}$$

for all $\varepsilon \in (0,1]$ and $k = 1,2,\ldots,N_\varepsilon$. Define

$$\Delta^\varepsilon(\varepsilon k) \overset{\triangle}{=} \Theta^\varepsilon(\varepsilon k) - Z^\varepsilon(\varepsilon k) \qquad \forall\, k = 0,1,2,\ldots,N_\varepsilon. \tag{5.5}$$

From (2.5) we know that $\bar{R} = R_{j+1} - (\tilde{R}_{j+1} + \hat{R}_{j+1})$, and thus, using (5.1) and (3.6),

$$Z^\varepsilon(\varepsilon k) = W^\varepsilon(\varepsilon k) - \varepsilon \sum_{j=0}^{k-1} R_{j+1} Z^\varepsilon(\varepsilon j) + \varepsilon \sum_{j=0}^{k-1}(\tilde{R}_{j+1} + \hat{R}_{j+1})Z^\varepsilon(\varepsilon j). \tag{5.6}$$

The quantity in brackets in the fourth term on the right side of (5.4) is clearly $O(\varepsilon)$ uniformly with respect to $k = 1,2,\ldots,N_\varepsilon$, while condition (C3a) ensures that the third term on the right side of (5.4) is $O(\varepsilon^{\frac{1}{2}})$ uniformly in $k = 1,2,\ldots,N_\varepsilon$. Subtracting (5.6) from (5.4) and taking magnitudes gives

$$|\Delta^\varepsilon(\varepsilon k)| \le \varepsilon \sum_{j=0}^{k-1} |R_{j+1}|\,|\Delta^\varepsilon(\varepsilon j)| + (|I_k^\varepsilon| + |J_k^\varepsilon| + O(\varepsilon^{\frac{1}{2}})), \tag{5.7}$$

where the constant implied by $O(\varepsilon^{\frac{1}{2}})$ is uniform with respect to $k = 1,2,\ldots,N_\varepsilon$, and

$$I_k^\varepsilon \overset{\triangle}{=} \varepsilon \sum_{j=0}^{k-1} \tilde{R}_{j+1} Z^\varepsilon(\varepsilon j), \quad J_k^\varepsilon \overset{\triangle}{=} \varepsilon \sum_{j=0}^{k-1} \hat{R}_{j+1} Z^\varepsilon(\varepsilon j) \quad \forall\, k = 1,2,\ldots,1+N_\varepsilon. \tag{5.8}$$

By (5.7), the fact that $\Delta^\varepsilon(0) = 0$, and the discrete-parameter version of Gronwall–Bellman's inequality (obtained by taking $\mu(\cdot)$ to be counting measure with weights $\varepsilon|R_{j+1}|$ over the nonnegative integers $j = 0,1,2,\ldots$ in Theorem 5.1 on page 498 of [8]), we get

$$\max_{0\le k\le N_\varepsilon+1} |\Delta^\varepsilon(\varepsilon k)| \le \exp\left(\varepsilon \sum_{j=1}^{N_\varepsilon} |R_j|\right) \max_{1\le k\le N_\varepsilon+1} (|I_k^\varepsilon| + |J_k^\varepsilon| + O(\varepsilon^{\frac{1}{2}})) \tag{5.9}$$

for all $\varepsilon \in (0,1]$. The proof proceeds by showing that, as $\varepsilon \searrow 0$, the right-hand side of (5.9) converges to zero a.s. Since $Z^\varepsilon(\cdot)$ and $\Theta^\varepsilon(\cdot)$ are both linear on intervals of the form $[(k-1)\varepsilon, k\varepsilon)$, $k = 1,2,\ldots,N_\varepsilon$ and continuous, this gives Theorem 3.6. Now products of geometrically L-mixing processes are geometrically L-mixing (see Remark 2.7), and thus one easily sees from condition (C1) and (2.5) that $\{|R_n|^2\}$ is geometrically L-mixing; in view of Theorem 8.5 and a standard use of the Borel–Cantelli lemma, it then follows that

$$\limsup_{\varepsilon\searrow 0} \varepsilon \sum_{j=1}^{N_\varepsilon} |R_j| < \infty, \quad \text{a.s.} \tag{5.10}$$

In light of (5.9) and (5.10), Theorem 3.6 will be established when we show that

$$\text{(5.11)}\quad \limsup_{\varepsilon\searrow 0}\max_{0\le k\le N_\varepsilon+1}|I^\varepsilon_k| = 0 \text{ a.s.} \quad\text{and}\quad \limsup_{\varepsilon\searrow 0}\max_{0\le k\le N_\varepsilon+1}|J^\varepsilon_k| = 0 \text{ a.s.}$$

To establish the first limit of (5.11), note from (5.2) that $Z^\varepsilon(\varepsilon j) = \varepsilon^{\frac{1}{2}}(I-\varepsilon\bar{R})^j\sum_{i=0}^{j-1}(I-\varepsilon\bar{R})^{-i-1}\xi^\varepsilon_{i+1}$ for all $j=0,1,\dots,N_\varepsilon{+}1$, from which we see that I^ε_k in (5.8) can be written as

$$\text{(5.12)}\ I^\varepsilon_k = \varepsilon^{\frac{3}{2}}\sum_{j=0}^{k-1}\tilde{R}_{j+1}(I-\varepsilon\bar{R})^j\Gamma^\varepsilon_j, \quad\text{where}\quad \Gamma^\varepsilon_j \triangleq \sum_{i=1}^{j}(I-\varepsilon\bar{R})^{-i}\xi^\varepsilon_i, \quad \Gamma^\varepsilon_0 \triangleq 0.$$

The following very special case of Theorem 1.1 of Gerencsér [10] is essential for establishing (5.11).

THEOREM 5.2. *Let $\{f_1(i)\}_{i=0}^\infty$ and $\{f_2(i)\}_{i=0}^\infty$ be real-valued nonrandom sequences, and let $\{u_1(i)\}_{i=0}^\infty$ and $\{u_2(i)\}_{i=0}^\infty$ be zero-mean geometrically L-mixing processes with rate $\lambda\in(0,1)$. Then there exists a constant $c\in(0,\infty)$ such that, for $l,k=1,2\dots$, with $l>k$, we have*

$$\left\|\sum_{j=k}^{l-1}\sum_{i=0}^{j-1} f_1(i)u_1(i)f_2(j)u_2(j)\right\|_4 \le c[\psi_1(l,k)+\psi_2(l,k)],$$

where (taking $f_1(i)\triangleq f_2(i)\triangleq 0$ when $i<0$)

$$\text{(5.13)}\qquad \begin{cases} \psi_1(l,k)\triangleq\left(\sum_{i=k}^{l} h(i,l,k)(|f_2(i-1)|+|f_2(i-2)|)\varphi_1^2(i)\right)^{\frac{1}{2}}, \\ \psi_2(l,k)\triangleq\sum_{i=1}^{l} h(i,l,k)|f_1(i-2)|, \end{cases}$$

$$\text{(5.14)}\ \varphi_1(i)\triangleq\left(\sum_{j=0}^{i-1}f_1^2(j)\right)^{\frac{1}{2}}, \qquad h(i,l,k)\triangleq\sum_{j=i\vee k}^{l}|f_2(j-1)|\lambda^{j-i}, \quad i=1,\dots,l.$$

Remark 5.3. The constant c is invariant with respect to $l,k=1,2,\dots$, and the sequences $\{f_1(i)\}$ and $\{f_2(i)\}$, but may depend on the quantities λ, $\sup_{i\ge 0}\|u_1(i)\|_8$ and $\sup_{j\ge 0}\|u_2(j)\|_8$. Theorem 1.1 of [10] is established in a continuous-parameter setting and gives general L^p-bounds on arbitrarily many multiple integrals of L-mixing processes. Theorem 5.2 stated here is a specialization of this result to the discrete-parameter case for an L^4-bound on two summations of geometrically L-mixing processes.

Since we can diagonalize $\bar{R}$ (by condition (C2)) there is no loss of generality in supposing that $d=1$. For each $\varepsilon\in(0,1)$ and $i,j=0,1,2,\dots$, define

$$\text{(5.15)}\qquad u_1(\varepsilon;i)\triangleq\xi^\varepsilon_{i+1} \qquad\text{and}\qquad u_2(j)\triangleq\tilde{R}_{j+1},$$

$$\text{(5.16)}\ f_1(\varepsilon;i)\triangleq\begin{cases}(1-\varepsilon\bar{R})^{-i-1}, & \text{if } i\ge 0,\\ 0, & \text{if } i<0,\end{cases} \qquad f_2(\varepsilon;j)\triangleq\begin{cases}(1-\varepsilon\bar{R})^{j}, & \text{if } j\ge 0,\\ 0, & \text{if } j<0,\end{cases}$$

and note by Remark 4.5 that $u_1(\varepsilon;\cdot)$ and $u_2(\cdot)$ are zero-mean geometrically L-mixing with some rate $\lambda\in(0,1)$. From (5.12), (5.15), and (5.16) we can write

$$E[|I_l^\varepsilon - I_k^\varepsilon|^4] = \varepsilon^6 E\left[\left(\sum_{j=k}^{l-1}\sum_{i=0}^{j-1} f_1(\varepsilon;i)u_1(\varepsilon;i)f_2(\varepsilon;j)u_2(j)\right)^4\right] \tag{5.17}$$

for all $1 \le k < l \le 1 + N_\varepsilon$. Now define $\psi_1(l,k;\varepsilon)$, $\psi_2(l,k;\varepsilon)$, $\varphi_1(i;\varepsilon)$, and $h(i,l,k;\varepsilon)$ exactly as in (5.13) to (5.14), but allowing for the parametrization by ε in (5.15) and (5.16). Put $\varepsilon_0 \overset{\triangle}{=} 1/2$ when $\bar{R} = 0$, and put $\varepsilon_0 \overset{\triangle}{=} 1/(2\bar{R})$ when $\bar{R} > 0$. In view of (5.14) and (5.16), we have $h(i,l,k;\varepsilon) \le \sum_{j=i\vee k}^{l} \lambda^{j-i}$, and clearly $\max_{1\le i\le N_\varepsilon}(1-\varepsilon\bar{R})^{-i} = O(1)$ for all $\varepsilon \in (0,\varepsilon_0]$. Now $\sum_{i=1}^{k}\{\sum_{j=k}^{l}\lambda^{j-i}\} = \{\sum_{i=1}^{k}\lambda^{k-i}\}(1-\lambda^{l-k+1})/(1-\lambda) = O(1-\lambda^{l-k+1}) = O(l-k)$. Thus, from (5.13) and $\lambda \in (0,1)$, there are constants $c_1, c_2 \in (0,\infty)$ such that

$$\psi_2(l,k;\varepsilon) \le c_1\left(\sum_{i=1}^{k}\sum_{j=k}^{l}\lambda^{j-i} + \sum_{i=k+1}^{l}\sum_{j=i}^{l}\lambda^{j-i}\right) \le c_2(l-k) \tag{5.18}$$

for all $\varepsilon \in (0,\varepsilon_0]$, and $1 \le k < l \le 1+N_\varepsilon$. Similarly, from the fact that $\max_{1\le i\le N_\varepsilon}(1-\varepsilon\bar{R})^{-2i} = O(1)$ for all $\varepsilon \in (0,\varepsilon_0]$, and (5.14), we get $\varphi_1^2(i;\varepsilon) \le c_3 i$ for all $\varepsilon \in (0,\varepsilon_0]$, $i = 1,\ldots,N_\varepsilon$. Then, from (5.13), and $h(i,l,k;\varepsilon) \le \sum_{j=i}^{l}\lambda^{j-i}$ (when $i \ge k$),

$$\psi_1^2(l,k;\varepsilon) \le c_4 \sum_{i=k}^{l}\left(\sum_{j=i}^{l}\lambda^{j-i}\right)\varphi_1^2(i;\varepsilon) \le c_5\sum_{i=k}^{l} i \le c_6(l^2-k^2) \tag{5.19}$$

for all $\varepsilon \in (0,\varepsilon_0]$, and $1 \le k < l \le 1 + N_\varepsilon$, where $c_3, c_4, c_5, c_6 \in (0,\infty)$ are constants. Using Theorem 5.2, (5.17), (5.18), and (5.19), we find a constant $c_7 \in (0,\infty)$ such that

$$E[|I_l^\varepsilon - I_k^\varepsilon|^4] \le c_7\varepsilon^6[(l^2-k^2)^2 + (l-k)^4] \le 2c_7\varepsilon^6[l^2-k^2]^2 \tag{5.20}$$

for all $\varepsilon \in (0,1]$, and $1 \le k < l \le 1 + N_\varepsilon$. Since $I_1^\varepsilon = 0$ (see (5.2), (5.8)), from (5.20) and Theorem 8.1(ii) (with $\gamma \overset{\triangle}{=} 2$, $\nu \overset{\triangle}{=} 4$, $h(i,j) \overset{\triangle}{=} j^2 - i^2$, $1 \le i \le j \le 1 + N_\varepsilon$), there are constants $c_8, c_9 \in (0,\infty)$ such that

$$E\left[\max_{1\le k\le N_\varepsilon+1}|I_k^\varepsilon|^4\right] = E\left[\max_{1\le k\le N_\varepsilon+1}|I_k^\varepsilon - I_1^\varepsilon|^4\right] \le c_8\varepsilon^6[h(1,1+N_\varepsilon)]^2 \le c_9\varepsilon^2 \tag{5.21}$$

for all $\varepsilon \in (0,\varepsilon_0]$. From (5.21) (with $\varepsilon \overset{\triangle}{=} 1/n$) and the Borel–Cantelli theorem,

$$\lim_{n\to\infty}\max_{1\le k\le n+1}\left|I_k^{1/n}\right| = 0 \quad \text{a.s.},$$

and thus

$$\lim_{\varepsilon\searrow 0}\max_{1\le k\le N_\varepsilon+1}\left|I_k^{1/N_\varepsilon}\right| = 0 \quad \text{a.s.} \tag{5.22}$$

To get the first limit of (5.11) from (5.22) we must fill the gaps between successive $1/N_\varepsilon$. For this we observe

$$\max_{1\le k\le N_\varepsilon+1}|I_k^\varepsilon| \le \max_{1\le k\le N_\varepsilon+1}\left|I_k^{1/N_\varepsilon}\right| + B_\varepsilon^1 + B_\varepsilon^2, \tag{5.23}$$

which results from the triangle inequality, (5.12), $N_\varepsilon^{\frac{3}{2}}\varepsilon^{\frac{3}{2}} \leq 1$, and the definition of B_ε^1 and B_ε^2 in (5.24) and (5.25). Now we need the following lemma.

Lemma 5.4 (proved in section 7). *Suppose* (C1)–(C3) *of section* 2. *Let* Γ_j^ε *be defined by* (5.12) *and* $N_\varepsilon \overset{\triangle}{=} \lfloor \varepsilon^{-1} \rfloor$. *Then* (a) *we have*

$$(5.24)\quad B_\varepsilon^1 \overset{\triangle}{=} \varepsilon^{\frac{3}{2}} \max_{1\leq k\leq N_\varepsilon+1} \left| \sum_{j=0}^{k-1} \tilde{R}_{j+1}((I-\varepsilon\bar{R})^j - (I-N_\varepsilon^{-1}\bar{R})^j)\Gamma_j^{N_\varepsilon^{-1}} \right| \ll \varepsilon^{\frac{1}{2}},$$

$$(5.25)\quad B_\varepsilon^2 \overset{\triangle}{=} \varepsilon^{\frac{3}{2}} \max_{1\leq k\leq N_\varepsilon+1} \left| \sum_{j=0}^{k-1} \tilde{R}_{j+1}(I-\varepsilon\bar{R})^j(\Gamma_j^\varepsilon - \Gamma_j^{N_\varepsilon^{-1}}) \right| \ll \varepsilon^{\frac{3}{14}},$$

and (b) *we have bounds identical to* (5.24) *and* (5.25)*, but with* $\hat{R}_{j+1}$ *in place of* $\tilde{R}_{j+1}$.

The first limit of (5.11) follows from (5.22), (5.23), and Lemma 5.4 (a). In the same way, using condition (C3b) and Lemma 5.4(b), we can establish the second limit in (5.11) (the proof is similar to, but easier than that of the first limit in (5.11) since the matrices $\hat{R}_{j+1}$ in the definition of J_k^ε are nonrandom, whereas the $\tilde{R}_{j+1}$ in the definition of I_k^ε are random—see (5.8)). □

6. Proof of Theorem 4.4.

Notation 6.1. In this section we shall require the following additional notation: $\mathcal{Z} \overset{\triangle}{=} \{\ldots,-2,-1,0,1,2,\ldots\}$, $\mathcal{Z}_+ \overset{\triangle}{=} \{1,2,\ldots\}$. Also, $\#A$ indicates the cardinality of a finite set A. For positive integer m, let $\Pi_2^m(P_1,P_2)$ denote the Prohorov distance between probability measures P_1 and P_2 on the metric space $\Re^m$ with metric given by the norm $|\cdot|$ (see Notation 2.1), and let $\mathcal{N}_m(b,Q)$ denote the normal distribution in $\Re^m$ with m-dimensional mean vector b and covariance Q.

Remark 6.2. Theorem 4.4 is a functional CLT with rate of convergence for the process $\{W^\varepsilon(\tau),\ \tau\in[0,T]\}$, which is derived from summing the geometrically L-mixing random vectors $\{\xi_n^\varepsilon\}$ (see Remark 4.5), and may be regarded as a generalization to a function-space setting of [11, Lemma A.2.1]. This latter result gives rates of convergence in a classical (i.e., nonfunctional) CLT for a sum of geometrically L-mixing random vectors in terms of bounds on characteristic functions over finite-dimensional Euclidean space. The function-space result is considerably more difficult to establish because the Prohorov distance for probability measures on $C[0,T]$ does not relate nicely to characteristic functions. Theorem 4.4 also extends the result of Yurinskii [27, section 2], which is a functional CLT with rate of convergence for sums of independent random vectors, and bears clear similarities to [14, Lemma A6.1], which is a functional CLT with rate of convergence for a sum of strong mixing random vectors subject to quite stringent boundedness conditions that do not generally apply to algorithms. We emphasize that Theorem 4.4 involves a combination of a *function-space* rate of convergence (as contrasted with the classical rate in [11]) for a sum of *dependent* random vectors (compared with the independent case in [27]), and subject to only weak boundedness (as contrasted with uniform boundedness in [14]). The combination of all these elements presents technical challenges not found in [11], [14], or [27], and is the main reason for the somewhat lengthy proofs of this section.

Without loss of generality we shall prove Theorem 4.4 with $T=1$. For all $k=1,2,\ldots$ and $\varepsilon\in(0,1]$ define the kd-dimensional vectors Ξ_k^ε and $\hat{\Xi}_k^0$ by

$$\Xi_k^\varepsilon \triangleq \begin{bmatrix} W^\varepsilon(\frac{1}{k}) \\ W^\varepsilon(\frac{2}{k}) \\ \vdots \\ W^\varepsilon(1) \end{bmatrix} \quad \text{and} \quad \hat{\Xi}_k^0 \triangleq \begin{bmatrix} \hat{W}^0(\frac{1}{k}) \\ \hat{W}^0(\frac{2}{k}) \\ \vdots \\ \hat{W}^0(1) \end{bmatrix}. \tag{6.1}$$

In order to establish Theorem 4.4 we need the following CLT giving weak convergence of Ξ_k^ε to $\hat{\Xi}_k^0$ as $\varepsilon \to 0$ and $k \to \infty$, where k increases slowly enough that it is much less than ε^{-1} (the slowly increasing k is needed when we establish weak convergence in a function space setting).

THEOREM 6.3. *Let Ξ_k^ε and $\hat{\Xi}_k^0$ be as defined in* (6.1). *Under conditions* (C1)–(C4) *of section* 2, *there are constants $c \in (0,\infty)$ and $\varepsilon_0 \in (0,1)$ such that, for all $\varepsilon \in (0,\varepsilon_0]$ and $k=1,2,\dots,\lfloor \varepsilon^{-\frac{1}{36}} \rfloor$, we have*

$$\Pi_2^{kd}(\mathcal{L}(\Xi_k^\varepsilon), \mathcal{L}(\hat{\Xi}_k^0)) \le c\varepsilon^{\frac{1}{48}}. \tag{6.2}$$

Proof of Theorem 6.3. Fix arbitrary $\varepsilon \in (0,1]$ and $k=1,2,\dots$. Recalling that we take $T=1$ in the proof, we can use (4.7) to write the lth element $W^\varepsilon(l/k)$ of Ξ_k^ε in the form

$$W^\varepsilon(l/k) = \varepsilon^{\frac{1}{2}} \left(\sum_{j=1}^{l} \left[\sum_{\nu \in J_j^{\varepsilon,k}} \xi_\nu^\varepsilon \right] + V_l^{\varepsilon,k} \right) \qquad \forall\, l = 1,2,\dots,k, \tag{6.3}$$

where $V_l^{\varepsilon,k}$ is a small "interpolation term" given by

$$V_l^{\varepsilon,k} \triangleq \left(\frac{l}{\varepsilon k} - \left\lfloor \frac{l}{\varepsilon k} \right\rfloor \right) \xi^\varepsilon_{\lfloor l/(\varepsilon k) \rfloor + 1} \qquad \forall\, l = 1,2,\dots,k, \tag{6.4}$$

and $J_j^{\varepsilon,k}$ is the block of consecutive integers given by

$$J_j^{\varepsilon,k} \triangleq \left\{ 1 + \left\lfloor \frac{j-1}{\varepsilon k} \right\rfloor, \dots, \left\lfloor \frac{j}{\varepsilon k} \right\rfloor \right\} \qquad \forall\, j = 1,2,\dots,k. \tag{6.5}$$

We will call the $J_j^{\varepsilon,k}$ the "basic blocks." We see from (6.3) that the lth row of Ξ_k^ε involves summation over the basic blocks $J_1^{\varepsilon,k}, \dots J_l^{\varepsilon,k}$.

Remark 6.4. The proof is based on the Markov–Bernstein technique of long blocks alternated with short blocks. The basic idea is as follows. If $\sum_{\nu=1}^m \rho_\nu$ is a sum of weakly-dependent random vectors $\{\rho_\nu\}$ (e.g., the elements of a geometric L-mixing sequence), then we can study its asymptotic properties by partitioning the interval of summation $1,2,\dots,m$ into long blocks G_i^m of integers, all of equal length p_m, separated by short blocks H_i^m of integers, likewise of equal length q_m, giving a pattern of adjacent blocks $G_1^m, H_1^m, G_2^m, H_2^m, \dots, G_r^m, H_r^m, H_{r+1}^m$. Here the integer $r \equiv r_m$ is equal to the *maximum* number of long block/short block pairs (of total length $p_m + q_m$) which one can fit completely into the interval of summation $1,2,\dots,m$, and H_{r+1}^m is a "remainder block" whose cardinality is less than that of a long block/short block pair, namely $p_m + q_m$. On the one hand, if the short blocks H_i^m are long enough, then terms of the form $\sum_{\nu \in G_i^m} \rho_\nu$ and $\sum_{\nu \in G_j^m} \rho_\nu$ for $i \ne j$ involve widely separated (hence approximately independent) summands, and are thus themselves almost independent. On the other hand, if the long blocks are much longer than the short blocks then the contribution to the total sum of the ρ_ν for ν belonging to the short blocks H_i^m is negligible, and thus $\sum_{\nu=1}^m \rho_\nu \approx \sum_{i=1}^{r_m}\{\sum_{\nu\in G_i^m}\rho_\nu\}$. We thus have a sum (over $i = 1,2,\dots,r_m$) of almost independent terms on the right-hand side, and we can use this independence to study the limiting properties of the original sum of random vectors.

In our problem matters are slightly more complicated because we do not have a fixed interval of summation. Instead, different entries $W^\varepsilon(l/k)$ of the vector Ξ_k^ε involve different intervals of summation, as may be seen from (6.3). We therefore partition not the overall intervals of summation for $W^\varepsilon(l/k)$, but rather the basic blocks $J_j^{\varepsilon,k}$ in (6.3), into adjacent long and short blocks $G_{i,j}^{\varepsilon,k}$ and $H_{i,j}^{\varepsilon,k}$ of consecutive integers ordered according to the pattern

$$G_{1,j}^{\varepsilon,k},\ H_{1,j}^{\varepsilon,k},\ G_{2,j}^{\varepsilon,k},\ H_{2,j}^{\varepsilon,k},\ G_{3,j}^{\varepsilon,k}, H_{3,j}^{\varepsilon,k},\dots, G_{r,j}^{\varepsilon,k},\ H_{r,j}^{\varepsilon,k},\ H_{r+1,j}^{\varepsilon,k}. \tag{6.6}$$

The blocks $G_{i,j}^{\varepsilon,k}$, $i = 1,2,\dots,r$ have common cardinality $p_{\varepsilon,k}$ and the alternate blocks $H_{i,j}^{\varepsilon,k}$, $i = 1,2,\dots,r$ have common cardinality $q_{\varepsilon,k}$ given by

$$p_{\varepsilon,k} \triangleq \lfloor k^{-1}\varepsilon^{-\frac{1}{3}}\rfloor, \qquad q_{\varepsilon,k} \triangleq \lfloor k^{-1}\varepsilon^{-\frac{1}{9}}\rfloor, \tag{6.7}$$

while r is chosen to completely fit the largest number of consecutive pairs of blocks $G_{i,j}^{\varepsilon,k}, H_{i,j}^{\varepsilon,k}$ (of total length $p_{\varepsilon,k} + q_{\varepsilon,k}$) into $J_j^{\varepsilon,k}$, namely

$$r \equiv r_{\varepsilon,k} \triangleq \lfloor (\# J_j^{\varepsilon,k})/(p_{\varepsilon,k} + q_{\varepsilon,k})\rfloor. \tag{6.8}$$

The $G_{i,j}^{\varepsilon,k}$ are long blocks and the $H_{i,j}^{\varepsilon,k}$ are short blocks for all $i = 1,2,\dots,r$, while the last block $H_{r+1,j}^{\varepsilon,k}$ is a "remainder block" whose cardinality is less than $p_{\varepsilon,k} + q_{\varepsilon,k}$. Thus,

$$\#\left(G_{i,j}^{\varepsilon,k}\right) = p, \quad \#\left(H_{i,j}^{\varepsilon,k}\right) = q, \quad \text{and} \quad 0 \le \#\left(H_{r+1,j}^{\varepsilon,k}\right) < p + q, \tag{6.9}$$

where, for brevity, p and q are now used for $p_{\varepsilon,k}$ and $q_{\varepsilon,k}$, respectively. Denoting the first integer of the block $G_{i,j}^{\varepsilon,k}$ by $b_{i,j}^{\varepsilon,k}$, from (6.6) we see that

$$b_{i,j}^{\varepsilon,k} \triangleq 1 + \left\lfloor \frac{j-1}{k\varepsilon} \right\rfloor + (i-1)(p+q) \qquad \forall\, i = 1,\dots,r+1, \tag{6.10}$$

and, for each $j = 1,2,\dots,k$, we clearly have $G_{i,j}^{\varepsilon,k} = [b_{i,j}^{\varepsilon,k}\ ,\ b_{i,j}^{\varepsilon,k} + p)$ and $H_{i,j}^{\varepsilon,k} = [b_{i,j}^{\varepsilon,k} + p\ ,\ b_{i,j}^{\varepsilon,k} + p + q)$, for all $i = 1,\dots,r$, while $H_{r+1,j}^{\varepsilon,k} = [b_{r,j}^{\varepsilon,k} + p + q\ ,\ j/(k\varepsilon)]$.

Remark 6.5. To summarize, the lth entry $W^\varepsilon(l/k)$ of Ξ_k^ε is partitioned into a sum of ξ_ν^ε over the basic blocks $J_j^{\varepsilon,k}$ indexed by $j = 1,2,\dots,l$ (see (6.3)), and each basic block $J_j^{\varepsilon,k}$ is itself partitioned into a sequence of long block/short block pairs $G_{i,j}^{\varepsilon,k}, H_{i,j}^{\varepsilon,k}$, indexed by $i = 1,2,\dots,r_{\varepsilon,k}$, and a remainder block $H_{r+1,j}^{\varepsilon,k}$ (see (6.6)).

For each $j=1,\dots,k$ and $i=1,\dots,r$, define the d-vectors

$$Y_{i,j}^{\varepsilon,k} \triangleq \sum_{\nu\in G_{i,j}^{\varepsilon,k}} \xi_\nu^\varepsilon, \qquad Z_{i,j}^{\varepsilon,k} \triangleq \sum_{\nu\in H_{i,j}^{\varepsilon,k}} \xi_\nu^\varepsilon, \qquad Z_{r+1,j}^{\varepsilon,k} \triangleq \sum_{\nu\in H_{r+1,j}^{\varepsilon,k}} \xi_\nu^\varepsilon, \tag{6.11}$$

and observe from (6.3) that

$$W^\varepsilon(l/k) = \varepsilon^{\frac{1}{2}} \sum_{j=1}^{l} \left\{ \sum_{i=1}^{r} (Y_{i,j}^{\varepsilon,k} + Z_{i,j}^{\varepsilon,k}) + Z_{r+1,j}^{\varepsilon,k} \right\} + \varepsilon^{\frac{1}{2}} V_l^{\varepsilon,k} \tag{6.12}$$

for all $l = 1, 2, \dots, k$. Next, define the kd-vector $\tilde{Y}^{\varepsilon,k}_{i,j}$ by concatenating the d-dimensional zero vector $(j-1)$ times, and then by joining to this the $(k-j+1)$-fold concatenation of the d-vector $Y^{\varepsilon,k}_{i,j}$, namely, for each $j = 1, 2, \dots, k$ and $i = 1, 2, \dots, r$,

$$\tilde{Y}^{\varepsilon,k}_{i,j} \triangleq (\underbrace{0, 0, \dots, 0}_{d(j-1)\ \text{0's}}, \underbrace{(Y^{\varepsilon,k}_{i,j})', \dots, (Y^{\varepsilon,k}_{i,j})'}_{(k-j+1)\ Y^{\varepsilon,k}_{i,j}\text{'s}})'. \tag{6.13}$$

Likewise, for each $j = 1, 2, \dots, k$ and $i = 1, 2, \dots, r+1$, put

$$\tilde{Z}^{\varepsilon,k}_{i,j} \triangleq (\underbrace{0, 0, \dots, 0}_{d(j-1)\ \text{0's}}, \underbrace{(Z^{\varepsilon,k}_{i,j})', \dots, (Z^{\varepsilon,k}_{i,j})'}_{(k-j+1)\ Z^{\varepsilon,k}_{i,j}\text{'s}})', \tag{6.14}$$

and let $\tilde{V}^{\varepsilon,k}$ be the kd-vector formed by concatenating the d-vectors $V^{\varepsilon,k}_l$, $l = 1, 2, \dots, k$:

$$\tilde{V}^{\varepsilon,k} \triangleq ((V^{\varepsilon,k}_1)', (V^{\varepsilon,k}_2)', \dots, (V^{\varepsilon,k}_k)')'. \tag{6.15}$$

From (6.1), (6.12), and (6.13) to (6.15), we find

$$\Xi^{\varepsilon}_k = \varepsilon^{\frac{1}{2}} \sum_{j=1}^{k} \left\{ \sum_{i=1}^{r} \tilde{Y}^{\varepsilon,k}_{i,j} + \sum_{i=1}^{r+1} \tilde{Z}^{\varepsilon,k}_{i,j} \right\} + \varepsilon^{\frac{1}{2}} \tilde{V}^{\varepsilon,k}. \tag{6.16}$$

Remark 6.6. In the context of summing independent random vectors, Yurinskii [27] introduced the trick of adding a concatenation of zero-vectors to get sums of vectors of common length kd. This motivates the definitions of (6.13) and (6.14). Observe, from (6.11), (6.13), and (6.14), that for all $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, k$, the kd-vectors $\tilde{Y}^{\varepsilon,k}_{i,j}$ and $\tilde{Z}^{\varepsilon,k}_{i,j}$ are derived by effectively summing ξ^{ε}_{ν} over the long blocks $G^{\varepsilon,k}_{i,j}$ and short blocks $H^{\varepsilon,k}_{i,j}$, respectively, while $\tilde{Z}^{\varepsilon,k}_{r+1,j}$ is obtained by summing ξ^{ε}_{ν} over the (infrequent) remainder blocks $H^{\varepsilon,k}_{r+1,j}$. In the light of (6.16), this suggests

$$\Xi^{\varepsilon}_k \approx \varepsilon^{\frac{1}{2}} \sum_{j=1}^{k} \sum_{i=1}^{r} \tilde{Y}^{\varepsilon,k}_{i,j},$$

and, since $G^{\varepsilon,k}_{i,j}$ is separated from $G^{\varepsilon,k}_{i_1,j_1}$ by blocks of length q or more when $(i,j) \neq (i_1, j_1)$, it seems plausible that the $\tilde{Y}^{\varepsilon,k}_{i,j}$ for different (i,j) are approximately independent. For this intuition to help in establishing Theorem 6.3 we must use the fact that $\{\xi^{\varepsilon}_{\nu}\}$ is geometrically L-mixing (see Remark 4.5). To this end, with q given by (6.7), and recalling (4.8), for each $j=1, \dots, k$ and $i=1, \dots, r$, define the d-vectors

$$Y^{\varepsilon,k}_{i,j}[q] \triangleq \sum_{\nu \in G^{\varepsilon,k}_{i,j}} \xi^{\varepsilon}_{\nu}[q], \qquad Z^{\varepsilon,k}_{i,j}[q] \triangleq \sum_{\nu \in H^{\varepsilon,k}_{i,j}} \xi^{\varepsilon}_{\nu}[q], \qquad Z^{\varepsilon,k}_{r+1,j}[q] \triangleq \sum_{\nu \in H^{\varepsilon,k}_{r+1,j}} \xi^{\varepsilon}_{\nu}[q]. \tag{6.17}$$

Motivated by (6.13) to (6.15), for each $j = 1, 2, \dots, k$, put

$$\tilde{Y}^{\varepsilon,k}_{i,j}[q] \triangleq (\underbrace{0, 0, \dots, 0}_{d(j-1)\ \text{0's}}, \underbrace{(Y^{\varepsilon,k}_{i,j}[q])', \dots, (Y^{\varepsilon,k}_{i,j}[q])'}_{(k-j+1) Y^{\varepsilon,k}_{i,j}[q]\text{'s}})' \quad \forall\, i = 1, \dots, r, \tag{6.18}$$

$$\tilde{Z}^{\varepsilon,k}_{i,j}[q] \triangleq (\underbrace{0,0,\dots,0}_{d(j-1)\ \text{0's}}, \underbrace{(Z^{\varepsilon,k}_{i,j}[q])',\dots,(Z^{\varepsilon,k}_{i,j}[q])'}_{(k-j+1) Z^{\varepsilon,k}_{i,j}[q]\text{'s}})' \quad \forall\, i=1,\dots,r+1, \tag{6.19}$$

$$\tilde{V}^{\varepsilon,k}[q] \triangleq ((V^{\varepsilon,k}_1[q])', (V^{\varepsilon,k}_2[q])', \dots, (V^{\varepsilon,k}_k[q])')', \tag{6.20}$$

where (motivated by (6.4)) $V^{\varepsilon,k}_l[q] \triangleq (l/(k\varepsilon) - \lfloor l/(\varepsilon k)\rfloor)\xi^{\varepsilon}_{\lfloor l/(\varepsilon k)\rfloor+1}[q]$ for all $l = 1,2,\dots,k$. Also put

$$\Xi^{\varepsilon}_k[q] \triangleq \varepsilon^{\frac{1}{2}} \sum_{j=1}^{k} \left\{ \sum_{i=1}^{r} \tilde{Y}^{\varepsilon,k}_{i,j}[q] + \sum_{i=1}^{r+1} \tilde{Z}^{\varepsilon,k}_{i,j}[q] \right\} + \varepsilon^{\frac{1}{2}} \tilde{V}^{\varepsilon,k}[q]. \tag{6.21}$$

Using the triangle inequality for the Prohorov metric, we can write

$$\begin{aligned} \Pi^{kd}_2(\mathcal{L}(\Xi^{\varepsilon}_k),\mathcal{L}(\hat{\Xi}^0_k)) \le{}& \Pi^{kd}_2(\mathcal{L}(\Xi^{\varepsilon}_k), \mathcal{L}(\Xi^{\varepsilon}_k[q])) \\ &+ \Pi^{kd}_2\Bigg(\mathcal{L}(\Xi^{\varepsilon}_k[q]),\ \mathcal{L}\Bigg(\varepsilon^{\frac{1}{2}} \sum_{j=1}^{k}\sum_{i=1}^{r} \tilde{Y}^{\varepsilon,k}_{i,j}[q]\Bigg)\Bigg) \\ &+ \Pi^{kd}_2\Bigg(\mathcal{L}\Bigg(\varepsilon^{\frac{1}{2}} \sum_{j=1}^{k}\sum_{i=1}^{r} \tilde{Y}^{\varepsilon,k}_{i,j}[q]\Bigg),\ \mathcal{N}_{kd}\Bigg(0, \varepsilon \sum_{j=1}^{k}\sum_{i=1}^{r} \operatorname{cov}(\tilde{Y}^{\varepsilon,k}_{i,j}[q])\Bigg)\Bigg) \\ &+ \Pi^{kd}_2\Bigg(\mathcal{N}_{kd}\Bigg(0, \varepsilon \sum_{j=1}^{k}\sum_{i=1}^{r} \operatorname{cov}(\tilde{Y}^{\varepsilon,k}_{i,j}[q])\Bigg),\ \mathcal{L}(\hat{\Xi}^0_k)\Bigg). \end{aligned} \tag{6.22}$$

To get Theorem 6.3 we must establish upper bounds for each term on the right-hand side of (6.22). We will write c, c_1, etc. for nonnegative finite constants that may vary from one use to the next.

First term on RHS of (6.22). Define the d-vectors $W^{\varepsilon}_q(l/k) \triangleq \varepsilon^{\frac{1}{2}} \sum_{\nu=1}^{\lfloor l/(k\varepsilon)\rfloor} \xi^{\varepsilon}_{\nu}[q] + V^{\varepsilon,k}_l[q]$ for all $l = 1,2,\dots,k$, (compare (4.7)). Using Cauchy–Schwarz for discrete sums and (4.9), we get constant c such that

$$E[|W^{\varepsilon}(l/k) - W^{\varepsilon}_q(l/k)|^2] \le \varepsilon E\left[\left(1 + \frac{l}{\varepsilon k}\right) \sum_{\nu=1}^{1+\lfloor l/(\varepsilon k)\rfloor} |\xi^{\varepsilon}_{\nu} - \xi^{\varepsilon}_{\nu}[q]|^2\right] \le c\varepsilon^{-1}\lambda^q \tag{6.23}$$

for each $\varepsilon \in (0,1]$, $k = 1,2,\dots$, and $1 \le l \le k$. From (6.21) it follows that $\Xi^{\varepsilon}_k[q]$ is given by the concatenation of the d-vectors $W^{\varepsilon}_q(l/k)$, $l = 1,2,\dots,k$ (in just the same way that Ξ^{ε}_k is given in (6.1) by the concatenation of the d-vectors $W^{\varepsilon}(l/k)$, $l = 1,\dots,k$), and so, from (6.23):

$$\|\Xi^{\varepsilon}_k - \Xi^{\varepsilon}_k[q]\|_2 = \left(\sum_{l=1}^{k} E[|W^{\varepsilon}(l/k) - W^{\varepsilon}_q(l/k)|^2]\right)^{\frac{1}{2}} \le c^{\frac{1}{2}} k^{\frac{1}{2}} \varepsilon^{-\frac{1}{2}} \lambda^{\frac{q}{2}}. \tag{6.24}$$

By (6.24) and Lemma 8.6(ii), for all $\varepsilon \in (0,1]$ and $k = 1,2,\dots$, we get

$$\Pi^{kd}_2(\mathcal{L}(\Xi^{\varepsilon}_k), \mathcal{L}(\Xi^{\varepsilon}_k[q])) \le (c^{\frac{1}{2}} k^{\frac{1}{2}} \varepsilon^{-\frac{1}{2}} \lambda^{\frac{q}{2}})^{\frac{2}{3}} = c^{\frac{1}{3}} k^{\frac{1}{3}} \varepsilon^{-\frac{1}{3}} \lambda^{\frac{q}{3}}. \tag{6.25}$$

Second term on RHS of (6.22). From (6.21), (6.19), and (6.20), one easily verifies that the kd-dimensional vector $(\Xi^{\varepsilon}_k[q] - \varepsilon^{\frac{1}{2}} \sum_{j=1}^{k} \sum_{i=1}^{r} \tilde{Y}^{\varepsilon,k}_{i,j}[q])$ is the concatenation of the sequence of d-dimensional vectors $\varepsilon^{\frac{1}{2}}(\sum_{j=1}^{m} \sum_{i=1}^{r+1} Z^{\varepsilon,k}_{i,j}[q] + V^{\varepsilon,k}_m[q])$ for all $m = 1, 2, \ldots, k$, and thus

$$(6.26)\quad \left\|\Xi_k^\varepsilon[q] - \varepsilon^{\frac{1}{2}} \sum_{j=1}^{k}\sum_{i=1}^{r} \tilde{Y}_{i,j}^{\varepsilon,k}[q]\right\|_2 = \varepsilon^{\frac{1}{2}}\left(\sum_{m=1}^{k}\left\|\sum_{i=1}^{r+1}\sum_{j=1}^{m} Z_{i,j}^{\varepsilon,k}[q] + V_m^{\varepsilon,k}[q]\right\|_2^2\right)^{\frac{1}{2}}$$

$$\leq \varepsilon^{\frac{1}{2}} \sum_{m=1}^{k}\left(\left\|\sum_{i=1}^{r}\sum_{j=1}^{m} Z_{i,j}^{\varepsilon,k}[q]\right\|_2 + \left\|\sum_{j=1}^{m} Z_{r+1,j}^{\varepsilon,k}[q]\right\|_2 + \left\|V_m^{\varepsilon,k}[q]\right\|_2\right),$$

where we have used the inequality $(\sum_{m=1}^{k} a_i^2)^{\frac{1}{2}} \leq \sum_{m=1}^{k} |a_i|$ and then Minkowski's inequality to get the inequality in (6.26). To bound the terms in brackets on the right of (6.26) we shall assume that $d = 1$; the general multivariate case only involves more cumbersome notation. Since $\{\xi_\nu^\varepsilon[q]\}$ is a q-dependent zero-mean process (see Remark 4.5) and distinct blocks $H_{i,j}^{\varepsilon,k}$ are separated by long blocks of cardinality greater than q (recall (6.6) and (6.9)), we see from (6.17) that $Z_{i,j}^{\varepsilon,k}[q]$ and $Z_{i_1,j_1}^{\varepsilon,k}[q]$ are independent zero-mean random variables when $(i, j) \neq (i_1, j_1)$, and thus

$$(6.27)\qquad \left\|\sum_{i=1}^{r}\sum_{j=1}^{m} Z_{i,j}^{\varepsilon,k}[q]\right\|_2 = \left(\sum_{i=1}^{r}\sum_{j=1}^{m} E|Z_{i,j}^{\varepsilon,k}[q]|^2\right)^{1/2}.$$

Now $\{\xi_\nu^\varepsilon\}$ is a geometrically L-mixing zero-mean process; hence Remark 4.6, (6.9), and (6.11) give a constant c such that $\|Z_{i,j}^{\varepsilon,k}\|_2 \leq cq^{\frac{1}{2}}$ for all $\varepsilon \in (0, 1]$, for all $k = 1, 2, \ldots$, for all $i = 1, \ldots, r$, and for all $j = 1, \ldots, k$. Also, from (4.9), (6.11), (6.17), and Minkowski's inequality, we find $\|Z_{i,j}^{\varepsilon,k} - Z_{i,j}^{\varepsilon,k}[q]\|_2 \leq c_1 q\lambda^q$, and thus

$$(6.28)\qquad \|Z_{i,j}^{\varepsilon,k}[q]\|_2 \leq \|Z_{i,j}^{\varepsilon,k}\|_2 + \|Z_{i,j}^{\varepsilon,k}[q] - Z_{i,j}^{\varepsilon,k}\|_2 \leq cq^{\frac{1}{2}} + c_1 q\lambda^q \leq c_2 q^{\frac{1}{2}}$$

for some constants c_1, c_2. Hence, by (6.27), we get $\|\sum_{i=1}^{r}\sum_{j=1}^{m} Z_{i,j}^{\varepsilon,k}[q]\|_2 \leq c_2(rmq)^{\frac{1}{2}}$ for all $\varepsilon \in (0, 1]$, for all $k = 1, 2, \ldots$, and for all $m = 1, \ldots, k$. Similarly, we have $\|\sum_{j=1}^{m} Z_{r+1,j}^{\varepsilon,k}[q]\|_2 \leq c_2(m(p+q))^{\frac{1}{2}}$. Now, from (6.8), (6.7), and $\#J_j^{\varepsilon,k} = \lfloor k^{-1}\varepsilon^{-1}\rfloor$ (see (6.5)), we can find some $\varepsilon_2 \in (0, 1]$ such that

$$(6.29)\qquad r \equiv r_{\varepsilon,k} \leq 2\varepsilon^{-\frac{2}{3}} \qquad \forall\, k = 1, 2, \ldots, \left\lfloor \varepsilon^{-\frac{1}{36}}\right\rfloor, \quad \forall\, \varepsilon \in (0, \varepsilon_2]$$

(e.g., $\varepsilon_2 \overset{\triangle}{=} 1/16$). But clearly $\|V_m^{\varepsilon,k}[q]\|_2 = O(1)$ (uniformly with respect to ε, k, m); thus from (6.26), (6.7),

$$(6.30)\quad \left\|\Xi_k^\varepsilon[q] - \varepsilon^{\frac{1}{2}} \sum_{j=1}^{k}\sum_{i=1}^{r} \tilde{Y}_{i,j}^{\varepsilon,k}[q]\right\|_2 \leq c_3\varepsilon^{\frac{1}{2}} \sum_{m=1}^{k}\{(rmq)^{\frac{1}{2}} + (m(p+q))^{\frac{1}{2}} + O(1)\}$$

$$\leq c_4\varepsilon^{\frac{1}{2}}k\{(\varepsilon^{-\frac{2}{3}}kk^{-1}\varepsilon^{-\frac{1}{9}})^{\frac{1}{2}} + (kk^{-1}\varepsilon^{-\frac{1}{3}})^{\frac{1}{2}} + O(1)\} \leq c_5 k\varepsilon^{\frac{1}{9}}$$

for some constants c_3, c_4, c_5; hence, by Lemma 8.6(ii), for each $\varepsilon \in (0, \varepsilon_2]$ we have

$$(6.31)\qquad \Pi_2^{kd}\left(\mathcal{L}(\Xi_k^\varepsilon[q]), \mathcal{L}\left(\varepsilon^{\frac{1}{2}} \sum_{j=1}^{k}\sum_{i=1}^{r} \tilde{Y}_{i,j}^{\varepsilon,k}[q]\right)\right) \leq (c_5 k\varepsilon^{\frac{1}{9}})^{\frac{2}{3}} = c_5^{\frac{2}{3}} k^{\frac{2}{3}} \varepsilon^{\frac{2}{27}}$$

for all $k = 1, 2, \ldots, \lfloor \varepsilon^{-\frac{1}{36}}\rfloor$.

Third term on RHS of (6.22). Here we shall need the following special case of a finite-dimensional CLT for sums of independent random vectors due to Yurinskii [27, Theorem 1]. Note that the upper-bound we state here is somewhat less precise than the bound given in [27, Theorem 1] but is easier to use.

THEOREM 6.7. *There is a constant* $\gamma \in (0,\infty)$ *with the following property. For all positive integers* m *and finite sequences* $\{u_1, u_2, \dots, u_n\}$ *of zero-mean independent* $\Re^m$*-valued random vectors we have*

$$\Pi_2^m\left(\mathcal{L}\left(\sum_{\alpha=1}^{n} u_\alpha\right), \mathcal{N}_m\left(0, \sum_{\alpha=1}^{n} \operatorname{cov}(u_\alpha)\right)\right) \le \gamma m^{\frac{1}{3}}(\mu_n)^{\frac{2}{9}} \quad \textit{when} \quad \mu_n \overset{\triangle}{=} \sum_{\alpha=1}^{n} E|u_\alpha|^3 < 1.$$

The $\tilde{Y}_{i,j}^{\varepsilon,k}[q]$ given by (6.18) are zero-mean and independent as i and j vary (since $\{\xi_\nu^\varepsilon[q]\}$ is a q-dependent process, and the long blocks $G_{i,j}^{\varepsilon,k}$ are separated by either short blocks $H_{i,j}^{\varepsilon,k}$ or short block/remainder block pairs $H_{r,j}^{\varepsilon,k}, H_{r+1,j}^{\varepsilon,k}$, all of which have length at least q). In Theorem 6.7 we will identify m with kd, n with kr, the summation index α with (i,j) (for $j = 1, \dots, k$ and $i = 1, \dots, r$), u_α with $\varepsilon^{\frac{1}{2}}\tilde{Y}_{i,j}^{\varepsilon,k}[q]$, and μ_n with $\mu_{kr} \overset{\triangle}{=} \sum_{j=1}^{k}\sum_{i=1}^{r} \varepsilon^{\frac{3}{2}} E|\tilde{Y}_{i,j}^{\varepsilon,k}[q]|^3$. From (6.18) we have $|\tilde{Y}_{i,j}^{\varepsilon,k}[q]|^2 = (k - j+1)|Y_{i,j}^{\varepsilon,k}[q]|^2 \le k|Y_{i,j}^{\varepsilon,k}[q]|^2$, and, exactly as for (6.28), we can use Remark 4.6 to find constant c_1 such that $\|Y_{i,j}^{\varepsilon,k}[q]\|_3 \le c_1 p^{\frac{1}{2}}$ for all $\varepsilon \in (0,1]$, for all $k = 1, 2, \dots$, for all $i = 1, \dots, r$, and for all $j = 1, \dots, k$. Thus, from (6.7), we have $E|\tilde{Y}_{i,j}^{\varepsilon,k}[q]|^3 \le c_1^3 k^{\frac{3}{2}} p^{\frac{3}{2}} = c_1^3 \varepsilon^{-\frac{1}{2}}$; hence (6.29) gives $\mu_{kr} = \sum_{j=1}^{k}\sum_{i=1}^{r} \varepsilon^{\frac{3}{2}} E|\tilde{Y}_{i,j}^{\varepsilon,k}[q]|^3 \le 2c_1^3 k \varepsilon^{\frac{1}{3}}$. Define

$$\tilde{T}^{\varepsilon,k} \overset{\triangle}{=} \varepsilon \sum_{j=1}^{k}\sum_{i=1}^{r} \operatorname{cov}(\tilde{Y}_{i,j}^{\varepsilon,k}[q]) \equiv \sum_{\alpha} \operatorname{cov}(u_\alpha), \tag{6.32}$$

and fix some $\varepsilon_3 \in (0, \varepsilon_2]$ such that $2c_1^3 k \varepsilon^{\frac{1}{3}} < 1$ for all $\varepsilon \in (0, \varepsilon_3]$, $k = 1, 2, \dots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$. Then Theorem 6.7 gives a constant c_2 such that, for all $\varepsilon \in (0, \varepsilon_3]$, $k = 1, 2, \dots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$, we have

$$\Pi_2^{kd}\left(\mathcal{L}\left(\varepsilon^{\frac{1}{2}} \sum_{j=1}^{k}\sum_{i=1}^{r} \tilde{Y}_{i,j}^{\varepsilon,k}[q]\right), \mathcal{N}_{kd}(0, \tilde{T}^{\varepsilon,k})\right) \le c_1 (kd)^{\frac{1}{3}} (k\varepsilon^{\frac{1}{3}})^{\frac{2}{9}} = c_2 k^{\frac{5}{9}} \varepsilon^{\frac{2}{27}}. \tag{6.33}$$

Fourth term on RHS of (6.22). To simplify the notation, we use the following convention. If B is a $kd \times kd$ matrix, then $B_{m,n}$ denotes the (m,n)th of the $d \times d$ submatrices into which B can be partitioned for all $m, n = 1, 2, \dots, k$. If v is a kd-vector, then v_m denotes the d-vector consisting of the $(d(m-1)+1)$th to (dm)th elements of v. Fix arbitrary $\varepsilon \in (0,1]$ and $k = 1, 2, \dots$. Define the $kd \times kd$ matrices $\hat{T}^{\varepsilon,k}$ and T^k such that their (m,n)th d-dimensional submatrices are respectively given by

$$\hat{T}_{m,n}^{\varepsilon,k} \overset{\triangle}{=} \varepsilon \sum_{j=1}^{m\wedge n}\sum_{i=1}^{r} \sum_{\nu \in G_{i,j}^{\varepsilon,k}} A(\theta^0((\nu-1)\varepsilon)) \quad \text{and} \quad T_{m,n}^{k} \overset{\triangle}{=} (\operatorname{cov}\hat{\Xi}_k^0)_{m,n} \tag{6.34}$$

for all $m, n = 1, 2, \dots, k$, where $A(\cdot)$ is given by condition (C4). Then, from (6.1) and (3.8), we see that $\mathcal{L}(\hat{\Xi}_k^0) = \mathcal{N}_{kd}(0, T^k)$, so that the fourth term on the right side of (6.22) is bounded as follows:

$$\begin{aligned} \Pi_2^{kd}(\mathcal{N}_{kd}(0, \tilde{T}^{\varepsilon,k}), \mathcal{L}(\hat{\Xi}_k^0)) &\le \Pi_2^{kd}(\mathcal{N}_{kd}(0, \tilde{T}^{\varepsilon,k}), \mathcal{N}_{kd}(0, \hat{T}^{\varepsilon,k})) \\ &\quad + \Pi_2^{kd}(\mathcal{N}_{kd}(0, \hat{T}^{\varepsilon,k}), \mathcal{N}_{kd}(0, T^k)). \end{aligned} \tag{6.35}$$

We will now use Theorem 8.7 to bound each of the terms on the right side of (6.35). In order to deal with the first term we find an expression for $|\tilde{T}^{\varepsilon,k}_{m,n} - \hat{T}^{\varepsilon,k}_{m,n}|$. From (6.18),

$$[\mathrm{cov}(\tilde{Y}^{\varepsilon,k}_{i,j}[q])]_{m,n} = \begin{cases} \mathrm{cov}(Y^{\varepsilon,k}_{i,j}[q]), & \text{when } j \le (m \wedge n), \\ 0, & \text{when } (m \wedge n) < j, \end{cases}$$

and thus, using (6.32),

$$\tilde{T}^{\varepsilon,k}_{m,n} = \varepsilon \sum_{j=1}^{m\wedge n} \sum_{i=1}^{r} \mathrm{cov}(Y^{\varepsilon,k}_{i,j}[q]). \tag{6.36}$$

On combining (6.36), (6.34), and (6.17), we get, for all $m, n = 1, 2, \ldots, k$,

$$|\tilde{T}^{\varepsilon,k}_{m,n} - \hat{T}^{\varepsilon,k}_{m,n}| = \varepsilon \left| \sum_{j=1}^{m\wedge n} \sum_{i=1}^{r} \left[\mathrm{cov}\left(\sum_{\nu \in G^{\varepsilon,k}_{i,j}} \xi^{\varepsilon}_{\nu}[q] \right) - \sum_{\nu \in G^{\varepsilon,k}_{i,j}} A(\theta^0((\nu-1)\varepsilon)) \right] \right|. \tag{6.37}$$

To upper-bound the term in square braces on the right-hand side of (6.37), partition the long block $G^{\varepsilon,k}_{i,j}$ into *adjacent* blocks of integers $Q^{\varepsilon,k}_{\alpha,i,j}$ ordered according to the pattern

$$Q^{\varepsilon,k}_{1,i,j}, Q^{\varepsilon,k}_{2,i,j}, \ldots, Q^{\varepsilon,k}_{l,i,j}, Q^{\varepsilon,k}_{l+1,i,j}, \tag{6.38}$$

where $l \overset{\triangle}{=} \lfloor p/q^2 \rfloor$. The blocks $Q^{\varepsilon,k}_{\alpha,i,j}$, $\alpha = 1, \ldots, l$ have common cardinality q^2, and $Q^{\varepsilon,k}_{l+1,i,j}$ is a "remainder block" whose cardinality is less than q^2. Recalling (6.7), fix some $\varepsilon_{41} \in (0, \varepsilon_3]$ such that

$$l \equiv l_{\varepsilon,k} \overset{\triangle}{=} \lfloor p/q^2 \rfloor \le 2k\varepsilon^{-\frac{1}{9}} \quad \forall\, k = 1, 2, \ldots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor, \quad \forall\, \varepsilon \in (0, \varepsilon_{41}]. \tag{6.39}$$

If $s^{\varepsilon,k}_{\alpha,i,j}$ denotes the first integer of the block $Q^{\varepsilon,k}_{\alpha,i,j}$ then clearly

$$s^{\varepsilon,k}_{\alpha,i,j} \overset{\triangle}{=} b^{\varepsilon,k}_{i,j} + (\alpha-1)q^2 \quad \forall\, \alpha = 1, \ldots, l+1 \tag{6.40}$$

(recall that $b^{\varepsilon,k}_{i,j}$ is the first integer in $G^{\varepsilon,k}_{i,j}$), so that $Q^{\varepsilon,k}_{\alpha,i,j} \overset{\triangle}{=} [s^{\varepsilon,k}_{\alpha,i,j}, s^{\varepsilon,k}_{\alpha,i,j} + q^2)$ for all $\alpha = 1, \ldots, l$, while $Q^{\varepsilon,k}_{l+1,i,j} \overset{\triangle}{=} [s^{\varepsilon,k}_{l+1,i,j}, b^{\varepsilon,k}_{i,j} + p)$. Henceforth, to lighten the notation, we write Q_α for $Q^{\varepsilon,k}_{\alpha,i,j}$ and s_α for $s^{\varepsilon,k}_{\alpha,i,j}$ when convenient. Also, recalling (2.7) and (2.3), put

$$\tilde{H}_n(\theta)[s] \overset{\triangle}{=} E[\tilde{H}_n(\theta)|\mathcal{F}^n_{n-s}] = \tilde{b}_n[s] - \tilde{R}_n[s]\theta \quad \forall\, s, n = 1, 2, \ldots, \quad \theta \in \Re^d, \tag{6.41}$$

and observe from (4.8) that $\xi^{\varepsilon}_{\nu}[q] = \tilde{H}_{\nu}(\theta^0((\nu-1)\varepsilon))[q]$ for all $\nu \in G^{\varepsilon,k}_{i,j}$. One then sees from (6.37) that, for each $m, n = 1, \ldots, k$,

$$|\tilde{T}^{\varepsilon,k}_{m,n} - \hat{T}^{\varepsilon,k}_{m,n}| \le \varepsilon \sum_{j=1}^{m\wedge n} \sum_{i=1}^{r} \{\mathrm{I}^{\varepsilon,k}_{i,j} + \mathrm{II}^{\varepsilon,k}_{i,j} + \mathrm{III}^{\varepsilon,k}_{i,j} + \mathrm{IV}^{\varepsilon,k}_{i,j}\}, \tag{6.42}$$

where

$$\mathrm{I}_{i,j}^{\varepsilon,k} \triangleq \left|\mathrm{cov}\left(\sum_{\nu\in G_{i,j}^{\varepsilon,k}} \xi_\nu^\varepsilon[q]\right) - \sum_{\alpha=1}^{l+1}\mathrm{cov}\left(\sum_{\nu\in Q_\alpha} \xi_\nu^\varepsilon[q]\right)\right|, \tag{6.43}$$

$$\mathrm{II}_{i,j}^{\varepsilon,k} \triangleq \left|\sum_{\alpha=1}^{l+1}\left\{\mathrm{cov}\left(\sum_{\nu\in Q_\alpha} \xi_\nu^\varepsilon[q]\right) - \mathrm{cov}\left(\sum_{\nu\in Q_\alpha} \tilde{H}_\nu(\theta^0(\varepsilon s_\alpha))[q]\right)\right\}\right|, \tag{6.44}$$

$$\mathrm{III}_{i,j}^{\varepsilon,k} \triangleq \left|\sum_{\alpha=1}^{l+1}\left\{\mathrm{cov}\left(\sum_{\nu\in Q_\alpha}\tilde{H}_\nu(\theta^0(\varepsilon s_\alpha))[q]\right) - \left(\#Q_\alpha\right)A(\theta^0(\varepsilon s_\alpha))\right\}\right|, \tag{6.45}$$

$$\mathrm{IV}_{i,j}^{\varepsilon,k} \triangleq \left|\sum_{\alpha=1}^{l+1}(\#Q_\alpha)A(\theta^0(\varepsilon s_\alpha)) - \sum_{\nu\in G_{i,j}^{\varepsilon,k}} A(\theta^0((\nu-1)\varepsilon))\right|. \tag{6.46}$$

Remark 6.8. Our calculation of an upper-bound for $|\tilde{T}_{m,n}^{\varepsilon,k} - \hat{T}_{m,n}^{\varepsilon,k}|$ is closely based on the proof of Lemma 3.1 of Khas'minskii [13], which motivates introduction of the blocks Q_α and the upper-bound (6.42) in terms of the quantities (6.43) to (6.46). The idea is that over the blocks Q_α we "freeze" the time-varying quantity $\theta^0(\varepsilon(\nu-1))$, $\nu \in G_{i,j}^{\varepsilon,k}$, at the value $\theta^0(\varepsilon s_\alpha)$ corresponding to the first member s_α of Q_α. The block Q_α is long enough for the averaging postulated in condition (C4) to come into play (where we identify $\{n_0,\dots,N+n_0\}$ and θ in (2.6) with Q_α and $\theta^0(\varepsilon s_\alpha)$, respectively) and provide an upper bound for (6.45). At the same time, Q_α is short enough to ensure that $\theta^0((\nu-1)\varepsilon) \approx \theta^0(\varepsilon s_\alpha)$, and this will give us upper-bounds for (6.44) and (6.46). Finally, we shall use L-mixing of $\{\xi_\nu^\varepsilon\}$ to bound (6.43). We now proceed to make this intuition precise. For the term in (6.43), put

$$B_{i,j}^{\varepsilon,k} \triangleq (G_{i,j}^{\varepsilon,k} \times G_{i,j}^{\varepsilon,k}) - \bigcup_{\alpha=1}^{l+1}(Q_\alpha \times Q_\alpha) \tag{6.47}$$

(recall that Q_α is short for $Q_{\alpha,i,j}^{\varepsilon,k}$) and let $B_+ \triangleq \{(\nu,\mu) \in B_{i,j}^{\varepsilon,k} : \mu > \nu\}$, $D_\alpha \triangleq \{(\nu,\mu) \in \mathcal{Z}^2 : -\infty < \nu < s_\alpha \le \mu < \infty\}$ for all $\alpha = 2,\dots,l+1$. It follows from (6.47) that

$$B_+ \subset \bigcup_{\alpha=2}^{l+1} D_\alpha. \tag{6.48}$$

By (6.39), (6.48), and Remark 8.4, we have constants $\lambda \in (0,1)$ and c, c_1 with

$$\mathrm{I}_{i,j}^{\varepsilon,k} = \left|\sum_{(\nu,\mu)\in B_{i,j}^{\varepsilon,k}} E\big[(\xi_\nu^\varepsilon[q])(\xi_\mu^\varepsilon[q])'\big]\right| \le c\sum_{(\nu,\mu)\in B_+}\lambda^{\mu-\nu} \tag{6.49}$$

$$\le c\sum_{\alpha=2}^{l+1}\left\{\sum_{(\nu,\mu)\in D_\alpha}\lambda^{\mu-\nu}\right\} \le c\sum_{\alpha=2}^{l+1}\left\{\sum_{n=1}^{\infty} n\lambda^n\right\} \le c_1 l = c_1 l_{\varepsilon,k} \le c_1 k\varepsilon^{-\frac{1}{9}}$$

for all $\varepsilon \in (0,\varepsilon_{41}]$, $k=1,\dots,\lfloor\varepsilon^{-\frac{1}{36}}\rfloor$, $i=1,\dots,r$, $j=1,\dots,k$. Next, we upper-bound the term in (6.44): put $\eta_\nu^1 \triangleq \xi_\nu^\varepsilon[q] \equiv \tilde{H}_\nu(\theta^0((\nu-1)\varepsilon))[q]$ and $\eta_\nu^2 \triangleq \tilde{H}_\nu(\theta^0(\varepsilon s_\alpha))[q]$ for all $\nu \in Q_\alpha$. Then

$$\operatorname{cov}\left(\sum_{\nu\in Q_\alpha} \eta^1_\nu\right) - \operatorname{cov}\left(\sum_{\nu\in Q_\alpha} \eta^2_\nu\right) = \sum_{\nu,\mu\in Q_\alpha} E\left[(\eta^1_\nu - \eta^2_\nu)(\eta^1_\mu)' + (\eta^2_\nu)(\eta^1_\mu - \eta^2_\mu)'\right]. \tag{6.50}$$

Since $\{\theta^0(\tau),\ \tau \in [0,1]\}$ is Lipschitz continuous and $\{\tilde{R}_n[s]\}$, $\{\eta^1_n\}$ are clearly L_2-bounded (uniformly in $s, n = 1, 2, \ldots$), we find constant c such that

$$\left|E\left[(\eta^1_\nu - \eta^2_\nu)(\eta^1_\mu)'\right]\right| \leq \|\tilde{R}_\nu[q]\|_2 \, \|\eta^1_\mu\|_2 |\theta^0(\varepsilon s_\alpha) - \theta^0(\varepsilon(\nu-1))| \leq c\varepsilon(\#Q_\alpha) \tag{6.51}$$

for all $\mu,\ \nu \in Q_\alpha$, with an identical bound for $|E[(\eta^2_\nu)(\eta^1_\mu - \eta^2_\mu)']|$. From (6.39), (6.50), and (6.51), we find

$$\mathrm{II}^{\varepsilon,k}_{i,j} \leq c\varepsilon l(\#Q_\alpha)^3 \leq c\varepsilon(2k\varepsilon^{-\frac{1}{9}})(k^{-2}\varepsilon^{-\frac{2}{9}})^3 \leq c_1\varepsilon^{\frac{2}{9}} \tag{6.52}$$

for all $\varepsilon \in (0, \varepsilon_{41}]$, $k = 1, 2, \ldots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$, $i = 1, \ldots, r$, $j = 1, \ldots, k$. For the term given by (6.45), we need the following lemma.

Lemma 6.9 (proved in section 7). *Suppose conditions* (C1)–(C4) *of section* 2, *and let* $\{\tilde{H}_n(\theta)[s]\}$ *be defined by* (6.41). *Then there are constants* $C \in [0,\infty)$ *and* $\lambda \in (0,1)$ *such that*

$$\left|\operatorname{cov}\left(\sum_{n=n_0}^{N+n_0} \tilde{H}_n(\theta)[s]\right) - A(\theta)(N+1)\right| \leq C[1 + |\theta|^2][1 + (N+1)^2\lambda^s]$$

for all $\theta \in \Re^d$, $s, N, n_0 = 1, 2, 3, \ldots$.

By Lemma 6.9 (identifying θ, s, and the interval $\{n_0, \ldots, n_0 + N\}$ in Lemma 6.9 with $\theta^0(\varepsilon s_\alpha)$, q, and Q_α, respectively), uniform boundedness of $\theta^0(((n-1)\varepsilon) \wedge T)$ (in (ε, n)), and recalling $\#(Q_\alpha) = q^2$, we find constants c, c_1, such that

$$\mathrm{III}^{\varepsilon,k}_{i,j} \leq c\sum_{\alpha=1}^{l+1}[1 + q^4\lambda^q] \leq c_1 l \leq c_1 k\varepsilon^{-\frac{1}{9}} \tag{6.53}$$

for all $\varepsilon \in (0, \varepsilon_{41}]$ and $k = 1, 2, \ldots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$, $i = 1, \ldots, r$, $j = 1, \ldots, k$. Now we bound the term given by (6.46). From Lemma 2.9 we see that $\tau \to A(\theta^0(\tau))$ is globally Lipschitz continuous over the interval $\tau \in [0,1]$, and hence there are constants c, c_1 such that, for all $\varepsilon \in (0, \varepsilon_{41}]$ and $k = 1, 2, \ldots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$,

$$\begin{aligned}\mathrm{IV}^{\varepsilon,k}_{i,j} &\leq c\sum_{\alpha=1}^{l+1}\sum_{\nu\in Q_\alpha} |\varepsilon s_\alpha - (\nu-1)\varepsilon| \leq c_1\varepsilon l\,(\#Q_\alpha)^2 \\ &\leq c_1\varepsilon(k\varepsilon^{-\frac{1}{9}})(k^{-2}\varepsilon^{-\frac{2}{9}})^2 \leq c_1\varepsilon^{\frac{4}{9}}.\end{aligned} \tag{6.54}$$

Combining (6.42), and (6.49) to (6.54), and using (6.29) gives constants c, c_1, such that for all $\varepsilon \in (0, \varepsilon_{41}]$ and all $k = 1, 2, \ldots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$,

$$\begin{aligned}\max_{1\leq m,n\leq k} |\tilde{T}^{\varepsilon,k}_{m,n} - \hat{T}^{\varepsilon,k}_{m,n}| &\leq c\varepsilon\sum_{j=1}^{k}\sum_{i=1}^{r}\{k\varepsilon^{-\frac{1}{9}} + \varepsilon^{\frac{2}{9}} + k\varepsilon^{-\frac{1}{9}} + \varepsilon^{\frac{4}{9}}\} \\ &\leq c_1(\varepsilon kr)(k\varepsilon^{-\frac{1}{9}}) \leq c_1\varepsilon^{\frac{1}{6}}.\end{aligned} \tag{6.55}$$

Thus, there must exist some $\varepsilon_{42} \in (0, \varepsilon_{41}]$ such that $kd^{\frac{1}{2}} \max_{1\le m,n\le k} |\tilde{T}^{\varepsilon,k}_{m,n} - \hat{T}^{\varepsilon,k}_{m,n}| \le c_1\varepsilon^{\frac{1}{6}}kd^{\frac{1}{2}} < 1$ for all $\varepsilon \in (0, \varepsilon_{42}]$ and $k = 1, 2 \dots \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$, and we can use Theorem 8.7 with (6.55) to find a constant c such that

$$\Pi_2^{kd}(\mathcal{N}_{kd}(0, \tilde{T}^{\varepsilon,k}), \mathcal{N}_{kd}(0, \hat{T}^{\varepsilon,k})) \le ck^{\frac{3}{4}}(\varepsilon^{\frac{1}{6}})^{\frac{1}{4}} \le c\varepsilon^{\frac{1}{48}} \tag{6.56}$$

for all $\varepsilon \in (0, \varepsilon_{42}]$ and $k = 1, 2, \dots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$. To bound the second term on the right side of (6.35), fix some $1 \le m, n \le k$. One sees from (3.8) that $\{\hat{W}^0(\tau),\ 0 \le \tau \le 1\}$ is a zero-mean Gaussian process with independent increments and covariance $\int_0^\tau A(\theta^0(s))\,ds$. Thus, from (6.1) and (6.34), we have $T^k_{m,n} = \varepsilon \int_0^{m\wedge n/(k\varepsilon)} A(\theta^0(u\varepsilon))\,du$, and hence

$$\begin{aligned} T^k_{m,n} - \hat{T}^{\varepsilon,k}_{m,n} &= \varepsilon \sum_{j=1}^{m\wedge n} \sum_{i=1}^{r} \sum_{\nu \in H^{\varepsilon,k}_{i,j}} A(\theta^0((\nu - 1)\varepsilon)) \\ &\quad + \varepsilon \left(\int_0^{m\wedge n/(k\varepsilon)} [A(\theta^0(u\varepsilon)) - A(\theta^0(\lfloor u \rfloor \varepsilon))]\,du \right). \end{aligned} \tag{6.57}$$

Now, the last term on the right of (6.57) is clearly $O(\varepsilon)$ uniformly with respect to $1 \le m, n \le k \le \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$ (since Lemma 2.9 ensures that $\tau \to A(\theta^0(\tau))$ is globally Lipschitz continuous over $\tau \in [0, 1]$). Hence, using (6.57), there must be constants c, c_1 such that for $\varepsilon \in (0, \varepsilon_{42}]$ and $k = 1, 2, \dots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$, we have

$$\begin{aligned} \max_{1\le m,n\le k} |\hat{T}^{\varepsilon,k}_{m,n} - T^k_{m,n}| &\le c\varepsilon kr(\# H^{\varepsilon,k}_{i,j}) + O(\varepsilon) \\ &\le c\varepsilon k(2\varepsilon^{-\frac{2}{3}})k^{-1}\varepsilon^{-\frac{1}{9}} + O(\varepsilon) \le c_1\varepsilon^{\frac{2}{9}}. \end{aligned} \tag{6.58}$$

By (6.58) there is clearly some $\varepsilon_{43} \in (0, \varepsilon_{42}]$ such that $kd^{\frac{1}{2}} \max_{1\le m,n\le k} |\hat{T}^{\varepsilon,k}_{m,n} - T^k_{m,n}| < 1$ for all $\varepsilon \in (0, \varepsilon_{43}]$ and $k = 1, 2 \dots \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$, and hence, by Theorem 8.7 and (6.58), we get

$$\Pi_2^{kd}(\mathcal{N}_{kd}(0, \hat{T}^{\varepsilon,k}), \mathcal{N}_{kd}(0, T^k)) \le c_1k^{\frac{3}{4}}(\varepsilon^{\frac{2}{9}})^{\frac{1}{4}} \le c_1\varepsilon^{\frac{5}{144}} \tag{6.59}$$

for all $\varepsilon \in (0, \varepsilon_{43}]$ and $k = 1, 2 \dots \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$. Combining (6.32), (6.35), (6.56), and (6.59), we find a constant c such that

$$\Pi_2^{kd}\left(\mathcal{N}_{kd}\left(0, \varepsilon \sum_{j=1}^{k} \sum_{i=1}^{r} \operatorname{cov} \tilde{Y}^{\varepsilon;k}_{i,j}[q] \right), \mathcal{L}(\hat{\Xi}^0_k) \right) \le c\varepsilon^{\frac{1}{48}} \tag{6.60}$$

for all $\varepsilon \in (0, \varepsilon_{43}]$, $k = 1, 2, \dots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$. Finally, we see that the four terms on the right of (6.22) are upper-bounded by (6.25), (6.31), (6.33), and (6.60), respectively. Since $\lambda \in (0, 1)$ we have $\lambda^{\varepsilon^{-\beta}} = O(\varepsilon^\alpha)$ for constants $\alpha, \beta \in (0, \infty)$. Thus, there must be constants c, c_1, such that

$$\Pi_2^{kd}(\mathcal{L}(\Xi^\varepsilon_k), \mathcal{L}(\hat{\Xi}^0_k)) \le c(k^{\frac{1}{3}}\varepsilon^{-\frac{1}{3}}\lambda^{\frac{q}{3}} + k^{\frac{2}{3}}\varepsilon^{\frac{2}{27}} + k^{\frac{5}{9}}\varepsilon^{\frac{2}{27}} + \varepsilon^{\frac{1}{48}}) \le c_1\varepsilon^{\frac{1}{48}}$$

for all $\varepsilon \in (0, \varepsilon_{43}]$ and $k = 1, 2, \dots, \lfloor \varepsilon^{-\frac{1}{36}} \rfloor$. Now Theorem 6.3 follows with $\varepsilon_0 \overset{\triangle}{=} \varepsilon_{43}$. $\square$

Proof of Theorem 4.4. For each $k = 1, 2, 3, \dots$, $\varepsilon \in (0, 1]$ define continuous process $\{W^\varepsilon_k(\tau),\ \tau \in [0, 1]\}$ by

$$W^\varepsilon_k(\tau) \overset{\triangle}{=} \begin{cases} W^\varepsilon(\tau) & \text{for } \tau = i/k,\ i = 0, 1, \dots, k, \\ \text{linear interpolation,} & \text{otherwise.} \end{cases} \tag{6.61}$$

In the same way, define $\{\hat{W}^0_k(\tau),\ 0\le\tau\le 1\}$ in terms of $\{\hat{W}^0(\tau),\ 0\le\tau\le 1\}$. Also, put $\gamma \triangleq 72$ and $k(\varepsilon)\triangleq\lfloor\varepsilon^{-\frac{1}{\gamma}}\rfloor$ for all $\varepsilon\in(0,1]$. By the triangle inequality we can write

$$(6.62)\quad \begin{aligned}\Pi_C(\mathcal{L}(W^\varepsilon),\mathcal{L}(\hat{W}^0)) \le{}& \Pi_C(\mathcal{L}(W^\varepsilon),\mathcal{L}(W^\varepsilon_{k(\varepsilon)})) + \Pi_C(\mathcal{L}(W^\varepsilon_{k(\varepsilon)}),\mathcal{L}(\hat{W}^0_{k(\varepsilon)}))\\ &+\Pi_C(\mathcal{L}(\hat{W}^0_{k(\varepsilon)}),\mathcal{L}(\hat{W}^0)).\end{aligned}$$

Now upper-bound each term on the right side of (6.62) as follows.

Third term on RHS of (6.62). By the definition of $\{\hat{W}^0_k(\tau),\ 0\le\tau\le 1\}$ (following (6.61)), we have

$$(6.63)\quad \|\hat{W}^0-\hat{W}^0_{k(\varepsilon)}\|_C \le 2\max_{i=0,1,\dots,k(\varepsilon)-1}\left\{\max_{\frac{i}{k(\varepsilon)}\le\tau\le\frac{i+1}{k(\varepsilon)}}\left|\hat{W}^0(\tau)-\hat{W}^0\left(\frac{i}{k(\varepsilon)}\right)\right|\right\},$$

and thus, from the Chebyshev inequality,

$$(6.64)\quad \begin{aligned}&\hat{P}[\|\hat{W}^0-\hat{W}^0_{k(\varepsilon)}\|_C\ge\varepsilon^{\frac{1}{8\gamma}}]\\ &\le\sum_{i=0}^{k(\varepsilon)-1}\hat{P}\left\{\max_{\frac{i}{k(\varepsilon)}\le\tau\le\frac{(i+1)}{k(\varepsilon)}}\left|\hat{W}^0(\tau)-\hat{W}^0\left(\frac{i}{k(\varepsilon)}\right)\right|\ge\frac{1}{2}\varepsilon^{\frac{1}{8\gamma}}\right\}\\ &\le 16\varepsilon^{-\frac{1}{2\gamma}}\sum_{i=0}^{k(\varepsilon)-1}\hat{E}\left\{\max_{\frac{i}{k(\varepsilon)}\le\tau\le\frac{(i+1)}{k(\varepsilon)}}\left|\hat{W}^0(\tau)-\hat{W}^0\left(\frac{i}{k(\varepsilon)}\right)\right|^4\right\}.\end{aligned}$$

By Doob's maximal L_p inequality [8, Proposition 2.16, p. 63] applied to the nonnegative submartingale $\{|\hat{W}^0(\tau)-\hat{W}^0(i/k(\varepsilon))|,\ \tau\in[i/k(\varepsilon),1]\}$, we get

$$(6.65)\quad \begin{aligned}&\hat{E}\left\{\max_{\frac{i}{k(\varepsilon)}\le\tau\le\frac{(i+1)}{k(\varepsilon)}}\left|\hat{W}^0(\tau)-\hat{W}^0\left(\frac{i}{k(\varepsilon)}\right)\right|^4\right\}\\ &\le\left(\frac{4}{3}\right)^4\hat{E}\left|\hat{W}^0\left(\frac{i+1}{k(\varepsilon)}\right)-\hat{W}^0\left(\frac{i}{k(\varepsilon)}\right)\right|^4\ll k^{-2}(\varepsilon).\end{aligned}$$

Combining (6.64) and (6.65) gives

$$\hat{P}[\|\hat{W}^0-\hat{W}^0_{k(\varepsilon)}\|_C\ge\varepsilon^{\frac{1}{8\gamma}}]\ll k^{-1}(\varepsilon)\varepsilon^{-\frac{1}{2\gamma}}\le\varepsilon^{\frac{1}{2\gamma}}\le\varepsilon^{\frac{1}{8\gamma}},$$

and thus, by Lemma 8.6(i), we get

$$(6.66)\quad \Pi_C(\mathcal{L}(\hat{W}^0),\mathcal{L}(\hat{W}^0_{k(\varepsilon)}))\ll\varepsilon^{\frac{1}{8\gamma}}.$$

Second term on RHS of (6.62). We need the next result, which is suggested by the arguments on page 246 of Yurinskii [27].

LEMMA 6.10 (proved in section 7). *For* Ξ^ε_k *and* $\hat{\Xi}^0_k$ *given by* (6.1), *and* W^ε *and* $\hat{W}^0$ *given by* (6.61), *we have*

$$\Pi_C(\mathcal{L}(W^\varepsilon_k),\mathcal{L}(\hat{W}^0_k))\le\Pi_2^{kd}(\mathcal{L}(\Xi^\varepsilon_k),\mathcal{L}(\hat{\Xi}^0_k))\quad\forall\, k=1,2,\dots,\quad\varepsilon\in(0,1].$$

From Lemma 6.10, Theorem 6.3, and the fact that $k(\varepsilon)<\lfloor\varepsilon^{-\frac{1}{36}}\rfloor$,

$$(6.67)\quad \Pi_C(\mathcal{L}(W^\varepsilon_{k(\varepsilon)}),\mathcal{L}(\hat{W}^0_{k(\varepsilon)}))\le\Pi_2^{kd}(\mathcal{L}(\Xi^\varepsilon_{k(\varepsilon)}),\mathcal{L}(\hat{\Xi}^0_{k(\varepsilon)}))\ll\varepsilon^{\frac{1}{48}}.$$

First term on RHS of (6.62). For continuous $\varphi : [0,1] \to \Re$ and $0 \le \tau_1 \le \mu_1 < \mu_2 \le \tau_2 \le 1$ we have $\max_{\mu_1 \le \tau \le \mu_2} |\varphi(\tau) - \varphi(\mu_1)| \le 2 \max_{\tau_1 \le \tau \le \tau_2} |\varphi(\tau) - \varphi(\tau_1)|$. Thus, from (3.5) and (4.6), we get

$$\begin{aligned}
&\max_{\frac{i}{k(\varepsilon)} \le \tau \le \frac{(i+1)}{k(\varepsilon)}} \left| W^\varepsilon(\tau) - W^\varepsilon\left(\frac{i}{k(\varepsilon)}\right)\right| \\
&\quad \le 2 \max_{\lfloor \frac{i}{\varepsilon k(\varepsilon)} \rfloor \le k \le 1 + \lfloor \frac{(i+1)}{\varepsilon k(\varepsilon)} \rfloor} \left| W^\varepsilon(\varepsilon k) - W^\varepsilon\left(\varepsilon \left\lfloor \frac{i}{\varepsilon k(\varepsilon)} \right\rfloor\right)\right| \\
&\quad \le 2 \max_{1 + \lfloor \frac{i}{\varepsilon k(\varepsilon)} \rfloor \le k \le 1 + \lfloor \frac{i+1}{\varepsilon k(\varepsilon)} \rfloor} \varepsilon^{\frac{1}{2}} \left| \sum_{j = 1 + \lfloor i/(\varepsilon k(\varepsilon)) \rfloor}^{k} \xi_j^\varepsilon \right|
\end{aligned} \tag{6.68}$$

for all $i = 0, 1, \ldots, k(\varepsilon) - 1$. Now fix integers M, N with $1 \le M < N$. We see from Remark 4.6 that there is a constant $C_1 \in [0, \infty)$, not depending on the sequence $\{\xi_j^\varepsilon\}$, or M, N, such that $E[|\sum_{j=i}^k \xi_j^\varepsilon|^4] \le C_1[k - i + 1]^2$ for all $M \le i \le k \le N$ and each $\varepsilon \in (0,1]$. Hence, by Theorem 8.1(i) (with $\gamma \overset{\triangle}{=} 2$, $\nu \overset{\triangle}{=} 4$, $g(i,k) \overset{\triangle}{=} k - i + 1$, $M \le i \le k \le N$), there is some constant $A(4,2) \in [0, \infty)$ such that $E[\max_{M \le k \le N} |\sum_{j=M}^k \xi_j^\varepsilon|^4] \le A(4,2)[g(M,N)]^2$ for all $1 \le M < N$ and $\varepsilon \in (0,1]$. Taking $M \overset{\triangle}{=} 1 + \lfloor i/(\varepsilon k(\varepsilon)) \rfloor$, $N \overset{\triangle}{=} 1 + \lfloor (i+1)/(\varepsilon k(\varepsilon)) \rfloor$, we see from (6.68) that there are constants $C_2, C_3 \in [0, \infty)$ such that

$$E\left[\max_{\frac{i}{k(\varepsilon)} \le \tau \le \frac{(i+1)}{k(\varepsilon)}} \left| W^\varepsilon(\tau) - W^\varepsilon\left(\frac{i}{k(\varepsilon)}\right)\right|^4\right] \le \varepsilon^2 C_2 [g(M,N)]^2 \le C_3 k(\varepsilon)^{-2} \tag{6.69}$$

for all $\varepsilon \in (0,1]$. Now (6.63) and (6.64) continue to hold with $\hat{P}$, $\hat{E}$, and $\hat{W}^0$ replaced by P, E, and W^ε, respectively. Thus, we can repeat the argument which gave (6.66), but using (6.69) in place of (6.65), to see

$$\Pi_C(\mathcal{L}(W^\varepsilon), \mathcal{L}(W^\varepsilon_{k(\varepsilon)})) \ll \varepsilon^{\frac{1}{8\gamma}}. \tag{6.70}$$

The conclusion follows from (6.62) and the upper-bounds given by (6.66), (6.67), and (6.70). □

7. Proofs of technical lemmas.

Proof of Lemma 2.9. It is enough to take $d = 1$. From (C4) we have $A(\theta) = \lim_{N \to \infty} \operatorname{cov}(\sum_{\nu=1}^N \tilde{H}_\nu(\theta))$, and thus (see (2.7)),

$$\begin{aligned}
|A(\theta_1) - A(\theta_2)| \le 2 \limsup_{N \to \infty} \frac{1}{N} \sum_{\nu=1}^N \sum_{\nu=1}^N |E[\tilde{b}_\nu \tilde{R}_\mu]|\, |\theta_1 - \theta_2| \\
+ \limsup_{N \to \infty} \frac{1}{N} \sum_{\nu=1}^N \sum_{\nu=1}^N |E[\tilde{R}_\nu \tilde{R}_\mu]|\, |\theta_1^2 - \theta_2^2|.
\end{aligned}$$

From Lemma 8.3(i) it easily follows that the double sums on the right-hand side are $O(N)$, as required for local Lipschitz continuity. □

Proof of Lemma 4.8. Since $P(|\gamma_n| \ge n^{\frac{1}{7}}) \le E\big[|\gamma_n|^8\big]/n^{\frac{8}{7}}$, we must have that $\sum_{n=1}^\infty P(|\gamma_n| \ge n^{\frac{1}{7}}) < \infty$. By Borel–Cantelli, for a.a. ω there exists an integer $L(\omega)$ such that $|\gamma_n(\omega)| \le n^{\frac{1}{7}}$ a.s. for all $n > L(\omega)$. Hence we can find $C(\omega)$ such that $|\gamma_n| \leq C(\omega) n^{\frac{1}{7}}$ for all $n \geq 1$, and the result follows from the monotonicity of $n \to n^{\frac{1}{7}}$. □

Proof of Lemma 4.11. Recall the notation defined in (4.22) to (4.25). Clearly,

$$\max_{0 \leq \tau \leq 1} |W_1^r(\tau) - W_2^r(\tau)| \leq \varepsilon_r^{-\frac{1}{2}} \int_{\tau_r}^{1} |\zeta_{t/\varepsilon_r}^{\varepsilon_r} - \zeta_{t/\varepsilon_r}^{\varepsilon_r}[q(r)]| \, dt.$$

Hence, by Jensen's inequality,

$$\max_{0 \leq \tau \leq 1} |W_1^r(\tau) - W_2^r(\tau)|^4 \leq \varepsilon_r^{-1} \int_0^{1/\varepsilon_r} I[\tau_r/\varepsilon_r, 1/\varepsilon_r](u) \, |\zeta_u^{\varepsilon_r} - \zeta_u^{\varepsilon_r}[q(r)]|^4 \, du.$$

Then, by (4.23), (4.22), (4.9), the fact that $(\tau_r/\varepsilon_r) \geq 1 + q(r)$, and Fubini,

$$E\left[\max_{0 \leq \tau \leq 1} |W_1^r(\tau) - W_2^r(\tau)|^4\right] \ll \varepsilon_r^{-1} \int_0^{1/\varepsilon_r} \lambda^{4q(r)} \, du \ll \exp(2(r^\sigma - r^2)).$$

By the mean value theorem for the mapping $\alpha \to r^\alpha$ (r constant), there exists $\gamma \in [\sigma, 2]$ such that $r^\sigma - r^2 = (\sigma - 2)(\ln r) r^\gamma \leq (\sigma - 2) r^\sigma$, so $\exp\left(2(r^\sigma - r^2)\right) \leq \varepsilon_r^{2(2-\sigma)} \leq \varepsilon_r$ (since $\sigma < 3/2$). □

Proof of Lemma 4.12. Recall the notation defined in (4.22)–(4.25). From (3.5),

$$W^{\varepsilon_r}(\tau) = \varepsilon_r^{\frac{1}{2}} \int_0^{\tau/\varepsilon_r} \zeta_u^{\varepsilon_r} \, du, \quad 0 \leq \tau \leq 1. \tag{7.1}$$

Thus, from (7.1) and (4.24),

$$W^{\varepsilon_r}(\tau) - W_1^r(\tau) = \begin{cases} \varepsilon_r^{\frac{1}{2}} \displaystyle\int_0^{\tau/\varepsilon_r} \zeta_u^{\varepsilon_r} \, du, & \text{if } 0 \leq \tau \leq \tau_r, \\ \varepsilon_r^{\frac{1}{2}} \displaystyle\int_0^{\tau_r/\varepsilon_r} \zeta_u^{\varepsilon_r} \, du, & \text{if } \tau_r \leq \tau \leq 1. \end{cases} \tag{7.2}$$

Therefore, from (7.2) and (4.23) we have

$$\max_{0 \leq \tau \leq 1} |W^{\varepsilon_r}(\tau) - W_1^r(\tau)| = \varepsilon_r^{\frac{1}{2}} \max_{1 \leq k \leq 1 + \lfloor \tau_r/\varepsilon_r \rfloor} \left| \sum_{j=1}^{k} \xi_j^{\varepsilon_r} \right|. \tag{7.3}$$

By Remark 4.6 and Theorem 8.1(i) (with $\nu = 4$, $\gamma = 2$, and $g(i,j) \overset{\triangle}{=} j - i + 1$), we easily see there is a constant $c \in [0, \infty)$, not depending on r, such that

$$E\left[\max_{1 \leq k \leq 1 + \lfloor \tau_r/\varepsilon_r \rfloor} \left| \sum_{j=1}^{k} \xi_j^{\varepsilon_r} \right|^4\right] \leq c \left(\frac{\tau_r}{\varepsilon_r}\right)^2. \tag{7.4}$$

Combining (7.4) with (7.3) and using the fact that $(\tau_r/\varepsilon_r) \ll \varepsilon_{r-1}^{-1}$ gives

$$\begin{aligned} E\left[\max_{0 \leq \tau \leq 1} |W^{\varepsilon_r}(\tau) - W_1^r(\tau)|^4\right] &\ll \varepsilon_r^2 \left(\frac{\tau_r}{\varepsilon_r}\right)^2 \\ &\ll \left(\frac{\varepsilon_r}{\varepsilon_{r-1}}\right)^2 = \exp\left(2(r-1)^\sigma - 2r^\sigma\right). \end{aligned} \tag{7.5}$$

Now $r^\sigma - (r-1)^\sigma = \sigma s^{\sigma-1}$ for some $s \in [r-1, r]$, and hence $\exp(2((r-1)^\sigma - r^\sigma)) \le \exp(-2\sigma(r-1)^{\sigma-1})$. Observe that, since $\sigma > 1$, we have $5\ln r \ll \sigma(r-1)^{\sigma-1}$, or, equivalently, $\exp\left(-\sigma(r-1)^{\sigma-1}\right) \ll r^{-5}$ for all $r \ge 1$. Now the result follows from (7.5). □

Proof of Lemma 5.4. We use the notation defined in the proof of Lemma 3.6. For (5.24) we have

$$\varepsilon^{\frac{3}{2}} \max_{1\le k\le N_\varepsilon+1} \left| \sum_{j=0}^{k-1} \tilde{R}_{j+1} \left((I-\varepsilon\bar{R})^j - (I-N_\varepsilon^{-1}\bar{R})^j\right) \Gamma_j^{N_\varepsilon^{-1}} \right| \tag{7.6}$$

$$\le \varepsilon^{\frac{3}{2}} N_\varepsilon \cdot \max_{0\le j\le N_\varepsilon} |\tilde{R}_{j+1}| \cdot \max_{0\le j\le N_\varepsilon} \left|(I-\varepsilon\bar{R})^j - (I-N_\varepsilon^{-1}\bar{R})^j\right| \cdot \max_{0\le j\le N_\varepsilon} |\Gamma_j^{N_\varepsilon^{-1}}|.$$

We bound the $\max_j |\cdot|$ factors on the right of (7.6). For the third factor, we see from $\max_{1\le i\le n} |I - n^{-1}\bar{R}|^{-2i} = O(1)$ (uniformly in n) and Remark 4.6, that there is a constant $c_1 \in (0,\infty)$ such that

$$E\left[\left|\sum_{k=i}^{j} (I - n^{-1}\bar{R})^{-k} \xi_k^{n^{-1}}\right|^8\right] \le c_1(j-i+1)^4 \quad \forall\, 1\le j\le k\le n. \tag{7.7}$$

From (5.12), (7.7), and Theorem 8.1(i) (with $\nu \overset{\triangle}{=} 8$, $\gamma \overset{\triangle}{=} 4$, $g(i,j) \overset{\triangle}{=} j-i+1$, $1 \le i \le j \le n$), we have

$$E\left[\max_{0\le j\le n} \left|\Gamma_j^{n^{-1}}\right|^8\right] \ll n^4, \tag{7.8}$$

and thus, from Borel–Cantelli,

$$\max_{0\le j\le n} |\Gamma_j^{n^{-1}}| = O(n^{\frac{6}{7}}) \text{ a.s.} \tag{7.9}$$

so that the third $\max_j |\cdot|$ factor on the right of (7.6) is $O(\varepsilon^{-\frac{6}{7}})$ a.s. It follows from Lemma 4.8 that the first $\max_j |\cdot|$ factor on the right of (7.6) is $O(\varepsilon^{-\frac{1}{7}})$ a.s. while the second $\max_j |\cdot|$ factor is easily shown by Taylor's theorem to be $O(\varepsilon)$. Thus, from (7.6), the quantity in (5.24) is a.s. $O(\varepsilon^{\frac{1}{2}}\varepsilon^{-\frac{1}{7}}\varepsilon\varepsilon^{-\frac{6}{7}}) = O(\varepsilon^{\frac{1}{2}})$. The proofs for (5.25) and part (b) are similar and are omitted. □

Proof of Lemma 6.9. From Remark 2.7 we know that $\{\tilde{b}_n\}$ and $\{\tilde{R}_n\}$ are geometrically L-mixing and zero-mean. Thus (recalling (2.7) and (6.41)), there are constants $\lambda \in (0,1)$ and $C_1 \in [0,\infty)$ such that $\|\tilde{H}_n(\theta)\|_2 \le C_1[1+|\theta|]$, $\|\tilde{H}_n(\theta)[s]\|_2 \le C_1[1+|\theta|]$, and $\|\tilde{H}_n(\theta) - \tilde{H}_n(\theta)[s]\|_2 \le C_1[1+|\theta|]\lambda^s$ for all $\theta \in \Re^d$, for all $s, n = 1, 2, \ldots$. Put

$$D^1(n,m,s,\theta) \overset{\triangle}{=} (\tilde{H}_n(\theta) - \tilde{H}_n(\theta)[s])(\tilde{H}_m(\theta))'.$$

By Cauchy–Schwarz and the preceding bounds,

$$|ED^1(n,m,s,\theta)| \le \|\tilde{H}_n(\theta) - \tilde{H}_n(\theta)[s]\|_2 \|\tilde{H}_m(\theta)\|_2 \le C_2[1+|\theta|^2]\lambda^s,$$

and the same upper-bound clearly holds for the expectation of

$$D^2(n,m,s,\theta) \overset{\triangle}{=} \tilde{H}_n(\theta)[s](\tilde{H}_m(\theta) - \tilde{H}_m(\theta)[s])'.$$

Then

$$\begin{aligned}(7.10)\qquad &\left|\operatorname{cov}\left(\sum_{n=n_0}^{N+n_0}\tilde{H}_n(\theta)\right)-\operatorname{cov}\left(\sum_{n=n_0}^{N+n_0}\tilde{H}_n(\theta)[s]\right)\right|\\ &=\left|\sum_{n=n_0}^{N+n_0}\sum_{m=n_0}^{N+n_0}E[D^1(n,m,s,\theta)+D^2(n,m,s,\theta)]\right|\\ &\le 2C_2[1+|\theta|^2](N+1)^2\lambda^s.\end{aligned}$$

The result now follows from condition (C4) and (7.10). □

Proof of Lemma 6.10. Fix some $k=1,2,\ldots$, and let $|\cdot|_\infty$ denote the maximum norm on $\Re^{kd}$ (i.e., $|x|_\infty \overset{\triangle}{=} \max_{1\le i\le kd}|x^i|$). Let $\Pi_\infty^{kd}(P_1,P_2)$ denote Prohorov distance between probability measures P_1 and P_2 on $\Re^{kd}$ with norm $|\cdot|_\infty$, and let $C_k[0,1]$ be the subset of $C[0,1]$ comprising all continuous functions $f:[0,1]\to\Re^d$, with $f(0)=0$, which are piecewise linear with break-points at i/k, $i=1,\ldots,k-1$ (see right-hand side of (6.61)). Then the metric spaces $(\Re^{kd},|\cdot|_\infty)$ and $(C_k[0,1],\|\cdot\|_C)$ are homeomorphic. Since the paths of $\{W_k^\varepsilon(\tau),\ \tau\in[0,1]\}$ and $\{\hat{W}_k^0(\tau),\ \tau\in[0,1]\}$ are in $C_k[0,1]$, it easily follows from (6.1) and (6.61) that $\Pi_C(\mathcal{L}(W_k^\varepsilon),\mathcal{L}(\hat{W}_k^0))=\Pi_\infty^{kd}(\mathcal{L}(\Xi_k^\varepsilon),\mathcal{L}(\hat{\Xi}_k^0))$. Now, for $\eta>0$ and closed $A\subset\Re^{kd}$, put $A^\eta\overset{\triangle}{=}\{x\in\Re^{kd}:|x-a|<\eta$ for some $a\in A\}$ and $A_\infty^\eta\overset{\triangle}{=}\{x\in\Re^{kd}:|x-a|_\infty<\eta$ for some $a\in A\}$. Since $|x|_\infty\le|x|$, we have $A^\eta\subset A_\infty^\eta$, and from this, together with (8.1), we get $\Pi_\infty^{kd}(\mathcal{L}(\Xi_k^\varepsilon),\mathcal{L}(\hat{\Xi}_k^0))\le\Pi_2^{kd}(\mathcal{L}(\Xi_k^\varepsilon),\mathcal{L}(\hat{\Xi}_k^0))$, as required. □

8. Useful results. In this section we collect for easy reference some simple adaptations of results from probability which are used in the previous sections. The following maximal inequalities, due to Móricz [23, Theorem 1] and Longnecker and Serfling [22], are needed for lines (5.21), (6.69), (7.4), and (7.8).

THEOREM 8.1. *Suppose that M and N are integers, $1\le M<N<\infty$, and $\mathcal{Y}$ is a normed vector space with norm $\|\cdot\|$.*

(i) *Let $\{z_k,\ k=M,M+1,\ldots,N\}$ be arbitrary $\mathcal{Y}$-valued random variables. Suppose there are constants $c,\nu\in(0,\infty)$, $\gamma\in(1,\infty)$, and an $\Re$-valued mapping $g(i,j)$ defined for $M\le i\le j\le N$, such that $E[\|\sum_{k=i}^j z_k\|^\nu]\le c[g(i,j)]^\gamma$ for all $M\le i\le j\le N$, and $g(i,j)+g(j+1,k)\le g(i,k)$ for all $M\le i\le j<k\le N$. Then there is a constant $A(\nu,\gamma)\in[0,\infty)$ such that $E[\max_{M\le n\le N}\|\sum_{k=M}^n z_k\|^\nu]\le cA(\nu,\gamma)[g(M,N)]^\gamma$. The constant $A(\nu,\gamma)$ depends on ν and γ only.*

(ii) *Let $\{Q_k,\ k=M,M+1,\ldots,N\}$ be arbitrary $\mathcal{Y}$-valued random variables. Suppose there are constants $c,\nu\in(0,\infty)$, $\gamma\in(1,\infty)$, and an $\Re$-valued mapping $h(i,j)$ defined for $M\le i\le j\le N$, such that $E[\|Q_j-Q_i\|^\nu]\le c[h(i,j)]^\gamma$ for all $M\le i\le j\le N$, and $h(i,j)+h(j,k)\le h(i,k)$ for all $M\le i\le j\le k\le N$. Then, for the constant $A(\nu,\gamma)$ of (i), we have $E[\max_{M\le n\le N}\|Q_n-Q_M\|^\nu]\le cA(\nu,\gamma)[h(M,N)]^\gamma$.*

Remark 8.2. The constant $A(\nu,\gamma)$ has nothing to do with M,N, the random variables $\{z_k\}$, $\{Q_k\}$, or the functions $g(\cdot,\cdot)$, $h(\cdot,\cdot)$ in (i) and (ii). Indeed, (2.1)–(2.3) of [22] give $A(\nu,\gamma)$ explicitly as a function of $\nu>0$ and $\gamma>1$ only. Notice that (ii) follows upon applying (i) of Theorem 8.1 to the sequence $\{z_k,\ k=M,M+1,\ldots,N-1\}$ defined by $z_k\overset{\triangle}{=}Q_{k+1}-Q_k$ for all $k=M,M+1,\ldots,N-1$, with $g(i,j)\overset{\triangle}{=}h(i,j+1)$ for all $M\le i\le j\le N-1$.

By trivially adapting the arguments for Lemma A.1.2(a)(b) of [11], we get the following lemma.

Lemma 8.3. *Suppose that* $\{z_n^1,\ n=1,2,\ldots\}$ *and* $\{z_n^2,\ n=1,2,\ldots\}$ *are* $\Re$*-valued zero-mean geometrically* L*-mixing processes with respect to the system* $(\mathcal{F}_n,\mathcal{F}_n^+)$*. Then there are constants* $C_1,C_2\in[0,\infty)$ *and* $\lambda\in(0,1)$ *such that* (i) $|E\{z_n^1 z_m^2\}|\le C_1\lambda^{|n-m|}$ *for all* $m,n=1,2,\ldots,$ *and* (ii) $|E\{z_n^1[s]z_m^2[s]\}|\le C_2\lambda^{|n-m|}$ *for all* $m,n,s=1,2,\ldots$ *(recall* (2.3)*).*

Remark 8.4. Since $\theta^0((n-1)\varepsilon\wedge T)$ is uniformly bounded (in $\varepsilon\in(0,1]$, $n=1,2,\ldots$) and $\{\tilde{b}_n\}$ and $\{\tilde{R}_n\}$ are zero-mean geometrically L-mixing (by Remark 2.7), we see from (4.8) and Lemma 8.3(ii) that there are constants $C\in[0,\infty)$ and $\lambda\in(0,1)$ such that $|E[(\xi_\nu^\varepsilon[s])(\xi_\mu^\varepsilon[s])']|\le C\ \lambda^{|\mu-\nu|}$ for all $\varepsilon\in(0,1]$, for all $s,\mu,\nu=1,2,\ldots$.

The following result, which is a special case of Theorem 1.1 of Gerencsér [9], is repeatedly used.

Theorem 8.5. *Suppose that* $\{u_j\}$ *is a zero-mean* $\Re^{l\times r}$*-valued geometrically* L*-mixing process, and* $\{A_j\}$ *is a nonrandom sequence of* $d\times l$ *matrices. Then, for each* $p\in[2,\infty)$, $\|\sum_{j=1}^N A_ju_j\|_p\ll(\sum_{j=1}^N|A_j|^2)^{\frac{1}{2}}$*. The constant implied by* $\ll$ *depends only upon* p, $\sup_{j\ge1}\|u_j\|_p$*, and the rate* λ *of the geometrically* L*-mixing process* $\{u_j\}$*, in particular, does NOT depend on* N *or the sequence* $\{A_j\}$*.*

In the remainder of this section we summarize some relevant facts about the *Prohorov metric* for probability measures on a metric space. Suppose that (S,ρ) is a metric space, and let P_1 and P_2 be two probability measures on $(S,\mathcal{B}(S))$. Define the number

$$\Pi(P_1,P_2)\overset{\triangle}{=}\inf\{\eta\in(0,\infty):\ P_1(A)\le P_2(A^\eta)+\eta\ \text{ for all closed } A\subset S\}, \tag{8.1}$$

where $A^\eta\overset{\triangle}{=}\{s\in S:\ \rho(s,a)\le\eta\ \text{ for some } a\in A\}$ for $A\subset S$. All we need to know is that the mapping $\Pi(\cdot,\cdot)$ defined by (8.1) is indeed a metric in the set of all probability measures on $(S,\mathcal{B}(S))$, called the *Prohorov metric*, and when (S,ρ) is a *separable* metric space then $\Pi(\cdot,\cdot)$ is a metric for the topology of weak convergence of probability measures on $(S,\mathcal{B}(S))$. See [8, Theorem 3.1, p. 108]. Thus, one can use the Prohorov metric to quantify *rates* of weak convergence of probability measures. In fact, this is the significance of condition (4.3) in Theorem 4.2. We also make repeated use of the following simple result which is an immediate consequence of (8.1) and the Chebyshev inequality.

Lemma 8.6. *Suppose that* X *and* Y *are random variables defined on* $(\Omega,\mathcal{F},P)$ *with values in a metric space* (S,ρ)*.* (i) *If, for some* $\beta\in(0,\infty)$*, we have* $P\{\rho(X,Y)\ge\beta\}\le\beta$*, then* $\Pi(\mathcal{L}(X),\mathcal{L}(Y))\le\beta$*.* (ii) *If, for some* $\beta\in(0,\infty)$ *and* $c\in[1,\infty)$*, we have* $\|\rho(X,Y)\|_c\le\beta$*, then* $\Pi(\mathcal{L}(X),\mathcal{L}(Y))\le\beta^{\frac{c}{c+1}}$*.*

The next result (used for (6.56) and (6.59)) is a special case of Dehling [7, Theorem 7, p. 400], and upper-bounds the Prohorov distance between Gaussian distributions on Euclidean space.

Theorem 8.7. *There exists a constant* $\beta\in[0,\infty)$ *such that*

$$\Pi_2^{kd}\left(\mathcal{N}_{kd}(0,T),\mathcal{N}_{kd}(0,S)\right)\le\beta k^{\frac{3}{4}}d^{\frac{5}{8}}\left(\max_{1\le m,n\le k}|T_{m,n}-S_{m,n}|\right)^{\frac{1}{4}}$$

for all $k,d=1,2,\ldots,$ *and all* $kd\times kd$ *symmetric positive semidefinite matrices* T *and* S *such that* $kd^{\frac{1}{2}}\max_{1\le m,n\le k}|T_{m,n}-S_{m,n}|<1$*.*

Acknowledgment. Many thanks indeed to the referee for an exceptionally careful and detailed review which led to numerous improvements in this work.

REFERENCES

[1] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, 1990.

[2] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[3] J. A. Bucklew, T. G. Kurtz, and W. A. Sethares, *Weak convergence and local stability properties of fixed step size recursive algorithms*, IEEE Trans. Inform. Theory, 30 (1993), pp. 966–978.

[4] P. E. Caines, *Linear Stochastic Systems*, Wiley, New York, 1988.

[5] J. Chover, *On Strassen's version of the log log law*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 8 (1967), pp. 83–90.

[6] M. H. A. Davis and R. B. Vinter, *Stochastic Modelling and Control*, Chapman and Hall, London, UK, 1985.

[7] H. Dehling, *Limit theorems for sums of weakly dependent Banach space valued random variables*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 63 (1983), pp. 393–432.

[8] S. N. Ethier and T. G. Kurtz, *Markov Processes Characterization and Convergence*, Wiley, New York, 1986.

[9] L. Gerencsér, *On a class of mixing processes*, Stochastics Stochastics Rep., 26 (1989), pp. 165–191.

[10] L. Gerencsér, *Multiple integrals with respect to L-mixing processes*, Statist. Probab. Lett., 17 (1993), pp. 73–85.

[11] A. J. Heunis, *Asymptotic properties of prediction error estimators in approximate system identification*, Stochastics, 24 (1988), pp. 1–43.

[12] A. J. Heunis, *Rates of convergence for an adaptive filtering algorithm driven by stationary dependent data*, SIAM J. Control Optim., 32 (1994), pp. 116–139.

[13] R. Z. Khas'minskii, *On stochastic processes defined by differential equations with a small parameter*, Theory Probab. Appl., 11 (1966), pp. 211–228.

[14] M. A. Kouritzin and A. J. Heunis, *Rates of convergence in a central limit theorem for stochastic processes defined by differential equations with a small parameter*, J. Multivariate Anal., 43 (1992), pp. 58–109.

[15] M. A. Kouritzin and A. J. Heunis, *A law of the iterated logarithm for stochastic processes defined by differential equations with a small parameter*, Ann. Probab., 22 (1994), pp. 659–679.

[16] J. Kuelbs, *A strong convergence theorem for Banach space valued random variables*, Ann. Probab., 4 (1976), pp. 744–771.

[17] P. R. Kumar and P.P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[18] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.

[19] H. J. Kushner and A. Shwartz, *Weak convergence and asymptotic properties of adaptive filters with constant gains*, IEEE Trans. Inform. Theory, 30 (1984), pp. 177–182.

[20] L. Ljung, *Convergence analysis of parametric identification methods*, IEEE Trans. Automat. Control, 23 (1978), pp. 770–783.

[21] L. Ljung and P. E. Caines, *Asymptotic normality of prediction error estimators for approximate system models*, Stochastics, 3 (1979), pp. 29–46.

[22] M. Longnecker and R. J. Serfling, *General moment and probability inequalities for the maximum partial sum*, Studia Sci. Math. Hungar., 30 (1977), pp. 129–133.

[23] F. A. Móricz, *Moment inequalities and the strong law of large numbers*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 35 (1976), pp. 299–314.

[24] H. Pezezhki-Esfahani and A.J. Heunis, *Strong diffusion approximations for recursive stochastic algorithms*, IEEE Trans. Inform. Theory, 43 (1997), pp. 512–523.

[25] V. Strassen, *An invariance principle for the law of the iterated logarithm*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 3 (1964), pp. 211–226.

[26] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, UK, 1991.

[27] V. V. Yurinskii, *On the error of the Gaussian approximation for convolutions*, Theory Probab. Appl., 22 (1977), pp. 236–247.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 571–593

PROPERTIES OF A MULTIVALUED MAPPING ASSOCIATED WITH SOME NONMONOTONE COMPLEMENTARITY PROBLEMS*

YUN-BIN ZHAO† AND GEORGE ISAC‡

Abstract. Using the homotopy invariance property of the degree and a newly introduced concept of the interior-point-ε-exceptional family for continuous functions, we prove an alternative theorem concerning the existence of a certain interior-point of a continuous complementarity problem. Based on this result, we develop several sufficient conditions to assure some desirable properties (nonemptyness, boundedness, and upper-semicontinuity) of a multivalued mapping associated with continuous (nonmonotone) complementarity problems corresponding to semimonotone, $\mathrm{P}(\tau,\alpha,\beta)$-, quasi-$\mathrm{P}_*$-, and exceptionally regular maps. The results proved in this paper generalize well-known results on the existence of central paths in continuous P_0 complementarity problems.

Key words. nonlinear complementarity problems, central path, interior-point-ε-exceptional family, weakly univalent maps, generalized monotonicity

AMS subject classifications. 90C30, 90C33

PII. S0363012998345196

1. Introduction.

Consider the nonlinear complementarity problem (NCP)

$$f(x) \geq 0,\ x \geq 0,\ x^T f(x) = 0,$$

where f is a continuous function from R^n into itself. This problem has now gained much importance because of its many applications in optimization, economics, engineering, etc. (see [8, 12, 16, 18]).

There are several equivalent formulations of the NCP in the form of a nonlinear equation $F(x) = 0$, where F is a continuous function from R^n into R^n. Given such an equation $F(x) = 0$, the most used technique is to perturb F to a certain F_ε, where ε is a positive parameter, and then consider the equation $F_\varepsilon(x) = 0$. If $F_\varepsilon(x) = 0$ has a unique solution denoted by $x(\varepsilon)$ and $x(\varepsilon)$ is continuous in ε, then the solutions $\{x(\varepsilon)\}$ describe, depending on the nature of $F_\varepsilon(x)$, a short path denoted by $\{x(\varepsilon) : \varepsilon \in (0, \bar{\varepsilon}]\}$, or a long path $\{x(\varepsilon) : \varepsilon \in (0, \infty)\}$. If a short path $\{x(\varepsilon) : \varepsilon \in (0, \bar{\varepsilon}]\}$ is bounded, then for any subsequence $\{\varepsilon_k\}$ with $\varepsilon_k \to 0$, the sequence $\{x(\varepsilon_k)\}$ has at least one accumulation point, and by the continuity each of the accumulation points is a solution to the NCP. Thus, a path can be viewed as a certain continuous curve associated with the solution set of the NCP. Based on the path, we may construct various computational methods for solving the NCP, such as interior-point path-following methods (see, e.g., [15, 25, 26, 27, 28, 32, 39]), regularization methods (see [8, 10, 11, 41]), and noninterior path-following methods (see [1, 2, 3, 5, 7, 17, 21]). The most common interior-point path-following method is based on the central path. The curve $\{x(\varepsilon) : \varepsilon \in (0, \infty)\}$ is said to be the central path if for each $\varepsilon > 0$ the vector $x(\varepsilon)$ is the unique solution to the system

$$x(\varepsilon) > 0,\ f(x(\varepsilon)) > 0,\ X(\varepsilon)f(x(\varepsilon)) = \varepsilon e, \tag{1}$$

*Received by the editors September 28, 1998; accepted for publication (in revised form) March 14, 2000; published electronically September 20, 2000.
http://www.siam.org/journals/sicon/39-2/34519.html

†Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, China (zyb@lsec.cc.ac.cn, ybzhao@se.cuhk.edu.hk).

‡Department of Mathematics and Computer Sciences, Royal Military College of Canada, Kingston, Ontario, K7K 7B4, Canada (isac-g@rmc.ca).

where $X(\varepsilon) = \operatorname{diag}(x(\varepsilon)), e = (1, \ldots, 1)^T$, and $x(\cdot)$ is continuous on $(0, \infty)$.

In the case when f is a monotone function and the NCP is strictly feasible (i.e., there is a vector $u \in R^n$ such that $u > 0$ and $f(u) > 0$), the existence of the central path is well known (see, for example, [14, 25, 30, 31]). This existence result has been extended to some nonmonotone complementarity problems. Kojima, Mizuno, and Noma [27] proved that the central path exists if f is a uniform-P function. If f is a P_0-function satisfying a properness condition and the NCP is strictly feasible, Kojima, Megiddo, and Noma [25] showed that there exists a class of interior-point trajectories which includes the central path as a special case. If f is a P_0-function and NCP has a nonempty and bounded solution set, Chen, Chen, and Kanzow [4] and Gowda and Tawhid [13] proved that the NCP has a short central path $\{x(\varepsilon) : \varepsilon \in (0, \bar{\varepsilon})\}$. Under a certain properness condition, Gowda and Tawhid [13] showed that the NCP with a P_0-function has a long central path [13, Theorem 9]. It should be pointed out that noninterior-point trajectories have also been extensively studied in the recent literature (see [1, 2, 3, 5, 10, 11, 13, 17, 35, 37]).

However, for a general complementarity problem, the system (1) may have multiple solutions for a given $\varepsilon > 0$, and even if the solution is unique, it is not necessarily continuous in ε. As a result, the existence of the central path is not always guaranteed. We define the (multivalued) mapping $\mathcal{U} : (0, \infty) \to \mathcal{S}(R^n_{++})$ by

$$\mathcal{U}(\varepsilon) = \{x \in R^n_{++} : f(x) > 0, Xf(x) = \varepsilon e\}, \tag{2}$$

where $X = \operatorname{diag}(x)$ and $\mathcal{S}(R^n_{++})$ is the set of all subsets of R^n_{++}, the positive orthant of R^n. The main contribution of this paper is to describe several sufficient conditions which ensure that the multivalued mapping $\mathcal{U}(\varepsilon)$ has the following desirable properties.

(a) $\mathcal{U}(\varepsilon) \neq \emptyset$ *for each* $\varepsilon \in (0, \infty)$.

(b) *For any fixed* $\bar{\varepsilon} > 0$, *the set* $\bigcup_{\varepsilon \in (0, \bar{\varepsilon}]} \mathcal{U}(\varepsilon)$ *is bounded.*

(c) *If* $\mathcal{U}(\varepsilon) \neq \emptyset$, *then* $\mathcal{U}(\varepsilon)$ *is upper-semicontinuous at* ε. *(That is, for any sufficiently small* $\delta > 0$, *we have that* $\emptyset \neq \mathcal{U}(\varepsilon') \subseteq \mathcal{U}(\varepsilon) + \delta B$ *for all* ε' *sufficiently close to* ε, *where* $B = \{x \in R^n : \|x\| < 1\}$ *is the Euclidean unit ball.)*

(d) *If* $\mathcal{U}(\cdot)$ *is single-valued, then* $\mathcal{U}(\varepsilon)$ *is continuous at* ε *provided that* $\mathcal{U}(\varepsilon) \neq \emptyset$.

If the mapping $\mathcal{U}(\cdot)$ satisfies properties (a), (b), and (c), then the set $\bigcup_{\varepsilon \in (0,\infty)} \mathcal{U}(\varepsilon)$ can be viewed as an "interior band" associated with the solution set of the NCP. The "interior band" can be viewed as a generalization of the concept of the central path. Indeed, if $\mathcal{U}(\cdot)$ satisfies properties (a), (b), and (d), then the set $\bigcup_{\varepsilon \in (0,\infty)} \mathcal{U}(\varepsilon)$ coincides with the central path of the NCP.

There exist several ways of generating the central path of the NCP, including maximal monotone methods [14, 30], minimization methods [31], homeomorphism techniques [6, 14, 15, 25, 33], the parameterized Sard theorem [42], and weakly univalent properties of continuous functions [13, 35, 37]. In this paper, we develop a different method for the analysis of the existence of the central path. By means of the homotopy invariance property of the degree and a newly introduced concept of interior-point-ε-exceptional family for continuous functions, we establish an alternative theorem for the nonemptyness of the mapping $\mathcal{U}(\varepsilon)$. For a given $\varepsilon > 0$, the result states that there exists either an interior-point-ε-exceptional family for f or $\mathcal{U}(\varepsilon) \neq \emptyset$. Consequently, to show the nonemptyness of the mapping $\mathcal{U}(\cdot)$, it is sufficient to verify conditions under which the function f possesses no interior-point-ε-exceptional family for any $\varepsilon > 0$. Along with this idea, we provide several sufficient conditions that guarantee the aforementioned desirable properties of the multivalued mapping $\mathcal{U}(\cdot)$.

These sufficient conditions are related to several classes of (nonmonotone) functions such as semimonotone, quasi-P$_*$-, P(τ, α, β)-, and exceptionally regular maps. The results proved in the paper include several known results on the central path as special instances.

This paper is organized as follows. In section 2, we introduce some definitions and some basic results that will be utilized in the paper. In section 3, we show an essential alternative theorem that is useful in later derivations. In section 4, we establish some sufficient conditions to guarantee the nonemptyness, boundedness, and upper-semicontinuity of the map $\mathcal{U}(\varepsilon)$, and the existence of the central path. Some concluding remarks are given in section 5.

Notations: R^n_+ (respectively, R^n_{++}) denotes the space of n-dimensional real vectors with nonnegative components (respectively, positive components), and $R^{n\times n}$ stands for the space of $n \times n$ matrices. For any $x \in R^n$, we denote by $\|x\|$ the Euclidean norm of x, by x_i the ith component of x for $i = 1, \ldots, n$, and by $[x]_+$ the vector whose ith component is $\max\{0, x_i\}$. When $x \in R^n_+(R^n_{++})$, we also write it as $x \geq 0$ $(x > 0)$ for simplicity.

2. Preliminaries. We first introduce the concept of an E$_0$-function, which is a generalization of an E$_0$-matrix, i.e., semimonotone matrix, (see [8]). Recall that an $n \times n$ matrix M is said to be an E$_0$-matrix if for any $0 \neq x \geq 0$, there exists a component $x_i > 0$ such that $(Mx)_i \geq 0$. M is a strictly semimonotone matrix if for any $0 \neq x \geq 0$, there exists a component $x_i > 0$ such that $(Mx)_i > 0$.

DEFINITION 2.1. *A function $f : R^n \to R^n$ is said to be an E_0-function (i.e., semimonotone function) if for any $x \neq y$ and $x \geq y$ in R^n, there exists some i such that $x_i > y_i$ and $f_i(x) \geq f_i(y)$. f is a strictly semimonotone function if for any $x \neq y$ and $x \geq y$ in R^n, there exists some i such that $x_i > y_i$ and $f_i(x) > f_i(y)$.*

It is evident that $f = Mx + q$, where $M \in R^{n\times n}$ and $q \in R^n$, is an E$_0$-function if and only if M is an E$_0$-matrix. We recall that a function f is said to be a P$_0$(P)-function if for any $x \neq y$ in R^n

$$\max_{x_i \neq y_i} (x_i - y_i)(f_i(x) - f_i(y)) \geq 0 (> 0).$$

Clearly, a P$_0$-function is an E$_0$-function. However, the converse is not true (see [8, Example 3.9.2]). Thus the class of E$_0$-functions is larger than that of P$_0$-functions.

DEFINITION 2.2. (D1) [23, 24]. *A map $f : R^n \to R^n$ is said to be quasi monotone if for $x \neq y$ in R^n, $f(y)^T(x - y) > 0$ implies that $f(x)^T(x - y) \geq 0$.*

(D2) [26, 44]. *A map $f : R^n \to R^n$ is said to be a P_*-map if there exists a scalar $\kappa \geq 0$ such that for any $x \neq y$ in R^n we have*

$$(1 + \kappa) \sum_{i \in I_+(x,y)} (x_i - y_i)(f_i(x) - f_i(y)) + \sum_{i \in I_-(x,y)} (x_i - y_i)(f_i(x) - f_i(y)) \geq 0,$$

where

$$I_+(x, y) = \{i : (x_i - y_i)(f_i(x) - f_i(y)) > 0\}, \tag{3}$$

$$I_-(x, y) = \{i : (x_i - y_i)(f_i(x) - f_i(y)) < 0\}.$$

(D3) [26]. *M is said to be a P_*-matrix if there exists a scalar $\kappa \geq 0$ such that*

$$(1 + \kappa) \sum_{i \in I_+} x_i(Mx)_i + \sum_{i \in I_-} x_i(Mx)_i \geq 0,$$

where $I_+ = \{i : x_i(Mx)_i > 0\}$ *and* $I_- = \{i : x_i(Mx)_i < 0\}$.

Clearly, for a linear map $f(x) = Mx + q$, f is a P_*-map if and only if M is a P$_*$-matrix. Väliaho [40] showed that the class of P$_*$-matrices coincides with the class of sufficient matrices [8, 9]. A new equivalent definition of the P$_*$-matrix is given in [46]. The next concept is a generalization of the quasi monotone function and the P$_*$-map.

DEFINITION 2.3. [46] *A function* $f : R^n \to R^n$ *is said to be a quasi-*P_**-map if there exists a constant* $\tau \geq 0$ *such that the following implication holds for all* $x \neq y$ *in* R^n.

$$f(y)^T(x-y) - \tau \sum_{i \in I_+(x,y)} (x_i - y_i)(f_i(x) - f_i(y)) > 0 \Rightarrow f(x)^T(x-y) \geq 0,$$

where $I_+(x,y)$ *is defined by* (3).

From the above definition, it is evident that the class of quasi-P$_*$-maps includes quasi monotone functions and P$_*$-maps. (see [46] for details). The following concept of a $P(\tau, \alpha, \beta)$-map is also a generalization of the P$_*$-map. In [46], it is pointed out that monotone functions and P$_*$-maps are special cases of P(τ, α, β)-maps.

DEFINITION 2.4. [46] *A mapping* $f : R^n \to R^n$ *is said to be a* $P(\tau, \alpha, \beta)$*-map if there exist constants* $\tau \geq 0, \alpha \geq 0,$ *and* $0 \leq \beta < 1$ *such that the following inequality holds for all* $x \neq y$ *in* R^n :

$$(1+\tau) \max_{1 \leq i \leq n} (x_i - y_i)(f_i(x) - f_i(y)) + \min_{1 \leq i \leq n} (x_i - y_i)(f_i(x) - f_i(y)) + \alpha \|x - y\|^\beta \geq 0.$$

The concept of exceptional regularity that we are going to define next has a close relation to such concepts as copositive, R_0-, P$_0$-, and E$_0$-functions. It is shown that the exceptional regularity is a weak sufficient condition for the nonemptyness and the boundedness of the mapping $\mathcal{U}(\varepsilon)$ (see section 4.4 for details).

DEFINITION 2.5. *Let* f *be a function from* R^n *into* R^n. f *is said to be exceptionally regular if, for each* $\beta \geq 0$, *the following complementarity problem has no solution of norm* 1:

$$G(x) + \beta x \geq 0, \ x \geq 0, \ x^T(G(x) + \beta x) = 0,$$

where $G(x) = f(x) - f(0)$.

The following two results are employed to prove the main result of the next section. Let S be an open bounded set of R^n. We denote by $\overline{S}$ and $\partial(S)$ the closure and boundary of S, respectively. Let F be a continuous function from $\overline{S}$ into R^n. For any $y \in R^n$ such that $y \notin F(\partial(S))$, the symbol $\deg(F, S, y)$ denotes the topological degree associated with F, S, and y (see [34]).

LEMMA 2.1. [34] *Let* $S \subset R^n$ *be an open bounded set and* F, G *be two continuous functions from* $\overline{S}$ *into* R^n.

(i) *Let the homotopy* $H(x,t)$ *be defined as*

$$H(x,t) = tG(x) + (1-t)F(x), \ 0 \leq t \leq 1,$$

and let y *be an arbitrary point in* R^n. *If* $y \notin \{H(x,t) : x \in \partial S$ *and* $t \in [0,1]\}$, *then* $deg(G, S, y) = deg(F, S, y)$.

(ii) *If* $deg(F, S, y) \neq 0$, *then the equation* $F(x) = y$ *has a solution in* S.

The following upper-semicontinuity theorem of weakly univalent maps is due to Ravindran and Gowda [35].

LEMMA 2.2. [35] *Let $g : R^n \to R^n$ be weakly univalent; that is, g is continuous and there exist one-to-one continuous functions $g_k : R^n \to R^n$ such that $g_k \to g$ uniformly on every bounded subset of R^n. Suppose that $q^* \in R^n$ such that $g^{-1}(q^*)$ is nonempty and compact. Then for any given scalar $\delta > 0$ there exists a scalar $\gamma > 0$ such that for any weakly univalent function $h : R^n \to R^n$ and for any $q \in R^n$ with*

$$\sup_{\bar{\Omega}} \|h(x) - g(x)\| < \gamma, \ \|q - q^*\| < \gamma,$$

we have

$$\emptyset \neq h^{-1}(q) \subseteq g^{-1}(q^*) + \delta B,$$

where B denotes the open unit ball in R^n and $\Omega = g^{-1}(q^*) + \delta B$.

3. Interior-point-ε-exceptional family and an alternative theorem. We now introduce the concept of the interior-point-ε-exceptional family for a continuous function, which brings us to a new idea, to investigate the properties of the mapping $\mathcal{U}(\varepsilon)$ defined by (2), especially the existence of the central path for NCPs. This concept can be viewed as a variant of the exceptional family of elements which was originally introduced to study the solvability of complementarity problems and variational inequalities [19, 20, 36, 43, 44, 45, 46].

DEFINITION 3.1. *Let $f : R^n \to R^n$ be a continuous function. Given a scalar $\varepsilon > 0$, we say that a sequence $\{x^r\}_{r>0} \subset R^n_{++}$ is an interior-point-ε-exceptional family for f if $\|x^r\| \to \infty$ as $r \to \infty$ and for each x^r there exists a positive number $0 < \mu_r < 1$ such that*

$$f_i(x^r) = \frac{1}{2}\left(\mu_r - \frac{1}{\mu_r}\right) x_i^r + \frac{\varepsilon \mu_r}{x_i^r} \quad \text{for all } i = 1, \ldots, n. \tag{4}$$

Based on the above concept, we can prove the following result which plays a key role in the analysis of the paper.

THEOREM 3.1. *Let f be a continuous function from R^n into R^n. Then for each $\varepsilon > 0$ there exists either a point $x(\varepsilon)$ such that*

$$x(\varepsilon) > 0, \ f(x(\varepsilon)) > 0, \ x_i(\varepsilon) f_i(x(\varepsilon)) = \varepsilon, \ i = 1, \ldots, n \tag{5}$$

or an interior-point-ε-exceptional family for f.

Proof. Let $F(x) = (F_1(x), \ldots, F_n(x))^T$ be the Fischer–Burmeister function of f defined by

$$F_i(x) = x_i + f_i(x) - \sqrt{x_i^2 + f_i^2(x)}, \ i = 1, \ldots, n.$$

It is well known that x solves the NCP if and only if x solves the equation $F(x) = 0$. Given $\varepsilon > 0$, we perturb $F(x)$ to $F_\varepsilon(x)$ given by

$$[F_\varepsilon(x)]_i = x_i + f_i(x) - \sqrt{x_i^2 + f_i^2(x) + 2\varepsilon}, \ i = 1, \ldots, n. \tag{6}$$

It is easy to see that $x(\varepsilon)$ solves the equation $F_\varepsilon(x) = 0$ if and only if $x(\varepsilon)$ satisfies the system (5). We now consider the convex homotopy between the mapping $F_\varepsilon(x)$ and the identity mapping, that is,

$$H(x, t) = tx + (1 - t)F_\varepsilon(x), \ 0 \leq t \leq 1.$$

Let $r > 0$ be an arbitrary positive scalar. Consider the open bounded set $S_r = \{x \in R^n : \|x\| < r\}$. The boundary of S_r is given by $\partial S_r = \{x \in R^n : \|x\| = r\}$. There are only two cases.

Case 1. There exists a number $r > 0$ such that $0 \notin \{H(x,t) : x \in \partial S_r \text{ and } t \in [0,1]\}$. In this case, by (i) of Lemma 2.1, we have that $\deg(I, S_r, 0) = \deg(F_\varepsilon(x), S_r, 0)$, where I is the identity mapping. Since $\deg(I, S_r, 0) = 1$, from the above equation and (ii) of Lemma 2.1, we deduce that the equation $F_\varepsilon(x) = 0$ has a solution, denoted by $x(\varepsilon)$, which satisfies the system (5).

Case 2. For each $r > 0$, there exists some point $x^r \in \partial S_r$ and $t_r \in [0,1]$ such that

$$0 = H(x^r, t_r) = t_r x^r + (1 - t_r)F_\varepsilon(x^r). \tag{7}$$

If $t_r = 0$ for some $r > 0$, then the above equation reduces to $F_\varepsilon(x^r) = 0$, which implies that $x(\varepsilon) := x^r$ satisfies the system (5).

We now verify that $t_r \neq 1$. In fact, if $t_r = 1$ for some $r > 0$, then from (7) we have that $x^r = 0$, which is impossible since $x^r \in \partial S_r$.

Therefore, it is sufficient to consider the case of $0 < t_r < 1$ for all $r > 0$. In this case, it is easy to show that f actually has an interior-point-ε-exceptional family. Indeed, in this case, (7) can be written as

$$x_i^r + (1 - t_r)f_i(x^r) = (1 - t_r)\sqrt{(x_i^r)^2 + f_i^2(x^r) + 2\varepsilon},\ i = 1, \dots, n. \tag{8}$$

Squaring both sides of the above and simplifying, we have

$$x_i^r f_i(x^r) = \frac{1}{2}\left[(1 - t_r) - \frac{1}{1 - t_r}\right](x_i^r)^2 + (1 - t_r)\varepsilon,\ i = 1, \dots, n.$$

Since $t_r \in (0,1)$, the above equation implies that $x_i^r \neq 0$ for all $i = 1, \dots, n$. Denote $\mu_r = 1 - t_r$. We see from the above equation that

$$f_i(x^r) = \frac{1}{2}\left(\mu_r - \frac{1}{\mu_r}\right)x_i^r + \frac{\mu_r \varepsilon}{x_i^r},\ i = 1, \dots, n. \tag{9}$$

We further show that $x^r \in R^n_{++}$. In fact, it follows from (8) that

$$x_i^r + \mu_r f_i(x^r) > \mu_r\sqrt{2\varepsilon} > 0,\ i = 1, \dots, n. \tag{10}$$

On the other hand, by using (9) we obtain

$$x_i^r + \mu_r f_i(x^r) = \frac{1}{2}(\mu_r^2 + 1)x_i^r + \frac{\mu_r^2 \varepsilon}{x_i^r},\ i = 1, \dots, n.$$

Combining (10) and the above equation yields $x^r \in R^n_{++}$. Since $\|x^r\| = r$, it is clear that $\|x^r\| \to \infty$ as $r \to \infty$. Consequently, the sequence $\{x^r\}$ is an interior-point-ε-exceptional family for f. □

The above result shows that if f has no interior-point-ε-exceptional family for each $\varepsilon > 0$, then property (a) of the mapping $\mathcal{U}(\cdot)$ holds. From the result, it is interesting to study various practical conditions under which a continuous function does not possess an interior-point-ε-exceptional family for every $\varepsilon \in (0, \infty)$. In the next section, we provide several such conditions ensuring the aforementioned desirable properties of the mapping $\mathcal{U}(\cdot)$.

4. Sufficient conditions for properties of $\mathcal{U}(\cdot)$.

4.1. E$_0$-function. In this section, we prove that the multivalued mapping $\mathcal{U}(\cdot)$ has properties (a) and (b) if f is a continuous E$_0$-function satisfying a certain properness condition. Moreover, if $F_\varepsilon(x)$ given by (6) is weakly univalent, then property (c) also holds. Applied to P$_0$ complementarity problems, this existence result extends a recent result due to Gowda and Tawhid [13]. The following lemma is quite useful.

Lemma 4.1. *Let $f : R^n \to R^n$ be an E_0-function. Then for any sequence $\{u^k\} \subset R^n_{++}$ with $\|u^k\| \to \infty$, there exist an index i and a subsequence of $\{u^k\}$, denoted by $\{u^{k_j}\}$, such that $u_i^{k_j} \to \infty$ and $f_i(u^{k_j})$ is bounded below.*

Proof. This proof has appeared in several works, see [11, 13, 35, 38]. Let $\{u^k\} \subset R^n_{++}$ be a sequence satisfying $\|u^k\| \to \infty$. Choosing a subsequence if necessary, we may suppose that there exists an index set $I \subseteq \{1, \dots, n\}$ such that $u_i^k \to \infty$ for each $i \in I$, and $\{u_i^k\}$ is bounded for each $i \notin I$. Let $v^k \in R^n$ be a vector constructed as follows:

$$v_i^k = u_i^k \text{ for } i \notin I, \quad v_i^k = 0 \text{ for } i \in I.$$

Thus, $\{v^k\}$ is a bounded sequence. Clearly, $u^k \geq v^k$. Since f is an E$_0$-function, there exist an index $i \in I$ and a subsequence of $\{u^k\}$, denoted by $\{u^{k_j}\}$, such that $u_i^{k_j} > v_i^{k_j}$ and $f_i(u^{k_j}) \geq f_i(v^{k_j})$ for all j. Thus,

$$f_i(u^{k_j}) \geq \inf_j f_i(v^{k_j}).$$

Note that the right-hand side of the above inequality is bounded. The desired result follows. □

To show the main result of this subsection, we will make use of the following assumption which is weaker than several previously known conditions.

Condition 4.1. *For any sequence $\{x^k\}$ satisfying*

(i) *$\{x^k\} \subset R^n_{++}$, $\|x^k\| \to \infty$ and $[-f(x^k)]_+/\|x^k\| \to 0$, and*

(ii) *for each index i with $x_i^k \to \infty$, the corresponding sequence $\{f_i(x^k)\}$ is bounded above, and*

(iii) *there exists at least one index i_0 such that $x_{i_0}^k \to \infty$ and $\{f_{i_0}(x^k)\}$ is bounded, it holds that*

$$\max_{1 \leq i \leq n} x_i^{k_l} f_i(x^{k_l}) \to \infty$$

for some subsequence $\{x^{k_l}\}$.

As we see in the following result the above condition encompasses several particular cases; we omit the details.

Proposition 4.1. *Condition* 4.1 *is satisfied if one of the following conditions holds.*

(C1) *For any positive sequence $\{x^k\} \subset R^n_{++}$ with $\|x^k\| \to \infty$ and $[-f(x^k)]_+/\|x^k\| \to 0$, it holds that $\max_{1\leq i \leq n} x_i^{k_l} f_i(x^{k_l}) \to \infty$ for some subsequence $\{x^{k_l}\}$.*

(C2) *For any sequence $\{x^k\} \subset R^n_{++}$ with $\|x^k\| \to \infty$ and $\min_{1\leq i\leq n} f_i(x^k)/\|x^k\| \to 0$, it holds that $\max_{1\leq i\leq n} x_i^{k_l} f_i(x^{k_l}) \to \infty$ for some subsequence $\{x^{k_l}\}$.*

(C3) [22, 29] *For any sequence $\{x^k\}$ with $\|x^k\| \to \infty$, $[-x^k]_+/\|x^k\| \to 0$, and $[-f(x^k)]_+/\|x^k\| \to 0$, it holds that*

$$\liminf_{k\to\infty} (x^k)^T f(x^k)/\|x^k\| > 0.$$

(C4) [13] *For any sequence* $\{x^k\}$ *with* $\|x^k\| \to \infty$,

$$\liminf_{k\to\infty} \frac{\min_{1\le i\le n} x_i^k}{\|x^k\|} \ge 0, \quad \textit{and} \quad \liminf_{k\to\infty} \frac{\min_{1\le i\le n} f_i(x^k)}{\|x^k\|} \ge 0,$$

there exist an index j *and a subsequence* $\{x^{k_l}\}$ *such that* $x_j^{k_l} f_j(x^{k_l}) \to \infty$.

(C5) [6, 39] f *is a* R_0*-function.*

(C6) [14, 25, 30, 31] f *is monotone and the NCP is strictly feasible.*

(C7) [27] f *is a uniform P-function.*

Remark 4.1. The condition (C1) of the above proposition is weaker than each of the conditions (C2) through (C7). (C2) is weaker than each of the conditions (C4) through (C7). The concept of the R$_0$-function, a generalization of the R_0-matrix [8], was introduced in [39] and later modified in [6].

In what follows, we show under a properness condition that the short "interior band" $\bigcup_{\varepsilon\in(0,\bar\varepsilon]} \mathcal{U}(\varepsilon)$ is bounded for each given $\bar\varepsilon > 0$. The boundedness is important because it implies that the sequence $\{x(\varepsilon_k)\}$, where $x(\varepsilon_k) \in \mathcal{U}(\varepsilon_k)$ and $\varepsilon_k \to 0$, is bounded and each accumulation point of the sequence is a solution to the NCP provided that f is continuous. We impose the following condition on f.

Condition 4.2. *For any positive sequence* $\{x^k\} \subset R^n_{++}$ *such that* $\|x^k\| \to \infty$, $\lim_{k\to\infty}[-f(x^k)]_+ = 0$, *and the sequence* $\{f_i(x^k)\}$ *is bounded for each index* i *with* $x_i^k \to \infty$, *it holds that*

$$\max_{1\le i\le n} x_i^{k_l} f_i(x^{k_l}) \to \infty$$

for some subsequence $\{x^{k_l}\}$.

Clearly, Condition 4.2 is weaker than Condition 4.1 and thereby weaker than all conditions listed in Proposition 4.1. We now prove the boundedness of the short "interior band" under the above condition.

Lemma 4.2. *Suppose that Condition* 4.2 *is satisfied. If* $\mathcal{U}(\varepsilon) \ne \emptyset$ *for each* $\varepsilon > 0$, *then for any* $\bar\varepsilon > 0$ *the set* $\bigcup_{\varepsilon\in(0,\bar\varepsilon]} \mathcal{U}(\varepsilon)$ *is bounded, i.e., property* (b) *holds. Particularly,* $\mathcal{U}(\varepsilon)$ *is bounded for each* $\varepsilon > 0$.

Proof. Suppose that there exists some $\bar\varepsilon > 0$ such that $\bigcup_{\varepsilon\in(0,\bar\varepsilon]} \mathcal{U}(\varepsilon)$ is unbounded. Then there exists a sequence $\{x(\varepsilon_k)\}$, where $\varepsilon_k \in (0, \bar\varepsilon]$, such that $\|x(\varepsilon_k)\| \to \infty$ as $k \to \infty$. Since $x(\varepsilon_k) \in \mathcal{U}(\varepsilon_k)$, we deduce that $[-f(x(\varepsilon_k))]_+ = 0$ for all k, and that

$$0 < f_i(x(\varepsilon_k)) = \frac{\varepsilon_k}{x_i(\varepsilon_k)} < \frac{\bar\varepsilon}{x_i(\varepsilon_k)} \quad \text{for all } i = 1, \dots, n.$$

Thus, for each i such that $x_i(\varepsilon_k) \to \infty$, the sequence $\{f_i(x(\varepsilon_k))\}$ is bounded. By Condition 4.2, we deduce that there exists a subsequence $\{x(\varepsilon_{k_l})\}$ such that

$$\max_{1\le i\le n} x_i(\varepsilon_{k_l}) f_i(x(\varepsilon_{k_l})) \to \infty.$$

This is a contradiction since $x_i(\varepsilon_{k_l}) f_i(x(\varepsilon_{k_l})) = \varepsilon_{k_l} < \bar\varepsilon$ for all $i = 1, \dots, n$. □

The main result on E$_0$-functions is given as follows. Even for P$_0$-functions, this result is new.

Theorem 4.1. *Suppose that* f *is a continuous* E_0*-function and Condition* 4.1 *is satisfied. Then the properties* (a) *and* (b) *of the mapping* $\mathcal{U}(\varepsilon)$ *hold. Moreover, if* $F_\varepsilon(x)$ *defined by* (6) *is weakly univalent in* x, *then the mapping* $\mathcal{U}(\cdot)$ *is upper-semicontinuous, i.e., property* (c) *also holds.*

Proof. To prove property (a), by Theorem 3.1, it suffices to show that there exists no interior-point-ε-exceptional family of f for any $\varepsilon > 0$. Assume to the contrary that for certain $\varepsilon > 0$ the function f has an interior-point-ε-exceptional family $\{x^r\}$. Since $\|x^r\| \to \infty$, $\{x^r\} \subset R^n_{++}$, and f is an E$_0$-function, by Lemma 4.1 there exist some index m and a subsequence $\{x^{r_j}\}$, such that $x^{r_j}_m \to \infty$ and $f_m(x^{r_j})$ is bounded below. From (4), we have

$$0 > \frac{1}{2}\left(\mu_{r_j} - \frac{1}{\mu_{r_j}}\right) x^{r_j}_m = f_m(x^{r_j}) - \frac{\mu_{r_j}\varepsilon}{x^{r_j}_m}.$$

Since $x^{r_j}_m \to \infty$ and $f_m(x^{r_j})$ is bounded below, the right-hand side of the above equation is bounded below. It follows that $\lim_{j\to\infty} \mu_{r_j} = 1$.

On the other hand, we note that for any $0 < \mu < 1$ the function

$$\phi(t) = \frac{1}{2}\left(\mu - \frac{1}{\mu}\right) t + \frac{\mu\varepsilon}{t} \tag{11}$$

is monotonically decreasing with respect to the variable $t \in (0, \infty)$. Passing through a subsequence, we may suppose that there exists an index set $I \subseteq \{1, \dots, n\}$ such that $x^{r_j}_i \to \infty$ for each $i \in I$, and $\{x^{r_j}_i\}$ is bounded for each $i \notin I$.

If $i \notin I$, then there exists some scalar $C > 0$ such that $x^{r_j}_i \le C$ for all j. Since $\phi(t)$ is decreasing and $\mu_{r_j} \to 1$, we have

$$f_i(x^{r_j}) = \frac{1}{2}\left(\mu_{r_j} - \frac{1}{\mu_{r_j}}\right) x^{r_j}_i + \frac{\mu_{r_j}\varepsilon}{x^{r_j}_i} \ge \frac{1}{2}\left(\mu_{r_j} - \frac{1}{\mu_{r_j}}\right) C + \frac{\mu_{r_j}\varepsilon}{C} \to \frac{\varepsilon}{C} > 0.$$

Thus, for all sufficiently large j, we have

$$[-f_i(x^{r_j})]_+ = 0 \ \text{ for all } i \notin I.$$

If $i \in I$, by using (4) and the facts $\mu_{r_j} \to 1$ and $x^{r_j}_i \to \infty$, we have

$$\frac{f_i(x^{r_j})}{\|x^{r_j}\|} = \frac{1}{2}\left(\mu_{r_j} - \frac{1}{\mu_{r_j}}\right) \frac{x^{r_j}_i}{\|x^{r_j}\|} + \frac{\mu_{r_j}\varepsilon}{x^{r_j}_i \|x^{r_j}\|} \to 0,$$

which implies that

$$[-f_i(x^{r_j})]_+/\|x^{r_j}\| \to 0 \ \text{ for all } i \in I.$$

Therefore, $[-f(x^{r_j})]_+/\|x^{r_j}\| \to 0$. Moreover, it follows from (4) that

$$f_i(x^{r_j}) \le \frac{\mu_{r_j}\varepsilon}{x^{r_j}_i} \le \frac{\varepsilon}{x^{r_j}_i} \to 0 \text{ for all } i \in I,$$

which implies that $\{f_i(x^{r_j})\}$ is bounded above for all $i \in I$. Since $m \in I$ and $\{f_m(x^{r_j})\}$ is bounded below, the sequence $\{f_m(x^{r_j})\}$ is indeed bounded. From Condition 4.1, there is a subsequence of $\{x^{r_j}\}$, denoted also by $\{x^{r_j}\}$, such that

$$\max_{1\le i\le n} x^{r_j}_i f_i(x^{r_j}) \to \infty.$$

However, from (4) we have

$$x^{r_j}_i f_i(x^{r_j}) = \frac{1}{2}\left(\mu_{r_j} - \frac{1}{\mu_{r_j}}\right) (x^{r_j}_i)^2 + \mu_{r_j}\varepsilon \le \mu_{r_j}\varepsilon < \varepsilon \tag{12}$$

for all $i \in \{1, \dots, n\}$. This is a contradiction. Property (a) of $\mathcal{U}(\varepsilon)$ follows.

Since Condition 4.1 implies Condition 4.2, the boundedness of the set $\bigcup_{\varepsilon \in (0,\bar{\varepsilon}]} \mathcal{U}(\varepsilon)$ follows immediately from Lemma 4.2. It is known that $x(\varepsilon) \in \mathcal{U}(\varepsilon)$ if and only if $x(\varepsilon)$ is a solution to the equation $F_\varepsilon(x) = 0$, i.e., $\mathcal{U}(\varepsilon) = F_\varepsilon^{-1}(0)$. Since $\mathcal{U}(\varepsilon)$ is bounded, the set $F_\varepsilon^{-1}(0)$ is bounded (in fact, compact, since f is continuous). If $F_\varepsilon(x)$ is weakly univalent in x, by Lemma 2.2, for each scalar $\delta > 0$ there is a $\gamma > 0$ such that for any weakly univalent function $h : R^n \to R^n$ with

$$\sup_{x \in \bar{\Omega}} \|h(x) - F_\varepsilon(x)\| < \gamma, \text{ where } \Omega = F_\varepsilon^{-1}(0) + \delta B, \tag{13}$$

we have

$$\emptyset \neq h^{-1}(0) \subseteq F_\varepsilon^{-1}(0) + \delta B. \tag{14}$$

It is easy to see that for the given $\gamma > 0$ there exists a scalar $\beta > 0$ such that

$$\sup_{x \in \bar{\Omega}} \|F_{\varepsilon'}(x) - F_\varepsilon(x)\| < \gamma \text{ for all } |\varepsilon' - \varepsilon| < \beta.$$

Setting $h(x) := F_{\varepsilon'}(x)$ in (13) and (14), we obtain that $\emptyset \neq F_{\varepsilon'}^{-1}(0) \subseteq F_\varepsilon^{-1}(0) + \delta B$ for all $|\varepsilon' - \varepsilon| < \beta$, i.e., $\mathcal{U}(\varepsilon') \subseteq \mathcal{U}(\varepsilon) + \delta B$ for all ε' sufficiently close to ε. Thus, $\mathcal{U}(\varepsilon)$ is upper-semicontinuous. □

Ravindran and Gowda [35] showed that if f is a P_0-function, then $F_\varepsilon(x)$ given by (6) is a P-function in x, and hence the equation $F_\varepsilon(x) = 0$ has at most one solution $x(\varepsilon)$. In this case, the upper-semicontinuity of $\mathcal{U}(\cdot)$ reduces to the continuity of $x(\varepsilon)$. By the fact that every P_0-function is an E_0-function and is weakly univalent, we have the following result from Theorem 4.1.

COROLLARY 4.1. *Suppose that $f : R^n \to R^n$ is a continuous P_0-function and Condition* 4.1 *is satisfied. Then the central path exists and any slice of it is bounded, i.e., for each $\varepsilon > 0$ there exists a unique $x(\varepsilon)$ satisfying the system* (1)*, $x(\varepsilon)$ is continuous on* $(0, \infty)$*, and the set* $\{x(\varepsilon) : \varepsilon \in (0, \bar{\varepsilon}]\}$ *is bounded for each $\bar{\varepsilon} > 0$.*

When f is a P_0-function, Gowda and Tawhid [13, Theorem 9] showed that the (long) central path exists if condition (C4) of Proposition 4.1 is satisfied. Corollary 4.1 can serve as a generalization of the Gowda and Tawhid result. It is worth noting that the consequences of Corollary 4.1 remain valid if condition (C1) or (C2) of Proposition 4.1 holds.

4.2. Quasi-P_*-maps. The concept of the quasi-P_*-map that is a generalization of the quasi monotone function and the P_*-map was first introduced in [46] to study the solvability of the NCP. Under the strictly feasible assumption as well as the following condition, we can show the nonemptyness and the boundedness of $\mathcal{U}(\cdot)$ if f is a continuous quasi-P_*-map .

CONDITION 4.3. *For any sequence $\{x^k\} \subset R^n_{++}$ such that*

$$\|x^k\| \to \infty, \lim_{k \to \infty} [-f(x^k)]_+ = 0,$$

and $\{f(x^k)\}$ is bounded, it holds that

$$\max_{1 \le i \le n} x_i^{k_l} f_i(x^{k_l}) \to \infty$$

for some subsequence $\{x^{k_l}\}$.

Clearly, the above condition is weaker than Conditions 4.1 and 4.2. It is also weaker than Condition 3.8 in [4] and Condition 1.5(iii) in [25]. The following is the main result of this subsection.

THEOREM 4.2. *Let f be a continuous quasi-P_*-map with the constant $\tau \geq 0$ (see Definition* 2.3*). Suppose that Condition* 4.3 *is satisfied. If the NCP is strictly feasible, then property* (a) *of $\mathcal{U}(\varepsilon)$ holds. Moreover, if Condition* 4.2 *is satisfied, then property* (b) *holds, and if $F_\varepsilon(x)$ is weakly univalent in x, then property* (c) *also holds.*

While the nonemptyness of $\mathcal{U}(\varepsilon)$ is ensured under Condition 4.3, it is not clear if the boundedness of $\mathcal{U}(\varepsilon)$ can follow from this condition. However, from the implications Condition 4.1$\Rightarrow$ Condition 4.2 $\Rightarrow$ Condition 4.3, we have the next consequence.

COROLLARY 4.2. *Suppose that f is a continuous quasi-P_*-map and $F_\varepsilon(x)$ is weakly univalent in x. If the NCP is strictly feasible and Condition* 4.1 *or* 4.2 *is satisfied, then the mapping $\mathcal{U}(\cdot)$ has properties* (a), (b), *and* (c).

The proof of Theorem 4.2 is postponed until we have proved two technical lemmas.

LEMMA 4.3. *Let f satisfy Condition* 4.3*. Assume that $\{x^r\}_{r>0}$ is an interior-point-ε-exceptional family for f. If there exists a subsequence of $\{x^r\}$, denoted by $\{x^{r_k}\}$, such that for some $0 < \gamma < 1$,*

$$\lim_{k\to\infty}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)\|x^{r_k}\|^{1+\gamma} = 0, \tag{15}$$

then we have

$$\lim_{k\to\infty}\left(\min_{1\leq i\leq n} x_i^{r_k}\right) = 0.$$

Proof. Suppose that $\{x^{r_k}\}$ is an arbitrary subsequence of $\{x^r\}$ such that (15) holds. Since $\phi(t)$ defined by (11) is decreasing on $(0,\infty)$, for each $i \in \{1,\dots,n\}$ we have

$$f_i(x^{r_k}) \leq \frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)\min_{1\leq i\leq n} x_i^{r_k} + \frac{\mu_{r_k}\varepsilon}{\min_{1\leq i\leq n} x_i^{r_k}} \tag{16}$$

and

$$f_i(x^{r_k}) \geq \frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)\max_{1\leq i\leq n} x_i^{r_k} + \frac{\mu_{r_k}\varepsilon}{\max_{1\leq i\leq n} x_i^{r_k}}. \tag{17}$$

Suppose to the contrary that there exists a subsequence of $\{x^{r_k}\}$, denoted also by $\{x^{r_k}\}$, such that $\min_{1\leq i\leq n} x_i^{r_k} \geq \alpha > 0$ for all $k > 0$, where α is a constant. We derive a contradiction. Indeed, since $\mu_{r_k} - \frac{1}{\mu_{r_k}} < 0$, from (16) we have

$$f_i(x^{r_k}) \leq \frac{\mu_{r_k}\varepsilon}{\min_{1\leq i\leq n} x_i^{r_k}} \leq \frac{\varepsilon}{\alpha} \quad \text{for all } i = 1,\dots,n.$$

From (17) and the above relation, we obtain

$$\frac{\varepsilon}{\alpha} \geq f_i(x^{r_k}) \geq \frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)\max_{1\leq i\leq n} x_i^{r_k} \quad \text{for all } i = 1,\dots,n. \tag{18}$$

Since $\|x^{r_k}\| \to \infty$, we deduce from (15) that

$$\lim_{k\to\infty}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)\max_{1\leq i\leq n} x_i^{r_k} = 0.$$

Therefore, it follows from (18) that there exists a scalar c such that $c \le f_i(x^{r_k}) \le \varepsilon/\alpha$ for all $i = 1, \ldots, n$ and $\lim_{k\to\infty}[-f_i(x^{r_k})]_+ = 0$. By Condition 4.3, there exists a subsequence of $\{x^{r_k}\}$, denoted still by $\{x^{r_k}\}$, such that $\max_{1\le i\le n} x_i^{r_k} f_i(x^{r_k}) \to \infty$. However, from (12) we have that $x_i^{r_k} f(x_i^{r_k}) \le \mu_{r_k}\varepsilon < \varepsilon$ for all $i = 1, \ldots, n$. This is a contradiction. $\square$

LEMMA 4.4. *Let f satisfy Condition* 4.3. *Assume that $\{x^r\}$ is an interior-point-ε-exceptional family for f. Let $u > 0$ be an arbitrary vector in R^n. Then for any subsequence $\{x^{r_k}\}$ (where $r_k \to \infty$ as $k \to \infty$) there exists a subsequence of $\{x^{r_k}\}$, denoted still by $\{x^{r_k}\}$, such that $f(x^{r_k})^T(x^{r_k} - u) < 0$ for all sufficiently large k.*

Proof. Let $\{x^{r_k}\}$ be an arbitrary subsequence of $\{x^r\}$ (where $r_k \to \infty$ as $k \to \infty$). By using (4) we have

$$\begin{aligned} &f(x^{r_k})^T(x^{r_k} - u) \\ &= \frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)\|x^{r_k}\|^2 + n\varepsilon\mu_{r_k} - \frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)(x^{r_k})^T u - \sum_{i=1}^n \frac{\mu_{r_k}\varepsilon u_i}{x_i^{r_k}} \\ &= \frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)(\|x^{r_k}\|^2 - (x^{r_k})^T u) + \mu_{r_k}\varepsilon\left(n - \sum_{i=1}^n \frac{u_i}{x_i^{r_k}}\right). \end{aligned} \tag{19}$$

We suppose that $f(x^{r_k})^T(x^{r_k} - u) \ge 0$ for all sufficiently large k. We derive a contradiction. From (19), we have

$$0 \le f(x^{r_k})^T(x^{r_k} - u) \le \frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)(\|x^{r_k}\|^2 - (x^{r_k})^T u) + \mu_{r_k}\varepsilon n.$$

Since $\|x^{r_k}\| \to \infty$, for all sufficiently large k we have

$$0 \le \frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)(\|x^{r_k}\|^2 - (x^{r_k})^T u) + \mu_{r_k}\varepsilon n \le \mu_{r_k}\varepsilon n,$$

which implies that

$$\begin{aligned} &\lim_{k\to\infty}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)\|x^{r_k}\|^{1+\gamma} \\ &\quad = \lim_{k\to\infty}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)(\|x^{r_k}\|^2 - (x^{r_k})^T u)\frac{\|x^{r_k}\|^{1+\gamma}}{\|x^{r_k}\|^2 - (x^{r_k})^T u} \\ &\quad = 0 \end{aligned}$$

for any scalar $0 < \gamma < 1$. Thus, we see from Lemma 4.3 that

$$\min_{1\le i\le n} x_i^{r_k} \to 0. \tag{20}$$

Notice that

$$\frac{1}{2}\left(\mu_{r_k} - \frac{1}{\mu_{r_k}}\right)(\|x^{r_k}\|^2 - (x^{r_k})^T u) < 0$$

for all sufficiently large k. From (19), (20), and the above inequality, we have

$$f(x^{r_k})^T(x^{r_k} - u) \le \mu_{r_k}\varepsilon\left(n - \sum_{i=1}^n \frac{u_i}{x_i^{r_k}}\right) \le \mu_{r_k}\varepsilon\left(n - \frac{\min_{1\le i\le n} u_i}{\min_{1\le i\le n} x_i^{r_k}}\right) < 0$$

for all sufficiently large k. This is a contradiction. $\square$

We are now ready to prove the results of Theorem 4.2.

Proof of Theorem 4.2. To show property (a) of the mapping $\mathcal{U}(\varepsilon)$, by Theorem 3.1, it suffices to show that f has no interior-point-ε-exceptional family for any $\varepsilon > 0$. Assume to the contrary that there exists an interior-point-ε-exceptional family for f, denoted by $\{x^r\}$. By the strict feasibility of the NCP, there is a vector $u > 0$ such that $f(u) > 0$. Consider two possible cases.

Case (A). There exists a number $r_0 > 0$ such that

$$\max_{1\le i\le n}(x_i^r - u_i)(f_i(x^r) - f_i(u)) < 0 \quad \text{for all } r \ge r_0.$$

In this case, the index set $I_+(x^r, u)$ is empty. Since $f(u) > 0,\ x^r > 0$, and $\|x^r\| \to \infty$, it is easy to see that

$$f(u)^T(x^r - u) > 0$$

for all sufficiently large r. Since f is a quasi-P$_*$-map and $I_+(x^r, u)$ is empty, the above inequality implies that $f(x^r)^T(x^r - u) \ge 0$ for all sufficiently large r. However, by Lemma 4.4 there exists a subsequence of $\{x^r\}$, denoted by $\{x^{r_k}\}$, such that $f(x^{r_k})^T(x^{r_k} - u) < 0$ for all sufficiently large k. This is a contradiction.

Case (B). There exists a subsequence of $\{x^r\}$ denoted by $\{x^{r_j}\}$, where $r_j \to \infty$ as $j \to \infty$, such that

$$\max_{1\le i\le n}(x_i^{r_j} - u_i)(f_i(x^{r_j}) - f_i(u)) \ge 0 \quad \text{for all } j.$$

By using (4), for each i we have

$$\begin{aligned} A_i^{(r_j)} &:= (x_i^{r_j} - u_i)(f_i(x^{r_j}) - f_i(u)) \\ &= (x_i^{r_j} - u_i)\left(-\frac{1}{2}\left(\frac{1}{\mu_{r_j}} - \mu_{r_j}\right)x_i^{r_j} - f_i(u) + \frac{\mu_{r_j}\varepsilon}{x_i^{r_j}}\right). \end{aligned} \tag{21}$$

There exist a subsequence of $\{x^{r_j}\}$, denoted also by $\{x^{r_j}\}$, and a fixed index m such that

$$A_m^{(r_j)} := (x_m^{r_j} - u_m)(f_m(x^{r_j}) - f_m(u)) = \max_{1\le i\le n}(x_i^{r_j} - u_i)(f_i(x^{r_j}) - f_i(u)).$$

For each i such that $x_i^{r_j} \to \infty$, (21) implies that $A_i^{(r_j)} \to -\infty$. Since $A_m^{(r_j)} \ge 0$ for all j, we deduce that $\{x_m^{r_j}\}$ is bounded, i.e., there is a constant $\bar{\delta}$ such that $0 < x_m^{r_j} \le \bar{\delta}$ for all j.

If $x_m^{r_j} \le u_m$, setting $i = m$ in (21), we have

$$\begin{aligned} A_m^{(r_j)} &\le (u_m - x_m^{r_j})\left(\frac{1}{2}\left(\frac{1}{\mu_{r_j}} - \mu_{r_j}\right)x_m^{r_j} + f_m(u)\right) \\ &\le u_m\left(\frac{1}{2}\left(\frac{1}{\mu_{r_j}} - \mu_{r_j}\right)u_m + f_m(u)\right). \end{aligned} \tag{22}$$

If $u_m < x_m^{r_j} \le \bar{\delta}$, setting $i = m$ in (21), we obtain

$$A_m^{(r_j)} \le (x_m^{r_j} - u_m)\frac{\mu_{r_j}\varepsilon}{x_m^{r_j}} \le \mu_{r_j}\varepsilon < \varepsilon. \tag{23}$$

We consider two subcases, choosing a subsequence whenever it is necessary.

Subcase 1. $\mu_{r_j} \to 1$. From (22) and (23), for all sufficiently large j we have

$$A_m^{(r_j)} \leq \max\left\{\varepsilon, u_m\left(f_m(u)+\frac{u_m}{2}\right)\right\}.$$

Thus, for all sufficiently large j, we obtain

$$\begin{aligned} &f(u)^T(x^{r_j}-u)-\tau\max_{1\leq i\leq n}(x_i^{r_j}-u_i)(f_i(x^{r_j})-f_i(u)) \\ &\geq f(u)^T(x^{r_j}-u)-\tau\max\{\varepsilon, u_m(f_m(u)+u_m/2)\} \\ &> 0. \end{aligned}$$

The last inequality above follows from the fact that $f(u)>0, \{x^{r_j}\}\subset R^n_{++}$, and $\|x^{r_j}\|\to\infty$. Since f is a quasi-P$_*$-map, the above inequality implies that $f(x^{r_j})^T(x^{r_j}-u)\geq 0$ for all sufficiently large j, which is impossible according to Lemma 4.4.

Subcase 2. There exists a subsequence of $\{\mu_{r_j}\}$, denoted also by $\{\mu_{r_j}\}$, such that $\mu_{r_j}\leq\delta^*$ for all j, where $0<\delta^*<1$. In this case, from (22) and (23), we have

$$A_m^{(r_j)} \leq \max\left\{\varepsilon, u_m f_m(u)+\frac{u_m^2}{2}\left(\frac{1}{\mu_{r_j}}-\mu_{r_j}\right)\right\}.$$

It follows from (4) that

$$\begin{aligned} T^{(r_j)} &:= f(x^{r_j})^T(u-x^{r_j})-\tau\max_{1\leq i\leq n}(x_i^{r_j}-u_i)(f_i(x^{r_j})-f_i(u)) \\ &= \frac{1}{2}\left(\frac{1}{\mu_{r_j}}-\mu_{r_j}\right)(\|x^{r_j}\|^2-(x^{r_j})^Tu)+\mu_{r_j}\varepsilon\left(\sum_{i=1}^n\frac{u_i}{x_i^{r_j}}-n\right)-\tau A_m^{(r_j)}. \end{aligned}$$

We now show that $T^{(r_j)}>0$ for all sufficiently large j.

If $\varepsilon\leq u_m f_m(u)+\frac{u_m^2}{2}(\frac{1}{\mu_{r_j}}-\mu_{r_j})$, noting that $\mu_{r_j}\leq\delta^*$ and $\|x^{r_j}\|^2-(x^{r_j})^Tu-\tau u_m^2\to\infty$ as $j\to\infty$, we obtain

$$\begin{aligned} T^{(r_j)} &\geq \frac{1}{2}\left(\frac{1}{\mu_{r_j}}-\mu_{r_j}\right)(\|x^{r_j}\|^2-(x^{r_j})^Tu-\tau u_m^2)-\tau u_m f_m(u)-\mu_{r_j}\varepsilon n \\ &\geq \frac{1}{2}\left(\frac{1}{\delta^*}-\delta^*\right)(\|x^{r_j}\|^2-(x^{r_j})^Tu-\tau u_m^2)-\tau u_m f_m(u)-\delta^*\varepsilon n>0. \end{aligned}$$

If $\varepsilon>u_m f_m(u)+\frac{u_m^2}{2}(\frac{1}{\mu_{r_j}}-\mu_{r_j})$, by the same argument as the above, we can show that

$$T^{(r_j)}\geq\frac{1}{2}\left(\frac{1}{\delta^*}-\delta^*\right)(\|x^{r_j}\|^2-(x^{r_j})^Tu)-\delta\varepsilon n-\tau\varepsilon>0$$

for all sufficiently large j. Thus, by the quasi-P$_*$-property of f, we deduce from $T^{(r_j)}>0$ that $f(u)^T(u-x^{r_j})\geq 0$ for all sufficiently large j. It is a contradiction since $\{x^{r_j}\}\subset R^n_{++}, \|x^{r_j}\|\to\infty$, and $f(u)>0$.

The above contradictions show that f has no interior-point-ε-exceptional family for each $\varepsilon>0$. By Theorem 3.1, the set $\mathcal{U}(\varepsilon)\neq\emptyset$ for any $\varepsilon>0$. The boundedness of the short "interior band " follows from Lemma 4.2, and the upper-semicontinuity of $\mathcal{U}(\varepsilon)$ follows easily from Lemma 2.2. $\square$

The class of quasi-P$_*$-maps includes the quasi monotone functions as particular cases. The following result is an immediate consequence of Theorem 4.2.

Corollary 4.3. *Suppose that f is a continuous quasi monotone (in particular, pseudomonotone) function, and the NCP is strictly feasible.*

(i) *If Condition* 4.3 *is satisfied, then property* (a) *of $\mathcal{U}(\varepsilon)$ holds.*

(ii) *If Condition* 4.2 *is satisfied, then properties* (a) *and* (b) *of $\mathcal{U}(\varepsilon)$ hold.*

In the case when $F_\varepsilon(x)$ is univalent (continuous and one-to-one) in x, the equation $F_\varepsilon(x) = 0$ has at most one solution. Combining this fact and Theorem 4.2, we have the following result concerning the existence of the central path of the NCP. To our knowledge, this result can be viewed as the first existence result on the central path for the NCP with a (generalized) quasi monotone function. Up to now, there is no interior-point type algorithms designed for solving (generalized) quasi monotone complementarity problems.

Corollary 4.4. *Let f be a quasi-P_*-map, and $F_\varepsilon(x)$ is univalent in x. If the NCP is strictly feasible and Condition* 4.2 *is satisfied, then the central path exists and the set $\{x(\varepsilon) : \varepsilon \in (0, \bar{\varepsilon}]\}$ is bounded for any given $\bar{\varepsilon} > 0$.*

Particularly, if f is a P$_0$-function, then $F_\varepsilon(x)$ is univalent in x (see [35]). We have the following result.

Corollary 4.5. *Let f be a continuous P_0 and quasi-P_*-map. If the NCP is strictly feasible and Condition* 4.2 *is satisfied, then the conclusions of Corollary* 4.4 *are valid.*

4.3. $P(\tau, \alpha, \beta)$-maps.

It is well known (see [14, 25, 30, 31]) that the monotonicity combined with strict feasibility implies the existence of the central path. In this section, we extend the result to a class of nonmonotone complementarity problems. Our result states that if f is a P(τ, α, β) and P$_0$-map (see Definition 2.4), the central path exists provided that the NCP is strictly feasible. This result gives an answer to the question "What class of nonlinear functions beyond P$_*$-maps can ensure the existence of the central path if the NCP is strictly feasible?" We first show properties of the mapping $\mathcal{U}(\cdot)$ when f is a $P(\tau, \alpha, \beta)$-map.

Theorem 4.3. *Let f be a continuous $P(\tau, \alpha, \beta)$-map. If the NCP is strictly feasible, then properties* (a) *and* (b) *of $\mathcal{U}(\varepsilon)$ hold. Moreover, if $F_\varepsilon(x)$ is weakly univalent in x, property* (c) *also holds.*

Proof. Suppose that there exists a scalar $\varepsilon > 0$ such that f has an interior-point-ε-exceptional family denoted by $\{x^r\}$. Since $\{x^r\} \subset R^n_{++}$ and $\|x^r\| \to \infty$ as $r \to \infty$, there exist some p and a subsequence denoted by $\{x^{r_j}\}$, where $r_k \to \infty$ as $j \to \infty$, such that $\|x^{r_j}\| \to \infty$ and

$$x_p^{r_j} - u_p = \max_{1 \le i \le n} (x_i^{r_j} - u_i).$$

Clearly, $x_p^{r_j} \to \infty$ as $j \to \infty$. On the other hand, there exists a subsequence of $\{x^{r_j}\}$, denoted also by $\{x^{r_j}\}$, such that for some fixed index m and for all j we have

$$(x_m^{r_j} - u_m)(f_m(x^{r_j}) - f_m(u)) = \max_{1 \le i \le n} (x^{r_j} - u_i)(f_i(x^{r_j}) - f_i(u)).$$

By the definition of the P(τ, α, β)-map, we have

$$\begin{aligned}(x_p^{r_j}-u_p)&(f_p(x^{r_j})-f_p(u))\\ &\geq \min_{1\leq i\leq n}(x_i^{r_j}-u_i)(f_i(x^{r_j})-f_i(u))\\ &\geq -(1+\tau)\max_{1\leq i\leq n}(x_i^{r_j}-u_i)(f_i(x^{r_j})-f_i(u))-\alpha\|x^{r_j}-u\|^\beta\\ &= -(1+\tau)(x_m^{r_j}-u_m)(f_m(x^{r_j})-f_m(u))-\alpha\|x^{r_j}-u\|^\beta. \end{aligned} \tag{24}$$

From (4), we have that $f_p(x^{r_j})<\varepsilon/x_p^{r_j}$, and hence

$$B_p^{(r_j)} := \frac{(x_p^{r_j}-u_p)(f_p(x^{r_j})-f_p(u))}{\|x^{r_j}-u\|^\beta} \leq \frac{(x_p^{r_j}-u_p)}{\|x^{r_j}-u\|^\beta}\left(\frac{\varepsilon}{x_p^{r_j}}-f_p(u)\right). \tag{25}$$

It is easy to see that

$$\frac{\|x^{r_j}-u\|^\beta}{x_p^{r_j}-u_p} = \left(\frac{\|x^{r_j}-u\|}{x_p^{r_j}-u_p}\right)^\beta \cdot \frac{1}{(x_p^{r_j}-u_p)^{(1-\beta)}} \leq \frac{n^{\beta/2}}{(x_p^{r_j}-u_p)^{1-\beta}}. \tag{26}$$

Combining (25) and (26) leads to

$$B_p^{(r_j)} \to -\infty \quad \text{as } j\to\infty.$$

From

$$B_p^{(r_j)} \geq B_{min}^{r_j} := \min_{1\leq i\leq n}\frac{(x_i^{r_j}-u_i)(f_i(x^{r_j})-f_i(u))}{\|x^{r_j}-u\|^\beta},$$

we deduce that

$$B_{min}^{r_j} \to -\infty \quad \text{as } j\to\infty. \tag{27}$$

We now show that $\{x_m^{r_j}\}$ is bounded. Assume that there exists a subsequence of $\{x_m^{r_j}\}$, denoted still by $\{x_m^{r_j}\}$, such that $x_m^{r_j}\to\infty$. Then, from (21), we have

$$(x_m^{r_j}-u_m)(f_m(x^{r_j})-f_m(u)) \to -\infty,$$

and hence for all sufficiently large j we have

$$B_m^{(r_j)} := \frac{(x_m^{r_j}-u_m)(f_m(x^{r_j})-f_m(u))}{\|x^{r_j}-u\|^\beta} = \max_{1\leq i\leq n}\frac{(x_i^{r_j}-u_i)(f_i(x^{r_j})-f_i(u))}{\|x^{r_j}-u\|^\beta} < 0.$$

By (27) and the above relation, we obtain

$$(1+\tau)B_m^{(r_j)} + B_{min}^{r_j} \to -\infty \text{ as } j\to\infty. \tag{28}$$

However, since f is a P(τ,α,β)-map, we have

$$(1+\tau)B_m^{(r_j)} + B_{min}^{r_j} \geq -\alpha,$$

which contradicts (28). This contradiction shows that the sequence $\{x_m^{r_j}\}$ is bounded.

By using (4) and (24), we have

$$\begin{aligned}&-(x_p^{r_j}-u_p)\left(\frac{1}{2}\left(\frac{1}{\mu_{r_j}}-\mu_{r_j}\right)x_p^{r_j}+f_p(u)-\frac{\mu_{r_j}\varepsilon}{x_p^{r_j}}\right)\\ &\quad\geq (1+\tau)(x_m^{r_j}-u_m)\left(\frac{1}{2}\left(\frac{1}{\mu_{r_j}}-\mu_{r_j}\right)x_m^{r_j}+f_m(u)-\frac{\mu_{r_j}\varepsilon}{x_m^{r_j}}\right)-\alpha\|x^{r_j}-u\|^\beta.\end{aligned}$$

Multiplying both sides of the above inequality by $1/(x_p^{r_j} - u_p)$, rearranging terms, and using (26), we have

$$\begin{aligned}
&-\frac{1}{2}\left(\frac{1}{\mu_{r_j}} - \mu_{r_j}\right)\left(x_p^{r_j} + \frac{(1+\tau)x_m^{r_j}(x_m^{r_j} - u_m)}{x_p^{r_j} - u_p}\right)\\
&\geq f_p(u) - \frac{\mu_{r_j}\varepsilon}{x_p^{r_j}} + (1+\tau)\left(\frac{f_m(u)(x_m^{r_j} - u_m)}{x_p^{r_j} - u_p} - \frac{\mu_{r_j}\varepsilon(x_m^{r_j} - u_m)}{x_m^{r_j}(x_p^{r_j} - u_p)}\right) - \frac{\alpha\|x^{r_j} - u\|^\beta}{x_p^{r_j} - u_p}\\
&\geq f_p(u) - \frac{\varepsilon}{x_p^{r_j}} - \frac{(1+\tau)f_m(u)u_m}{x_p^{r_j} - u_p} - \frac{(1+\tau)\varepsilon}{x_p^{r_j} - u_p} - \frac{\alpha n^{\beta/2}}{(x_p^{r_j} - u_p)^{1-\beta}}.
\end{aligned}$$

For all sufficiently large j, the left-hand side of the above inequality is negative, but the right-hand side tends to $f_p(u) > 0$ as $j \to \infty$. This is a contradiction. The contradiction shows that f has no interior-point-ε-exceptional family for every $\varepsilon > 0$. By Theorem 3.1, property (a) of $\mathcal{U}(\varepsilon)$ follows. The proof of the boundedness of the set $\bigcup_{\varepsilon\in(0,\bar{\varepsilon}]}\mathcal{U}(\varepsilon)$ is not straightforward. It can be proved by the same argument as the above. Indeed, we suppose that $\{x(\varepsilon_k)\}_{0<\varepsilon_k<\bar{\varepsilon}} \subseteq \bigcup_{\varepsilon\in(0,\bar{\varepsilon}]}\mathcal{U}(\varepsilon)$ is an unbounded sequence. Replacing $\{x^{r_j}\}$ by $\{x(\varepsilon_k)\}$, using

$$f(x(\varepsilon_k)) = \frac{\varepsilon_k}{x(\varepsilon_k)} < \frac{\bar{\varepsilon}}{x(\varepsilon_k)}$$

instead of (4), and repeating the aforementioned proof, we can derive a contradiction. The upper-semicontinuity of $\mathcal{U}(\cdot)$ can be obtained by Lemma 2.2. The proof is complete. □

The class of $\mathrm{P}(\tau,\alpha,\beta)$-maps includes several particular cases such as $P(\tau,\alpha,0)$-, $\mathrm{P}(\tau,0,0)$-, and $\mathrm{P}(0,\alpha,\beta)$-maps. It is shown in [46] that the class of $\mathrm{P}(\tau,0,0)$-maps coincides with the class of P_*-maps. Therefore, f is said to be a P_*-map if and only if there exists a nonnegative scalar $\kappa \geq 0$ such that

$$(1+\kappa)\max_{1\leq i\leq n}(x_i - y_i)(f_i(x) - f_i(y)) + \min_{1\leq i\leq n}(x_i - y_i)(f_i(x) - f_i(y)) \geq 0.$$

Particularly, a matrix $M \in R^{n\times n}$ is a P_*-matrix if and only if there is a constant $\kappa \geq 0$ such that

$$(1+\kappa)\max_{1\leq i\leq n} x_i(Mx)_i + \min_{1\leq i\leq n} x_i(Mx)_i \geq 0.$$

This is an equivalent definition of the concept of a P_*-matrix (sufficient matrix) introduced by Kojima et al. [26] and Cottle, Pang, and Venkateswaran [9]. The following result follows immediately from Theorem 4.3.

Corollary 4.6. *Let f be a continuous P_0 and $P(\tau,\alpha,\beta)$-map. If the NCP is strictly feasible, then the central path exists and any slice of it is bounded.*

It is worth noting that each P_*-map is a P_0 and a $\mathrm{P}(\tau,\alpha,\beta)$-function. The following result is a straightforward consequence of the above corollary.

Corollary 4.7. *Let f be a continuous P_*-map. If the NCP is strictly feasible, then the central path exists and any slice of it is bounded.*

It should be pointed out that P_*-maps are also special instances of quasi-P_*-maps. A result similar to Corollary 4.3 can be stated for P_*-maps. However, as we have shown in Corollary 4.7, the additional conditions such as Conditions 4.1, 4.2, and 4.3 are not necessary for a P_*-map to guarantee the existence of the central path. While P_*-maps and quasi monotone functions are contained in the class of quasi-P_*-maps, Zhao and Isac [46] gave examples to show that a P_*-map, in general, is not a quasi monotone function, and vice versa.

4.4. Exceptionally regular functions. In section 4.1, we study the properties of the mapping $\mathcal{U}(\varepsilon)$ for E_0-functions satisfying a properness condition, i.e., Condition 4.1. In sections 4.2, we show properties of $\mathcal{U}(\varepsilon)$ for quasi-P_*-maps under the strictly feasible condition as well as some properness conditions. In the above section, properness assumptions are removed, and properties of $\mathcal{U}(\varepsilon)$ for $P(\tau, \alpha, \beta)$-maps are proved under the strictly feasible condition only. In this section, removing both the strictly feasible condition and properness conditions, we prove that properties of $\mathcal{U}(\varepsilon)$ hold if f is an exceptionally regular function. The exceptional regularity of a function (see Definition 2.5) was originally introduced in [46] to investigate the existence of a solution to the NCP.

Definition 4.1. [16] *A map $v : R^n \to R^n$ is said to be positively homogeneous of degree $\alpha > 0$ if $v(tx) = t^\alpha v(x)$ for all $x \in R^n$.*

When $\alpha = 1$, the above concept reduces to the standard concept of positive homogeneity. Under the assumption of positively homogeneous of degree $\alpha > 0$, we can show that properties (a) and (b) of $\mathcal{U}(\varepsilon)$ hold if f is exceptionally regular. See the following result.

Theorem 4.4. *Let f be a continuous and exceptionally regular function from R^n into R^n. If $G(x) = f(x) - f(0)$ is positively homogeneous of degree $\alpha > 0$, then properties* (a) *and* (b) *of $\mathcal{U}(\varepsilon)$ hold. Moreover, if $F_\varepsilon(x)$ is weakly univalent, property* (c) *also holds.*

Proof. Suppose that there is a scalar $\varepsilon > 0$ such that f has an interior-point-ε-exceptional family $\{x^r\}$. We derive a contradiction. Indeed, since $G(x)$ is positively homogeneous of degree $\alpha > 0$, we have

$$f(x^r) = f(0) + \|x^r\|^\alpha (f(x^r/\|x^r\|) - f(0)).$$

Without loss of generality, assume that $x^r/\|x^r\| \to \hat{x}$. From the above relation, we have

$$\lim_{r\to\infty} f(x^r)/\|x^r\|^\alpha = f(\hat{x}) - f(0) = G(\hat{x}). \tag{29}$$

From (4), we have

$$\frac{1}{2}\left(\frac{1}{\mu_r} - \mu_r\right) = -\frac{f_i(x^r)}{x_i^r} + \frac{\mu_r \varepsilon}{(x_i^r)^2} \quad \text{for all } i = 1, \dots, n. \tag{30}$$

Let $I_+(\hat{x}) = \{i : \hat{x}_i > 0\}$. Since $\|x^r\| \to \infty$ and $x_i^r/\|x^r\| \to \hat{x}_i$, we deduce that $x_i^r \to \infty$ for each $i \in I_+(\hat{x})$. We now show that

$$\lim_{r\to\infty} \frac{1}{2}\left(\frac{1}{\mu_r} - \mu_r\right) \frac{\|x^r\|}{\|x^r\|^\alpha} = \hat{\mu} \tag{31}$$

for some $\hat{\mu} \geq 0$. It is sufficient to show the existence of the above limit. Indeed, for each $i \in I_+(\hat{x})$, by using (30) and (29) we have

$$\lim_{r\to\infty} \frac{1}{2}\left(\frac{1}{\mu_r} - \mu_r\right) \frac{\|x^r\|}{\|x^r\|^\alpha} = \lim_{r\to\infty} \frac{\|x^r\|}{x_i^r} \left(-\frac{f_i(x^r)}{\|x^r\|^\alpha} + \frac{\mu_r \varepsilon}{x_i^r \|x^r\|^\alpha}\right) = -\frac{G_i(\hat{x})}{\hat{x}_i}.$$

Thus, (31) holds, with

$$\frac{G_i(\hat{x})}{\hat{x}_i} = -\hat{\mu} \quad \text{for all } i \in I_+(\hat{x}). \tag{32}$$

Now, we consider the case of $i \notin I_+(\hat{x})$. In this case, $\hat{x}_i = 0$. By using (4), (31), and (29), we see from $x_i^r/\|x^r\| \to 0$ that

$$\begin{aligned} 0 \le \lim_{r\to\infty} \frac{\mu_r \varepsilon}{x_i^r \|x^r\|^\alpha} &= \lim_{r\to\infty} \left(\frac{f_i(x^r)}{\|x^r\|^\alpha} + \frac{(1/\mu_r - \mu_r)x_i^r}{2\|x^r\|^\alpha} \right) \\ &= \lim_{r\to\infty} \left(\frac{f_i(x^r)}{\|x^r\|^\alpha} + \frac{(1/\mu_r - \mu_r)\|x^r\|}{2\|x^r\|^\alpha} \cdot \frac{x_i^r}{\|x^r\|} \right) \\ &= G_i(\hat{x}), \end{aligned}$$

i.e.,

$$G_i(\hat{x}) \ge 0 \quad \text{for all } i \notin I_+(\hat{x}).$$

Combining (32) and the above relation implies that f is not exceptionally regular. This is a contradiction. The contradiction shows that f has no interior-point-ε-exceptional family for each $\varepsilon > 0$, and hence property (a) of $\mathcal{U}(\varepsilon)$ follows from Theorem 3.1. Property (b) of $\mathcal{U}(\varepsilon)$ can be easily proved. Actually, suppose that there exists a sequence $\{x(\varepsilon_k)\}_{0<\varepsilon_k<\bar{\varepsilon}}$ with $\|x(\varepsilon_k)\| \to \infty$, where $x(\varepsilon_k) \in \mathcal{U}(\varepsilon_k)$. Without loss of generality, let $x(\varepsilon_k)/\|x(\varepsilon_k)\| \to \bar{x}$, where $\|\bar{x}\| = 1$. As in the proof of (29) we have

$$0 \le \lim_{k\to\infty} f(x(\varepsilon_k))/\|x(\varepsilon_k)\|^\alpha = G(\bar{x}).$$

Since $x(\varepsilon_k) \in \mathcal{U}(\varepsilon_k)$, we have that $x_i(\varepsilon_k) f_i(x(\varepsilon_k)) = \varepsilon_k$ for all $i = 1, \ldots, n$. Thus,

$$0 = \lim_{k\to\infty} \frac{x_i(\varepsilon_k) f_i(x(\varepsilon_k))}{\|x(\varepsilon_k)\|^{1+\alpha}} = \bar{x}_i G_i(\bar{x}) \quad \text{for all } i = 1, \ldots, n.$$

Therefore,

$$G_i(\bar{x}) = 0 \quad \text{whenever } \bar{x}_i > 0, \ \text{and } G_i(\bar{x}) \ge 0 \quad \text{whenever } \bar{x}_i = 0,$$

which contradicts the exceptional regularity of $f(x)$. □

It is not difficult to see that a strictly copositive map and a strictly semimonotone function are special cases of exceptionally regular maps. Hence, we have the following result.

COROLLARY 4.8. *Suppose that $G(x) = f(x) - f(0)$ is positively homogeneous of degree $\alpha > 0$. Then conclusions of Theorem 4.4 are valid if one of the following conditions holds.*

(i) *f is an E_0-function, and for each $0 \ne x \ge 0$ there exists an index i such that $x_i > 0$ and $f_i(x) \ne f_i(0)$.*

(ii) *f is strictly copositive, that is, $x^T(f(x) - f(0)) > 0$ for all $0 \ne x \ge 0$.*

(iii) *f is a strictly semimonotone function.*

Proof. Since each of the above conditions implies that $f(x)$ is exceptionally regular, the result follows immediately from Theorem 4.4. □

Motivated by Definition 2.5, we introduce the following concept.

DEFINITION 4.2. *$M \in R^{n\times n}$ is said to be an exceptionally regular matrix if for all $\beta \ge 0, M + \beta I$ is an R_0-matrix.*

It is evident that an exceptionally regular matrix is an R_0-matrix, but the converse is not true. The following result is an immediate consequence of Theorem 4.4 and its corollary.

COROLLARY 4.9. *Let $f = Mx + q$, where $M \in R^{n\times n}$, and q is an arbitrary vector in R^n. If one of the following conditions is satisfied, then properties* (a) *and* (b) *of the mapping $\mathcal{U}(\varepsilon)$ hold:*

(i) *$M \in R^{n\times n}$ is an exceptionally regular matrix.*

(ii) *M is a strictly copositive matrix.*

(iii) *M is a strictly semimonotone matrix.*

(iv) *M is an E_0-matrix, and for each $0 \neq x \geq 0$ there exists an index i such that $x_i > 0$ and $(Mx)_i \neq 0$ (possibly, $(Mx)_i < 0$).*

Furthermore, if M is also a P_0-matrix, then the central path of a linear complementarity problem exists and any slice of it is bounded.

The R_0-property of f has played an important role in the complementarity theory. We close this section by considering this situation. The concept of a nonlinear R_0-function was first introduced by Tseng [38] and later modified by Chen and Harker [6]. We now give a definition of the R_0-function that is different from those in [38] and [6].

Definition 4.3. *$f : R^n \to R^n$ is said to be an R_0-function if $x = 0$ is the unique solution to the following complementarity problem:*

$$G(x) = f(x) - f(0) \geq 0,\ x \geq 0,\ x^T G(x) = 0.$$

This concept is a natural generalization of the R_0-matrix [8]. In fact, for the linear function $f(x) = Mx + q$, it is easy to see that f is an R_0-function if and only if M is an R_0-matrix. In the case when f is an E_0-function, we have shown in Theorem 4.1 that there exists a subsequence $\{\mu_{r_k}\}$ such that $\mu_{r_k} \to 1$. Moreover, if G is positively homogeneous, then from (31) we deduce that $\hat{\mu} = 0$. By using these facts and the above R_0-property and repeating the proof of Theorem 4.4, we have the following result.

Theorem 4.5. *Suppose that $G(tx) = tG(x)$ for each scalar $t \geq 0$ and $x \in R^n$, and that f is an E_0 and R_0-function. Then the conclusions of Theorem* 4.4 *remain valid. Moreover, if f is a P_0 and R_0 -function, the central path exists and any slice of it is bounded.*

5. Conclusions. We introduced the concept of the interior-point-ε-exceptional family for continuous functions, which is important since it strongly pertains to the existence of an interior-point $x(\varepsilon) \in \mathcal{U}(\varepsilon)$ and the central path, even to the solvability of NCPs. By means of this concept, we proved that for every continuous NCP the set $\mathcal{U}(\varepsilon)$ is nonempty for each scalar $\varepsilon > 0$ if there exists no interior-point-ε-exceptional family for f. Based on the result, we established some sufficient conditions for the assurance of some desirable properties of the multivalued mapping $\mathcal{U}(\varepsilon)$ associated with certain nonmonotone complementarity problems. Since properties (a) and (b) of $\mathcal{U}(\varepsilon)$ imply that the NCP has a solution, the argument of this paper based on the interior-point-ε-exceptional family can serve as a new analysis method for the existence of a solution to the NCP.

It is worth noting that any point in $\mathcal{U}(\varepsilon)$ is strictly feasible, i.e., $x(\varepsilon) > 0$ and $f(x(\varepsilon)) > 0$. Therefore, the analysis method in this paper can also be viewed as a tool for investigating the strict feasibility of a complementarity problem. In fact, from Theorems 3.1, 4.1, 4.4, and 4.5, we have the following result.

Theorem 5.1. *Let f be a continuous function. Then the complementarity problem is strictly feasible whenever one of the following conditions holds.*

(i) *There exists a scalar $\varepsilon^* > 0$ such that f has no interior-point-ε^*-exceptional family.*

(ii) *f is an E_0-function and Condition* 4.1 *is satisfied.*

(iii) *$G(x) = f(x) - f(0)$ is positively homogeneous of degree $\alpha > 0$ and f is exceptionally regular.*

(iv) *$f(x) = Mx + q$, where M is an E_0 and R_0-matrix.*

It should be pointed out that the results and the argument of this paper can be easily extended to other interior-point paths. For instance, we can consider the existence of the path

$$(33) \qquad \{(x(\varepsilon), y(\varepsilon)) > 0 : \varepsilon > 0, y(\varepsilon) = f(x(\varepsilon)) + \varepsilon b, x_i(\varepsilon) y_i(\varepsilon) = \varepsilon a_i \text{ for all } i\}$$

(where b and $a > 0$ are fixed vectors in R^n) first studied by Kojima, Megiddo, and Noma [25]. (When $a = \varepsilon e, b = 0$, the above path reduces to the central path). This path can be studied by the concept of interior-point-$\varepsilon(a, b)$-exceptional family. For a continuous function $f : R^n \to R^n$, we say that a sequence $\{x^r\} \subset R^n_{++}$ is an interior-point-$\varepsilon(a, b)$-exceptional family for f if $\|x^r\| \to \infty$ as $r \to \infty$, and for each x^r there exists a positive number $\mu_r \in (0, 1)$ such that for each i

$$f_i(x^r) = -\varepsilon b_i + \frac{1}{2}\left[\mu_r - \frac{1}{\mu_r}\right] x_i^r + \frac{\mu_r \varepsilon a_i}{x_i^r}.$$

Using

$$F_i(x, \varepsilon) = x_i + (f_i(x) + \varepsilon b_i) - \sqrt{x_i^2 + (f_i(x) + \varepsilon b_i)^2 + 2\varepsilon a_i}$$

and arguing as in the same proof of Theorem 3.1, we can show that for any $\varepsilon > 0$ there exists either a point $x(\varepsilon)$ satisfying (33) or an interior-point-$\varepsilon(a, b)$-exceptional family for f. This result enables us to develop some sufficient conditions for the existence of the path (33).

Acknowledgments. The authors would like to thank the referees and Professor Jim Burke for their helpful suggestions and comments on an earlier version of this paper, which helped the authors to correct some mistakes and improve the presentation of the manuscript. They also thank Dr. Mustapha Ait Rami for his valuable comments.

REFERENCES

[1] J. Burke and S. Xu, *The global linear convergence of a non-interior-point path following algorithm for linear complementarity problems,* Math. Oper. Res., 23 (1998), pp. 719–734.

[2] J. Burke and S. Xu, *A non-interior predictor-corrector path following algorithm for the monotone linear complementarity problem*, Math. Program., 87 (2000), pp. 113–130.

[3] B. Chen and X. Chen, *A global and local superlinear continuation-smoothing method for P_0 and R_0 NCP or monotone NCP*, SIAM J. Optim., 9 (1999), pp. 624–645.

[4] B. Chen, X. Chen, and C. Kanzow, *A Penalized Fischer-Burmeister NCP-Function: Theoretical Investigation and Numerical Results*, Technical Report, Zur Angewandten Mathematik, Hamburger Beiträge, 1997.

[5] B. Chen and P. T. Harker, *A non-interior-point continuation method for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1168–1190.

[6] B. Chen and P. T. Harker, *Smooth approximations to nonlinear complementarity problems*, SIAM J. Optim., 7 (1997), pp. 403–420.

[7] C. Chen and O. L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.

[8] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[9] R. W. Cottle, J. S. Pang, and V. Venkateswaran, *Sufficient matrices and the linear complementarity problem*, Linear Algebra Appl., 114/115 (1989), pp. 231–249.

[10] F. Facchinei, *Structural and stability properties of P_0 nonlinear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 735–749.

[11] F. Facchinei and C. Kanzow, *Beyond monotonicity in regularization methods for nonlinear complementarity problems*, SIAM J. Control Optim., 37 (1999), pp. 1150–1161.

[12] M. C. Ferris and J. S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.

[13] M. S. Gowda and M. A. Tawhid, *Existence and limiting behavior of trajectories associated with P_0-equations*, Comput. Optim. Appl., 12 (1999), pp. 229–251.

[14] O. Güler, *Existence of interior points and interior-point paths in nonlinear monotone complementarity problems*, Math. Oper. Res., 18 (1993), pp. 128–147.

[15] O. Güler, *Path Following and Potential Reduction Algorithm for Nonlinear Monotone Complementarity Problems*, Technical Report, Department of Management Sciences, The University of Iowa, Iowa City, 1990.

[16] P. T. Harker and J. S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Programming, 48 (1990), pp. 161–220.

[17] K. Hotta and A. Yoshise, *Global convergence of a class of non-interior-point algorithms using Chen-Harker-Kanzow functions for nonlinear complementarity problems*, Math. Program., 86 (1999), pp. 105–133.

[18] G. Isac, *Complementarity Problems*, Lecture Notes in Math. 1528, Springer-Verlag, Berlin, 1992.

[19] G. Isac, V. Bulavski, and V. Kalashnikov, *Exceptional families, topological degree and complementarity problems*, J. Global Optim., 10 (1997), pp. 207–225.

[20] G. Isac and W. T. Obuchowska, *Functions without exceptional families of elements and complementarity problems*, J. Optim. Theory Appl., 99 (1998), pp. 147–163.

[21] C. Kanzow, *Some nonlinear continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.

[22] C. Kanzow, N. Yamashita, and M. Fukushima, *New NCP-functions and their properties*, J. Optim. Theory Appl., 94 (1997), pp. 115–135.

[23] S. Karamardian, *Complementarity problems over cones with monotone and pseudomonotone maps*, J. Optim. Theory Appl., 18 (1976), pp. 445–454.

[24] S. Karamardian and S. Schaible, *Seven kinds of monotone maps*, J. Optim. Theory Appl., 66 (1990), pp. 37–46.

[25] M. Kojima, N. Megiddo, and T. Noma, *Homotopy continuation methods for nonlinear complementarity problems*, Math. Oper. Res., 16 (1991), pp. 754–774.

[26] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Comput. Sci. 538, Springer-Verlag, New York, 1991.

[27] M. Kojima, M. Mizuno, and T. Noma, *A new continuation method for complementarity problems with uniform P-functions*, Math. Programming, 43 (1989), pp. 107–113.

[28] M. Kojima, M. Mizuno, and A. Yoshise, *A polynomial-time algorithm for linear complementarity problems*, Math. Programming, 44 (1989), pp. 1–26.

[29] Z. Q. Luo and P. Tseng, *A new class of merit functions for the nonlinear complementarity problem*, in Complementarity and Variational Problems: State of the Art, M.C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997, pp. 204–225.

[30] L. McLinden, *The complementarity problem for maximal monotone multifunctions*, in Variational Inequalities and Complementarity Problems, R.W. Cottle, F. Giannessi, and J.L. Lions, eds., John Wiley and Sons, New York, 1980, pp. 251–270.

[31] N. Megiddo, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming: Interior-Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.

[32] R. D. C. Monteiro and I. Adler, *Interior path following primal dual algorithms, Part I: Linear programming*, Math. Programming, 44 (1989), pp. 27–42.

[33] R. D. C. Monteiro and J. S. Pang, *Properties of an interior-point mapping for mixed complementarity problems*, Math. Oper. Res., 21 (1996), pp. 629–654.

[34] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[35] G. Ravindran and M. S. Gowda, *Regularization of P_0-functions in box variational inequality problems*, SIAM J. Optim., to appear.

[36] T. E. Smith, *A solution condition for complementarity problems with an application to spatial price equilibrium*, Appl. Math. Comput., 15 (1984), pp. 61–69.

[37] R. Sznajder and M. S. Gowda, *On the limiting behavior of the trajectory of regularized solutions of P_0 complementarity problems*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic

Publishers, Dordrecht, The Netherlands, 1998, pp. 317–379.

[38] P. Tseng, *Growth behavior of a class of merit functions for the nonlinear complementarity problems*, J. Optim. Theory Appl., 89 (1996), pp. 17–37.

[39] P. Tseng, *An infeasible path-following method for monotone complementarity problems*, SIAM J. Optim., 7 (1997), pp. 386–402.

[40] H. Väliaho, *P_* matrices are just sufficient*, Linear Algebra Appl., 239 (1996), pp. 103–108.

[41] V. Venkateswaran, *An algorithm for the linear complementarity problem with a P_0-matrix*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 967–977.

[42] D. Zhang and Y. Zhang, *On Constructing Interior-Point Path-Following Methods for Certain Semimonotone Linear Complementarity Problems*, Technical Report, Department of Computational and Applied Mathematics, Rice University, Houston, 1997.

[43] Y. B. Zhao, *Existence of a solution to nonlinear variational inequality under generalized positive homogeneity*, Oper. Res. Lett., 25 (1999), pp. 231–239.

[44] Y. B. Zhao and J. Han, *Exceptional family of elements for a variational inequality problem and its applications*, J. Global Optim., 14 (1999), pp. 313–330.

[45] Y. B. Zhao, J. Han, and H. D. Qi, *Exceptional families and existence theorems for variational inequality problems*, J. Optim. Theory Appl., 101 (1999), pp. 475–495.

[46] Y. B. Zhao and G. Isac, *Quasi-P_*-maps, $P(\tau,\alpha,\beta)$-maps, exceptional family of elements and complementarity problems*, J. Optim. Theory Appl., 105 (2000), pp. 213–231.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 594–634

THE VELOCITY TRACKING PROBLEM FOR NAVIER–STOKES FLOWS WITH BOUNDARY CONTROL*

M. D. GUNZBURGER† AND S. MANSERVISI‡

Abstract. We present some systematic approaches to the mathematical formulation and numerical approximation of the time-dependent optimal control problem of tracking the velocity for Navier–Stokes flows in a bounded, two-dimensional domain with boundary control. We study the existence of optimal solutions and derive an optimality system from which optimal solutions may be determined. We also define and analyze semidiscrete-in-time and full space-time discrete approximations of the optimality system and a gradient method for the solution of the fully discrete system. The results of some computational experiments are provided.

Key words. optimal control, Navier–Stokes equations, fluid mechanics

AMS subject classifications. 35B40, 35B37, 35Q30, 65M60

PII. S0363012999353771

1. Introduction.

In this paper, we study a class of optimal flow control problems for which the fluid motion is controlled by velocity forcing, i.e., injection or suction, along a portion of the boundary and the cost or objective functional is a measure of the discrepancy between the flow velocity and a given target velocity. The fluid motion is constrained to satisfy the time-dependent Navier–Stokes equations for viscous, incompressible flows. In order to regularize the problem, e.g., to limit the size of the control, a quadratic penalization functional depending on the control is added to the cost functional.

Time-dependent optimal flow control problems have been considered by numerous authors, e.g., [1], [4], [5], [6], [7], [13], [14], [15], [16], [22], [23], [24], [25], [26], [27], [29], [30], [31], and [33]. A number of these papers deal with distributed force control in the momentum equation; this type of control is not realizable in most practical situations. Some of these papers only treat questions concerning the existence of optimal solutions and the derivation of optimality systems from which optimal controls and states may be deduced. Others present formal derivations of optimality systems, define algorithms for the approximation of solutions of these systems, and the results of numerical experiments. The objective functionals considered in most of these papers differ from the one considered here, although in many cases, including the one of this paper, the analysis can be extended. Furthermore, in much of the literature, controls are assumed to be separable, i.e., to be the product of a function of time and a function of the space variables; in some cases the second function is assumed known so that control is effected only through a function depending on time. Finally, in some papers, the optimal control problems are set in function spaces that are not practical for numerical approximations.

*Received by the editors February 22, 1999; accepted for publication (in revised form) April 10, 2000; published electronically September 20, 2000.
http://www.siam.org/journals/sicon/39-2/35377.html

†Department of Mathematics, Iowa State University, Ames, IA 50011-2064 (gunzburg@iastate.edu). The research of this author was supported in part by the Air Force Office of Scientific Research under grant number F49620-95-1-0407.

‡Mathematical Institute, University of Oxford, 24-29 St. Giles', Oxford OX1 3LB, UK (manservi@maths.ox.ac.uk).

For example, in [1], partial results about optimal boundary control problems are given. The formulation is not suitable for some applications and the function space proposed for the solution may not allow easy finite element implementations. More consistent are the results in [16] for the drag minimization problem in unbounded domains. In that paper, the method used for the derivation of the optimality system is different from the one used here. Their analyses also apply to bounded domains and the method can be applied with small changes to other cost functionals. However, the function spaces used there are again not useful in practical calculations and no numerical algorithms or numerical analyses are given. In [4], algorithmic issues are discussed and numerical experiments are presented for separable boundary controls, but no numerical analyses are given.

The aim of this paper is to provide a complete and consistent analysis of an optimal boundary control problem for the Navier–Stokes equations and to present the results of some simple numerical experiments. The mathematical framework of this paper, as well as the algorithms developed and analyzed, can be extended to a broad class of practical boundary control problems for partial differential equations in bounded domains.

We now describe the problem of time-dependent boundary control for the Navier–Stokes system that models the velocity tracking problem through a quadratic functional. This problem reflects the desire to steer, over time, a candidate velocity field $\vec{u}$ to a given target velocity field $\vec{U}$ by appropriately controlling the velocity along a portion of the boundary of the flow domain. We consider a two-dimensional flow over the time interval $[0,T]$ in the physical domain Ω with boundary Γ with control effected over $\Gamma_c \subset \Gamma$. The equations considered here are the nondimensional incompressible Navier–Stokes equations

$$(1)\qquad \begin{cases} \vec{u}_t + (\vec{u}\cdot\nabla)\vec{u} - \nu\Delta\vec{u} + \nabla p = \vec{0} & \text{in } (0,T)\times\Omega, \\ \nabla\cdot\vec{u} = 0 \quad \text{in } (0,T)\times\Omega, \\ \vec{u} = \vec{g} \quad \text{on } (0,T)\times\Gamma_c, \\ \vec{u} = \vec{0} \quad \text{on } (0,T)\times(\Gamma\setminus\Gamma_c) \end{cases}$$

with initial velocity $\vec{u}(0,\vec{x}) = \vec{u}_0(\vec{x})$. The vector $\vec{u} = (u_1, u_2)$ denotes the velocity, p the pressure, and ν the constant kinematic viscosity coefficient. We note that for appropriate nondimensionalizations, the Reynolds number is equal to $1/\nu$. The boundary velocity control is denoted by $\vec{g}$ and is required to satisfy the compatibility conditions

$$(2)\qquad \int_{\Gamma_c} \vec{g}\cdot\vec{n}\, d\vec{x} = 0,$$

where $\vec{n}$ denotes the unit outward normal vector along Γ, and

$$(3)\qquad \vec{g}|_{t=0} = \vec{u}_0|_{\Gamma_c}.$$

Thus, the control is required to effect zero mass flow across the boundary and to match, at the initial time, the initial flow $\vec{u}_0$ on the boundary. The first of these is necessary in view of the incompressibility condition and the second in order to obtain the appropriate regularity for the solution of the Navier–Stokes system.

The optimal control problem is formulated as follows:

find a boundary control $\vec{g}$ and a velocity field $\vec{u}$ such that the cost functional

$$\mathcal{J}(\vec{u},\vec{g}) = \frac{\alpha}{2}\int_0^T \int_\Omega |\vec{u}-\vec{U}|^2 \, d\vec{x}dt + \frac{\beta}{2}\int_0^T \int_{\Gamma_c} (|\vec{g}|^2 + \beta_1|\vec{g}_t|^2 + \beta_2|\vec{g}_x|^2) \, d\vec{x}dt \tag{4}$$

is minimized subject to $(\vec{u},\vec{g})$ *satisfying* (1)–(2).

The minimization of the first term involving $(\vec{u}-\vec{U})$ in (4) is the real goal of the velocity tracking problem; the other terms have been introduced in order to bound the control function and to prove the existence of an optimal control. We can effectively limit the size of the control through an appropriate choice of the positive coefficients β, β_1, and β_2.

This paper is organized as follows. In the remainder of this section we introduce some notation and results that will be useful in what follows. In section 2, we give a precise definition of the optimal control problem and prove the existence of optimal solutions. Also, first-order necessary conditions are derived and optimal solutions are characterized as solution of a system of partial differential equations. Semidiscretizations in time and full space-time discretizations are treated in sections 3 and 4, respectively. Issues related to the numerical implementation of the fully discrete algorithms are discussed in section 5 and the results of some computational experiments are presented in section 6.

1.1. Notation and preliminary results.

We introduce the following standard notations over a bounded, connected, open set Ω in $\mathbb{R}^2$ with boundary $\Gamma \in C^2$. Let $\vec{n} = (n_1, n_2)$ and $\vec{\tau}$ denote the unit normal and tangent vectors, respectively. Let $I = (0,T)$, $Q = I \times \Omega$, $S = I \times \Gamma$, and $S_c = I \times \Gamma_c$, where Γ_c denotes the part of the boundary on which control is applied. Also, we denote $\Omega_0 = \Omega \times \{0\}$ and $\Gamma_0 = \Gamma \times \{0\}$.

We shall use the standard notations for the Sobolev spaces (and their vector-valued, i.e., $\mathbb{R}^2$-valued, counterparts) $H^m(\Omega)$ with norm $\|\cdot\|_m$; we also use the notations $L^2(\Omega) = H^0(\Omega)$ with $\|\cdot\| = \|\cdot\|_0$ and $\mathcal{D}(\Omega)$ for the space of distributions. Let $H_0^m(\Omega)$ denote the closure of $C_0^\infty(\Omega)$ under the norm $\|\cdot\|_m$ and $H_0^{-m}(\Omega)$ denote the dual space of $H_0^m(\Omega)$. We introduce the solenoidal spaces $\mathcal{V}(\Omega)$, $V(\Omega)$, and $W(\Omega)$ as

$$\mathcal{V}(\Omega) = \{\vec{u} \in C_0^\infty(\Omega) : \nabla\cdot\vec{u} = 0\},$$
$$V(\Omega) = \{\vec{u} \in H_0^1(\Omega) : \nabla\cdot\vec{u} = 0\},$$
$$W(\Omega) = \{\vec{u} \in L^2(\Omega) : \nabla\cdot\vec{u} = 0\}.$$

The dual space of $V(\Omega)$ is denoted by $V(\Omega)^*$. Also, we define

$$L_0^2(\Omega) = \left\{ p \in L^2(\Omega) : \int_\Omega p \, d\vec{x} = 0 \right\}.$$

Let X be a Banach space and (a,b) an open set of $\mathbb{R}$. We denote by $L^p((a,b);X)$ $(1 \le p < \infty)$ the space of functions $f(t) : (a,b) \to X$ such that f is measurable and

$$\|f\|_{L^p((a,b);X)} = \left(\int_a^b \|f(t)\|_X^p \right)^{1/p}$$

is finite. We also denote by $L^\infty((a,b);X)$ the space of functions f from (a,b) into X such that f is measurable and is bounded almost everywhere (a.e.) over (a,b) and we set

$$\|f\|_{L^\infty((a,b);X)} = \inf_{\|f(t)\| \le M a.e.} (M).$$

We define the following anisotropic Sobolev spaces. Let r and $s \geq 0$ and $Q = (a,b) \times \Omega$. We let

$$H^{r,s}(Q) = L^2((a,b); H^r(\Omega)) \cap H^s((a,b); L^2(\Omega)) \tag{5}$$

with the norm

$$\|u\|_{H^{r,s}} = (\|u\|^2_{L^2((a,b);H^r)} + \|u\|^2_{H^s((a,b);L^2)})^{1/2}.$$

For details about these spaces, see, e.g., [2], [12], [17], [19], and [28].

In order to define a weak form of the Navier–Stokes equations, we introduce two continuous bilinear forms

$$a(\vec{u},\vec{v}) = 2\nu \sum_{i,j=1}^{n} \int_\Omega D_{ij}(\vec{u}) D_{ij}(\vec{v})\, d\vec{x} \qquad \forall\, \vec{u},\, \vec{v} \in H^1(\Omega), \tag{6}$$

$$b(\vec{v},q) = -\int_\Omega q \nabla \cdot \vec{v}\, d\vec{x} \qquad \forall\, q \in L^2(\Omega), \quad \forall\, \vec{v} \in H^1(\Omega), \tag{7}$$

where $D_{ij}(\vec{v}) = \frac{1}{2}(\partial v_i/\partial x_j + \partial v_j/\partial x_i)$, and the continuous trilinear form

$$c(\vec{w};\vec{u},\vec{v}) = \sum_{i,j=1}^{n} \int_\Omega w_j \left(\frac{\partial u_i}{\partial x_j}\right) v_i\, d\vec{x} \qquad \forall\, \vec{w},\, \vec{u},\, \vec{v} \in H^1(\Omega).$$

We will also make use the following operators:

$$\begin{aligned}
&A : H^1(\Omega) \to H^{-1}(\Omega) && (8)\\
&\qquad \langle A\vec{u}, \vec{v}\rangle = a(\vec{u},\vec{v}) \quad \forall\, \vec{u} \in H^1(\Omega), \quad \forall\, \vec{v} \in H^1_0(\Omega), \\
&C : H^1(\Omega) \times H^1(\Omega) \to H^{-1}(\Omega) && (9)\\
&\qquad \langle C(\vec{w})\vec{u}, \vec{v}\rangle = c(\vec{w};\vec{u},\vec{v}) \quad \forall\, \vec{w}, \vec{u} \in H^1(\Omega), \quad \forall\, \vec{v} \in H^1_0(\Omega), \\
&B : H^1(\Omega) \to L^2_0(\Omega) && (10)\\
&\qquad \langle B\vec{u}, p\rangle = b(\vec{u},p) \quad \forall\, p \in L^2_0(\Omega), \quad \forall\, \vec{u} \in H^1(\Omega), \\
&B^* : L^2_0(\Omega) \to H^{-1}(\Omega) && (11)\\
&\qquad \langle \vec{u}, B^* p\rangle = b(\vec{u},p) \quad \forall\, p \in L^2_0(\Omega), \quad \forall\, \vec{u} \in H^1_0(\Omega).
\end{aligned}$$

We will denote by πA and πC the projections of these operators on $V(\Omega)$.

In the rest of the paper, we limit our domain $\Omega \subset \mathbb{R}^2$ to be an open bounded set with simply connected boundary $\Gamma \in C^2$. We define

$$\begin{aligned}
&curl(H^2)(\Omega) = \left\{ \vec{v} \in H^1(\Omega) \ : \ \nabla \cdot \vec{v} = 0, \quad \int_\Gamma \vec{v} \cdot \vec{n}\, d\vec{x} = 0 \right\}, \\
&H^1_n(\Gamma) = \left\{ \vec{g} \in H^1(\Gamma) \ : \ \int_\Gamma \vec{g} \cdot \vec{n}\, d\vec{x} = 0 \right\}, \\
&H^1_{n0}(\Gamma_c) = H^1_0(\Gamma_c) \cap H^1_n(\Gamma_c),
\end{aligned}$$

where Γ_c is part of the boundary Γ. The set $curl(H^2)(\Omega)$ is a closed subspace of $H^1(\Omega)$, and $H^1_n(\Gamma)$ and $H^1_{n0}(\Gamma)$ are closed subspaces of $H^1(\Gamma)$. For details concerning these subspaces, see, e.g., [12]. We remark that the space $H^1(\Gamma)$ can be decomposed in $H^1_n(\Gamma) \oplus (H^1_n(\Gamma))^\perp$, where $(H^1_n(\Gamma))^\perp$ is the space of vectors normal to the surface with constant length. If $\vec{g} \in H^1(\Gamma)$, one can write $\vec{g} = \vec{g}_1 + \vec{g}_2$, where

$$\vec{g}_2 = a\vec{n}, \qquad a = \frac{\int_\Gamma \vec{g} \cdot \vec{n}\, d\vec{x}}{\mu(\Gamma)},$$
$$\vec{g}_1 = \vec{\tau}(\vec{g} \cdot \vec{\tau}) + \vec{n}(\vec{g} \cdot \vec{n} - a),$$

where $\vec{g}_1 \in H^1_n(\Gamma)$ and $\vec{g}_2 \in (H^1_n(\Gamma))^\perp$.

Some useful properties of the trilinear form $c(\vec{u}; \vec{v}, \vec{w})$ can be summarized as follows (see [1] and [29]):

(i)

$$(12) \quad \begin{cases} c(\vec{u};\vec{v},\vec{w}) = -c(\vec{u};\vec{w},\vec{v}) & \forall\, \vec{u} \in V(\Omega), \quad \forall\, \vec{v},\vec{w} \in H^1(\Omega), \\ c(\vec{u};\vec{v},\vec{w}) = -c(\vec{u};\vec{w},\vec{v}) & \forall\, \vec{u} \in curl(H^2)(\Omega), \quad \forall\, \vec{v} \in H^1(\Omega), \quad \forall\, \vec{w} \in H^1_0(\Omega), \end{cases}$$

(ii)

$$(13) \quad \begin{cases} c(\vec{u};\vec{v},\vec{v}) = 0 & \forall\, \vec{u} \in V(\Omega), \quad \forall\, \vec{v} \in H^1(\Omega), \\ c(\vec{u};\vec{v},\vec{v}) = 0 & \forall\, \vec{u} \in curl(H^2)(\Omega), \quad \forall\, \vec{v} \in H^1_0(\Omega), \end{cases}$$

(iii)

$$(14) \quad \begin{cases} |c(\vec{u};\vec{v},\vec{w})| \le \sqrt{2}\|\vec{u}\|^{1/2}\|\nabla\vec{u}\|^{1/2}\|\vec{w}\|^{1/2}\|\nabla\vec{w}\|^{1/2}\|\nabla\vec{v}\| & \forall\, \vec{u},\vec{v},\vec{w} \in H^1(\Omega), \\ |c(\vec{u};\vec{v},\vec{w})| \le C\|\vec{u}\|_1\|\vec{v}\|_1\|\vec{w}\|_1 \quad \forall\, \vec{u},\vec{v},\vec{w} \in H^1(\Omega), \end{cases}$$

where C is independent of the functions $\vec{u}, \vec{v}$, and $\vec{w}$.

We remark that (iii) cannot be extended to a three-dimensional domain. Since we shall make extensive use of this result, all our results will hold only for two-dimensional domains.

The form $c(\vec{u};\vec{u},\vec{w})$ is differentiable in $\vec{u}$ from $V(\Omega)$ into $V(\Omega)^*$. We denote by $c'(\vec{u};\vec{v},\vec{w})$ its variation with respect to a variation $\vec{v}$ in $\vec{u}$ and by $c^*(\vec{u};\vec{v},\vec{w})$ the adjoint of $c(\vec{v};\vec{u},\vec{w})$ for the duality between $V(\Omega)$ and $V(\Omega)^*$. For details about these derived forms, one may refer to [1].

In the rest of the paper we shall use γ and γ_k to denote trace operators, i.e., $\gamma\vec{f} = \gamma_0\vec{f} = \vec{f}_\Gamma$ and $\gamma_0\partial^k_n\vec{f} = \gamma_k\vec{f}$, where $\partial_n\vec{f} = n_1\partial_1\vec{f} + n_2\partial_2\vec{f}$, with $\partial_j\vec{f} = \partial\vec{f}/\partial x_j$.

2. Formulation and analysis of the optimal control problem.

2.1. Weak formulation of the optimal control problem. We consider an open bounded set $\Omega \subset \mathbb{R}^2$ with a boundary $\Gamma \in C^2$. $\vec{U}(t,\vec{x})$ is said to be in the set of *admissible target velocities* U_{ad} if

$$(15) \qquad \begin{cases} \vec{U} = \vec{U}(t,\vec{x}) \in C([0,T];H^1(\Omega)), \\ \vec{F}_{\vec{U}}(t,\vec{x}) \in L^\infty((0,T);L^2(\Omega)), \end{cases}$$

where $\vec{F}_{\vec{U}} = \vec{U}_t - \nu\nabla^2\vec{U} + (\vec{U}\cdot\vec{\nabla})\vec{U}$.

Let $\vec{u} \in L^2((0,T);H^1(\Omega))$ and $p \in L^2((0,T);L^2_0(\Omega))$ denote the state variables, i.e., the velocity and pressure fields, respectively. Let the boundary control $\vec{g}$ belong to $L^2((0,T);H^1_{n0}(\Gamma_c))$ with $\vec{g}_t \in L^2((0,T);L^2(\Gamma_c))$. The state variables are constrained to satisfy the weak form of the Navier–Stokes system (1) for almost all t in $(0,T)$, i.e.,

$$(16)\qquad \begin{cases} \langle \vec{u}_t, \vec{v}\rangle + \nu a(\vec{u},\vec{v}) + c(\vec{u};\vec{u},\vec{v}) + b(\vec{v},p) = 0 \quad \forall\, \vec{v} \in H_0^1(\Omega),\\ b(\vec{u},q) = 0 \quad \forall\, q \in L_0^2(\Omega),\\ (\vec{u},\vec{s})_\Gamma = (\vec{g}(t,\vec{x}),\vec{s})_{\Gamma_c} \quad \forall\, \vec{s} \in H^{-1/2}(\Gamma),\\ \vec{u}(0,\vec{x}) = \vec{u}_0(\vec{x}) \in curl(H^2)(\Omega). \end{cases}$$

More precisely, let $\vec{g} \in H^{1,1}(S_c) \cap L^2((0,T);H^1_{n0}(\Gamma_c))$ and $\vec{u}_0, \in curl(H^2)(\Omega)$. Then, $(\vec{u},p) \in L^2((0,T);H^1(\Omega)) \times L^2((0,T);L^2_0(\Omega))$ is called a *weak solution for the Navier–Stokes equations* if it satisfies (16).

If $\vec{u}$ is a solution of (1), then it is also solution of the weak formulation (16). If $\vec{u}$ is solution of (16), then it satisfies (1) in the sense of distributions on $(0,T)$. If $\vec{g}$, $\vec{u}_0$ are given as above, then we can show that there exists a unique admissible weak solution $(\vec{u},p)$ of (16) such that $\vec{u} \in L^\infty((0,T);W(\Omega)) \cap L^2((0,T);H^1(\Omega))$ and $\vec{u}_t \in L^2((0,T);H^{-1}(\Omega))$; i.e, it is a.e. equal to a continuous function [12].

THEOREM 2.1. *Let $\Omega \subset \mathbb{R}^2$ be an open, bounded domain with boundary Γ of class C^2. Let $\vec{g}(t,\vec{x})$ be a function belonging to $H^{1/2,1}(S)$ satisfying the compatibility conditions*

$$(17)\qquad \int_\Gamma \vec{g}\cdot\vec{n}\, d\Gamma = 0,$$

$$(18)\qquad \vec{g}(0,\vec{x}) = \vec{u}_0|_\Gamma.$$

Then, there exists a unique $\vec{u} \in L^2((0,T);H^1(\Omega)) \cap L^\infty((0,T);L^2(\Omega))$ and $p \in L^2((0,T);L^2_0(\Omega))$ that are the solution of the nonhomogeneous Navier–Stokes problem

$$(19)\qquad \begin{cases} \langle \vec{u}_t, \vec{v}\rangle + \nu\, a(\vec{u},\vec{v}) + c(\vec{u};\vec{u},\vec{v}) + b(\vec{v},p) = 0 \qquad \forall\, \vec{v} \in H_0^1(\Omega),\\ b(\vec{u},q) = 0 \quad \forall\, q \in L_0^2(\Omega),\\ \vec{u} = \vec{g}(t,\vec{x}) \quad \forall\, \vec{x} \in \Gamma,\\ \vec{u}(0,\vec{x}) = \vec{u}_0(\vec{x}) \in curl(H^2)(\Omega) \end{cases}$$

for almost all $t \in (0,T)$. Moreover,

$$(20)\qquad \|\vec{u}\|^2_{L^2((0,T);H^1)} + \|\vec{u}\|^2_{L^\infty((0,T);L^2)} \le K(\|\vec{g}\|^2_{H^{1/2,1}(S)} + \|\vec{u}_0\|^2_{H^{1/2}(\Gamma)}),$$

where K depends on $\vec{g}$.

A proof can be found in [29]. It is worthwhile to recall that that if $\|\vec{g}\|^2_{H^{1/2,1}(S)}$ is uniformly bounded, then also the norms on the left-hand side of (20) are uniformly bounded.

From the previous discussion we can define more precisely the set of admissible solutions which we denote by A_d.

> *Given $T > 0$, $\vec{u}_0 \in curl(H^2)(\Omega)$, and $\vec{U} \in U_{ad}$, then $(\vec{u},p,\vec{g})$ is called an admissible solution for the optimal control problem if $(\vec{u},p,\vec{g}) \in L^2((0,T);H^1(\Omega)) \times L^2((0,T);L^2_0(\Omega)) \times \in H^{1,1}(S_c) \cap L^2(0,T;H^1_{n0}(\Gamma_c))$ is a solution of* (16)*, the control $\vec{g}$ satisfies the compatibility conditions* (17)*–*(18)*, and the functional $\mathcal{J}(\vec{u},\vec{g})$ is bounded.*

The optimal control problem can then be formulated as follows:

given $\vec{u}_0 \in curl(H^2)(\Omega)$ and $\vec{U} \in U_{ad}$, find $(\vec{u}, p, \vec{g}) \in A_d$ such that the control $\vec{g}$ minimizes the cost functional

$$\mathcal{J}(\vec{u}, \vec{g}) = \frac{\alpha}{2} \int_0^T \int_\Omega (\vec{u} - \vec{U})^2 \, d\vec{x}dt + \frac{\beta}{2} \int_0^T \int_{\Gamma_c} (\vec{g}^2 + \beta_1 \vec{g}_x^2 + \beta_2 \vec{g}_t^2) \, d\vec{x}dt \tag{21}$$

with $\alpha, \beta, \beta_1, \beta_2 > 0$.

The first term represents the goal of our optimization and the second term is the penalty term necessary to regularize the solution. The requirement that β must be positive has similarities with the distributed control case; see, e.g., [1] and [22]. The requirement that β_1, β_2 should be different from zero is necessary if we want $\vec{g} \in H^{1,1}(S_c)$.

2.2. Existence of an optimal solution. In this section, we prove that the optimal control problem (21) is well posed and has at least one solution.

THEOREM 2.2. *Given $T > 0$ and $\vec{u}_0 \in curl(H^2)(\Omega)$, then there exists a solution $(\vec{u}, p, \vec{g}) \in A_d$ of the optimal control problem* (21).

Proof. We consider the following equivalent problem. Let $\vec{g} \in H^{1,1}(S_c) \cap L^2(0, T; H^1_{n0}(\Gamma_c))$ and $(\widetilde{u}, \widetilde{p})$ satisfy the linear Stokes equation

$$\begin{cases} \langle \widetilde{u}_t, \vec{v} \rangle + \nu a(\widetilde{u}, \vec{v}) + b(\vec{v}, \widetilde{p}) = 0 \qquad \forall \, \vec{v} \in H^1_0(\Omega), \\ b(\widetilde{u}, q) = 0 \quad \forall \, q \in L^2_0(\Omega), \\ (\widetilde{u}, \vec{s})_\Gamma = (\vec{g}(t, \vec{x}), \vec{s})_{\Gamma_c} \quad \forall \, \vec{s} \in H^{-1/2}(\Gamma), \\ \widetilde{u}(0, \vec{x}) = u_0 \in curl(H^2)(\Omega). \end{cases} \tag{22}$$

The problem is now to find $\widetilde{u}$ and a solution $(\widehat{u}, \widehat{p})$ of the system

$$\begin{cases} \langle \widehat{u}_t, \vec{v} \rangle + \nu a(\widehat{u}, \vec{v}) + c(\widehat{u}; \widehat{u}, \vec{v}) + c(\widetilde{u}; \widetilde{u}, \vec{v}) \\ + c(\widehat{u}; \widetilde{u}, \vec{v}) + c(\widetilde{u}; \widehat{u}, \vec{v}) + b(\vec{v}, \widehat{p}) = 0 \qquad \forall \, \vec{v} \in H^1_0(\Omega), \\ \widehat{u} \in V(\Omega), \end{cases} \tag{23}$$

such that the control $\vec{g}$ minimizes the cost functional

$$\mathcal{J}(\widetilde{u} + \widehat{u}, \vec{g}) = \frac{\alpha}{2} \int_0^T \int_\Omega (\widetilde{u} + \widehat{u} - \vec{U})^2 \, d\vec{x}dt + \frac{\beta}{2} \int_0^T \int_{\Gamma_c} (\vec{g}^2 + \beta_1 \vec{g}_t^2 + \beta_2 \vec{g}_x^2) \, d\vec{x}dt. \tag{24}$$

An admissible solution can be found by solving the above systems with boundary conditions $\vec{g} = \gamma \vec{u}_0$ $\forall t \in [0, T]$. Since the admissible set is not empty and the set of values assumed by the functional is bounded from below, let $\vec{g}_n$ be a minimizing sequence for the problem in (22)–(24) and set $\widetilde{u}_n = \widetilde{u}(\vec{g}_n)$ and $\widehat{u}_n = \widehat{u}(\widetilde{u}_n, \vec{g}_n)$. In the rest of the proof, we denote g_t by g'. The sequences $\{\vec{g}_n\}$ and $\{\vec{g}'_n\}$ are uniformly bounded in $L^2((0, T); H^1_{n0}(\Gamma_c))$ and $L^2((0, T); L^2(\Gamma_c))$, respectively; in fact we can easily choose the sequences such that $\mathcal{J}(\vec{u}_n, \vec{g}_n) \leq M$ $\forall n$, where M is the value of the functional for an admissible solution. The boundary velocity $\widetilde{u}$ is thus uniformly bounded in $L^2((0, T); H^{1/2}_0(\Gamma_c))$ and the corresponding solutions $\widetilde{u}_n$ and $\widehat{u}_n$ are uniformly bounded in the set $L^\infty((0, T); L^2(\Omega)) \cap L^2((0, T); H^1(\Omega))$. Hence, there is a $(\widetilde{u}, \widehat{u}, \vec{g})$ and a subsequence of $(\widetilde{u}_m, \widehat{u}_m, \vec{g}_m)$ that converges weakly to $(\widetilde{u}, \widehat{u}, \vec{g})$. We write

$$\begin{array}{llll} \vec{g}_m & \to \vec{g} & \text{in} & L^2((0,T);H^1_{n0}(\Gamma_c)) & \text{weakly,} \\ \vec{g}'_m & \to \vec{g}' & \text{in} & L^2((0,T);L^2(\Gamma_c)) & \text{weakly,} \\ \widetilde{u}_m & \to \widetilde{u} & \text{in} & L^2((0,T);H^1(\Omega)) & \text{weakly,} \\ \widehat{u}_m & \to \widehat{u} & \text{in} & L^2((0,T);V(\Omega)) & \text{weakly,} \\ \widehat{u}_m & \to \widehat{u} & \text{in} & L^\infty((0,T);W(\Omega)) & *\text{-weakly.} \end{array}$$

Now, $(\widetilde{u},\widehat{u},\vec{g})$ satisfies the system of (22)–(23) and minimizes the functional. In fact, by the lower semicontinuity of the functional (24), we have

$$\mathcal{J}(\widetilde{u}+\widehat{u},\vec{g}) \le \liminf_{m\to\infty} \mathcal{J}(\widetilde{u}_m+\widehat{u}_m,\vec{g}_m).$$

Let $\vec{w}$ be in $\mathcal{V}(\Omega)$ and $\psi(t)$ be a continuously differentiable function on $[0,T]$ with $\psi(T)=0$. We multiply (22) and (23) by $\psi(\tau)\vec{w}$ and then integrate by parts in τ to obtain

$$\begin{aligned} -\int_0^T (\widehat{u}_m,\psi'(\tau)\vec{w})\,d\tau + \nu\int_0^T a(\widehat{u}_m,\psi(\tau)\vec{w})\,d\tau + \int_0^T c(\widehat{u}_m;\widehat{u}_m,\psi(\tau)\vec{w})\,d\tau \\ = \int_0^T (\widehat{f}_m,\psi(\tau)\vec{w})\,d\tau - \int_0^T (\widetilde{u}_m,\psi'(\tau)\vec{w})\,d\tau \\ +\nu\int_0^T a(\widetilde{u}_m,\psi(\tau)\vec{w})\,d\tau = (\vec{u}_0,\psi(0)\vec{w}). \end{aligned}$$

We can pass to the limit inside the linear and the nonlinear terms. In fact, the a priori estimate (see [11] or [32]) for $\widehat{u}$ in a fractional time order Sobolev space yields that $\widehat{u}_m$ converges strongly to $\widehat{u} \in L^2((0,T);V(\Omega))$. If $\psi \in \mathcal{D}((0,T))$, the limit $(\widehat{u},\widetilde{u},\vec{g})$ satisfies the Navier–Stokes equation (22) in the sense of distributions. Since $\mathcal{V}(\Omega)$ is dense in $H^1_0(\Omega)$, this is still true for any $\vec{w} \in H^1_0(\Omega)$ by a continuity argument. □

2.3. First-order necessary conditions. In this section, we derive the first-order necessary conditions. Let G be the set of all $\vec{g} \in H^{1,1}(S) \cap L^2(0,T;H^1_{n0}(\Gamma))$ satisfying the compatibility conditions in (17)–(18). For all $\vec{g} \in G$, the first-order necessary condition is available if the map

$$\vec{u}(\vec{g}) : G \to L^2((0,T);H^1(\Omega))$$

is Gâteaux differentiable. In the following theorem, we state and prove the existence of the Gâteaux derivative for directions $\widetilde{h}$ in $H^{1,1}(S) \cap L^2(0,T;H^1_{n0}(\Gamma))$.

THEOREM 2.3. *Given $\Omega \in C^2$, $\vec{u}_0 \in curl(H^2)$, and $\vec{g} \in G$, the mapping*

$$\vec{u}(\vec{g}) : G \to L^2((0,T);H^1(\Omega))$$

has a Gâteaux derivative $\frac{D\vec{u}}{D\vec{g}} \cdot \widetilde{h}$ in every direction $\widetilde{h} \in H^{1,1}(S) \cap L^2(0,T;H^1_{n0}(\Gamma))$ with $\widetilde{h} = 0$ at $t=0$. Furthermore, $\widetilde{w}(h) = \frac{D\vec{u}}{D\vec{g}} \cdot \widetilde{h}$ is the solution of the problem

$$(25)\quad \begin{cases} \langle \widetilde{w}_t,\vec{v}\rangle + \nu\, a(\widetilde{w},\vec{v}) + c(\vec{u};\widetilde{w},\vec{v}) + c(\vec{w};\widetilde{u},\vec{v}) + b(\vec{v},p) = 0 \quad \forall\, \vec{v} \in H^1_0(\Omega), \\ b(\widetilde{w},q) = 0 \quad \forall\, q \in L^2_0(\Omega), \\ (\widetilde{w}(t,\vec{x}),\vec{s}\,) = (\widetilde{h}(t,\vec{x}),\vec{s}\,) \quad \forall\, \vec{s} \in H^{-1/2}(\Gamma), \\ \widetilde{w}(0,\vec{x}) = 0, \quad \vec{x} \in \Omega, \end{cases}$$

where $\widetilde{w} \in L^\infty((0,T);L^2(\Omega)) \cap L^2((0,T);H^1(\Omega))$.

Proof. Let $\vec{g}$ and $\widetilde{h}$ be given in $H^{1,1}(S) \cap L^2(0,T;H^1_{n0}(\Gamma))$. We need to prove the following result:

$$\lim_{s\to 0}\left(\frac{\|(\vec{u}_{\vec{g}+s\widetilde{h}} - \vec{u}_{\vec{g}}) - s\widetilde{w}(\widetilde{h})\|_{L^2((0,T);H^1)}}{|s|}\right) = 0. \tag{26}$$

We set $\widetilde{u} = (\vec{u}_{\vec{g}+s\widetilde{h}} - \vec{u}_{\vec{g}}) - s\widetilde{w}(\widetilde{h})$ so that $\widetilde{u}$ is the solution of the evolution equation

$$\begin{cases} \dfrac{d\widetilde{u}}{dt} + \nu(\pi A)\widetilde{u} + (\pi C)(\vec{u}_{\vec{g}+s\widetilde{h}})\vec{u}_{\vec{g}+s\widetilde{h}} - (\pi C)(\vec{u}_{\vec{g}})\vec{u}_{\vec{g}} - (\pi C)'(\vec{u}_{\vec{g}})s\widetilde{w} = 0, \\ \widetilde{u} \in V(\Omega), \\ \widetilde{u}(0,\vec{x}) = 0, \quad \vec{x} \in \Omega. \end{cases} \tag{27}$$

If we define the function $\vec{k} \in L^2((0,T);H^{-1}(\Omega))$, as follows,

$$\vec{k} = (\pi C)(\vec{u}_{\vec{g}+s\widetilde{h}})\vec{u}_{\vec{g}+s\widetilde{h}} - (\pi C)(\vec{u}_{\vec{g}})\vec{u}_{\vec{g}} - (\pi C)'(\vec{u}_{\vec{g}})(\vec{u}_{\vec{g}+s\widetilde{h}} - \vec{u}_{\vec{g}}),$$

then (27) becomes

$$\begin{cases} \dfrac{d\widetilde{u}}{dt} + \nu(\pi A)\widetilde{u} + (\pi C)'(\vec{u}_{\vec{g}})\widetilde{u} = \vec{k}, \\ \widetilde{u} \in V(\Omega), \\ \widetilde{u}(t,\vec{x}) = 0, \quad \vec{x} \in \Gamma, \quad t \in (0,T), \\ \widetilde{u}(0,\vec{x}) = 0, \quad \vec{x} \in \Omega. \end{cases} \tag{28}$$

In order to estimate $\|\widetilde{u}\|_1$ we recall the following result, which can be easily found by standard techniques [29]. If $\vec{w}$ is the solution of

$$\begin{cases} \vec{w}_t + \nu(\pi A)\vec{w} + \delta[(\pi C)(\vec{w})\vec{u} + (\pi C)(\vec{u})\vec{w}] + \sigma(\pi C)(\vec{w})\vec{w} = \vec{f}, \\ \vec{w} \in V(\Omega), \end{cases} \tag{29}$$

with initial value $\vec{w}(0,\vec{x}) = 0$ and homogeneous boundary condition, then the solution $\vec{w}$ for all nonnegative real values of δ and σ has the following property: if $\vec{f} \in L^2((0,T);H^{-1}(\Omega))$ and $\vec{u} \in L^\infty((0,T);H^1(\Omega)) \cap L^2((0,T);V(\Omega))$, then the solution $\vec{w} \in L^\infty((0,T);W(\Omega)) \cap L^2((0,T);V(\Omega))$ and

$$\|\vec{w}\|_{L^\infty((0,T);L^2)} \le C_1\|\vec{f}\|_{L^2((0,T);H^{-1})}, \tag{30}$$

$$\|\vec{w}\|_{L^2((0,T);H^1)} \le C_2\|\vec{f}\|_{L^2((0,T);H^{-1})}, \tag{31}$$

where C_1, C_2, given $\vec{u}$, δ, and σ, are constants depending only on Ω and ν.

Using $(\pi C)'(\vec{u}) \cdot \widetilde{u} = (\pi C)(\vec{u})\widetilde{u} + (\pi C)(\widetilde{u})\vec{u}$ and (31) with $\sigma = 0, \delta = 1, \vec{f} = \vec{k} \in L^2((0,T);H^{-1}(\Omega))$, we obtain

$$\int_0^T \|\widetilde{u}\|_1^2 \, d\tau \le C_2 \int_0^T \|\vec{k}\|_{H^{-1}}^2 \, d\tau.$$

Now we need to evaluate the right-hand side term above. From the definition of the norm in $H^{-1}(\Omega)$, we have

$$\|\vec{k}\|_{H^{-1}} = \sup_{\|\vec{v}\|_{H_0^1(\Omega)}\leq 1} \frac{|\langle \vec{k}, \vec{v}\rangle|}{\|\vec{v}\|_1}.$$

The evaluation of the duality pairing on $H^{-1} \times H_0^1$ yields

$$\begin{aligned} |\langle \vec{k}, \vec{v}\rangle| &= |c(\vec{u}_{\vec{g}+s\widetilde{h}}; \vec{u}_{\vec{g}+s\widetilde{h}}, \vec{v}) - c(\vec{u}_{\vec{g}}; \vec{u}_{\vec{g}}, \vec{v}) - c'(\vec{u}_{\vec{g}}; \widehat{u}, \vec{v})| \\ &= |c(\vec{u}_{\vec{g}+s\widetilde{h}}; \vec{u}_{\vec{g}+s\widetilde{h}}, \vec{v}) - c(\vec{u}_{\vec{g}}; \vec{u}_{\vec{g}}, \vec{v}) - c(\vec{u}_{\vec{g}}; \widehat{u}, \vec{v}) - c(\widehat{u}; \vec{u}_{\vec{g}}, \vec{v})| \\ &= |c(\widehat{u}, \vec{u}_{\vec{g}+s\widetilde{h}}, \vec{v}) - c(\widehat{u}; \vec{u}_{\vec{g}}, \vec{v})| = |c(\widehat{u}; \widehat{u}, \vec{v})| \leq K\|\nabla \widehat{u}\|\|\widehat{u}\|\|\nabla \vec{v}\|, \end{aligned} \tag{32}$$

where $\widehat{u} = \vec{u}_{\vec{g}+s\widetilde{h}} - \vec{u}_{\vec{g}}$. Hence, the estimate for $\widetilde{u}$ yields

$$\int_0^T \|\widetilde{u}(t)\|_1^2\, dt \leq KC_2 \int_0^T \|\widehat{u}\|^2\|\widehat{u}\|_1^2\, dt, \tag{33}$$

where $\widehat{u} = \vec{u}_{\vec{g}+s\widetilde{h}} - \vec{u}_{\vec{g}}$ is the solution of the system

$$\begin{cases} \langle \widehat{u}_t, \vec{v}\rangle + \nu a(\widehat{u}, \vec{v}) + c(\vec{u}_{\vec{g}}; \widehat{u}, \vec{v}) + c(\widehat{u}; \vec{u}_{\vec{g}}, \vec{v}) \\ \quad + c(\widehat{u}; \widehat{u}, \vec{v}) + b(\vec{v}, p) = 0 \quad \forall\, \vec{v} \in H^1(\Omega), \\ b(\widehat{u}, q) = 0 \quad \forall\, q \in L_0^2(\Omega), \\ (\widehat{u}(t, \vec{x}), \vec{r}) = (s\widetilde{h}, \vec{r}) \quad \forall\, \vec{r} \in H^{-1/2}(\Gamma), \\ \widehat{u}(0, \vec{x}) = 0, \quad \vec{x} \in \Omega. \end{cases} \tag{34}$$

Again, in order to estimate the norm of $\widehat{u}$ in $L^\infty((0,T); L^2(\Omega))$ and in $L^2((0,T); H^1(\Omega))$, we set $\widehat{u}_2 = \widehat{u} - \widehat{u}_1$ and decompose the nonhomogeneous Navier–Stokes equation into two systems: a linear system defined by the Stokes problem

$$\begin{cases} \langle \widehat{u}_{1t}, \vec{v}\rangle + \nu a(\widehat{u}_1, \vec{v}) + b(\vec{v}, p) = 0 \qquad \forall\, \vec{v} \in H_0^1(\Omega), \\ b(\widehat{u}_1, q) = 0 \quad \forall\, q \in L_0^2(\Omega), \\ (\widehat{u}_1, \vec{r}) = (s\widetilde{h}, \vec{r}) \quad \forall\, \vec{r} \in H^{-1/2}(\Gamma), \\ \widehat{u}_1(0, \vec{x}) = u_0 \in curl(H^2)(\Omega) \end{cases} \tag{35}$$

and the homogeneous Navier–Stokes system

$$\begin{cases} \langle \widehat{u}_{2t}, \vec{v}\rangle + \nu a(\widehat{u}_2, \vec{v}) + c(\vec{u}_{\vec{g}} + \widehat{u}_1; \widehat{u}_2, \vec{v}) + c(\widehat{u}_2; \vec{u}_{\vec{g}} + \widehat{u}_1, \vec{v}) \\ + c(\widehat{u}_2; \widehat{u}_2, \vec{v}) = (\widehat{f}, \vec{v}) \qquad \forall\, \vec{v} \in V(\Omega), \\ \widehat{u}_2 \in V(\Omega), \end{cases} \tag{36}$$

where $\widehat{f} = -c(\widehat{u}_1; \widehat{u}_1, \vec{v})$. From the Stokes system, we have $\widehat{u}_1 \in L^\infty((0,T); L^2(\Omega))$ with $\|\widehat{u}_1\|_{L^\infty((0,T);L^2(\Omega))} \leq sC_3\|\widetilde{h}\|_{H^{1,1}}$ and

$$\int_0^T \|\widehat{u}_1\|_1^2\, dt \leq C_4|s|^2\|\widetilde{h}\|^2_{H^{1,1}(S)}$$

so that $\|\widehat{f}\|_{L^2((0,T);H^{-1})} \le C_5|s|^2\|\widetilde{h}\|^2_{H^{1,1}(S)}$. Also, by using (31) ($\delta = 1$, $\sigma = 1$) to estimate the solution in (36), we have $\|\widehat{u}_2\|_{L^\infty((0,T);L^2(\Omega))} \le sC_6\|\widetilde{h}\|_{H^{1,1}}$ and

$$\int_0^T \|\widehat{u}_2\|_1^2\,dt \le C_7|s|^2\|\widetilde{h}\|^2_{H^{1,1}}.$$

Further details on the above estimates can be found in [29]. Combining the above inequalities we obtain the estimates

$$\|\widehat{u}\|_{L^\infty((0,T);L^2(\Omega))} \le C_8 s\|\widetilde{h}\|_{H^{1,1}}, \tag{37}$$

$$\int_0^T \|\widehat{u}(t)\|_1^2\,dt \le C_9|s|^2\|\widetilde{h}\|^2_{H^{1,1}(S)}.$$

If we use (37) in (33), (26) is satisfied. From the regularity of $\widetilde{h}$, it follows that $\widetilde{w} \in L^\infty((0,T);L^2(\Omega)) \cap L^2((0,T);H^1(\Omega))$. □

The canonical extension $\widetilde{h}_c \to \widetilde{h}$ from $H^1_{n0}(\Gamma_c)$ to $H^1(\Gamma)$, where

$$\widetilde{h} = \begin{cases} \widetilde{h}_c, & \vec{x} \in \Gamma_c, \\ 0, & \vec{x} \in \Gamma \setminus \Gamma_c \end{cases}$$

is a continuous mapping; see [12]. This allows us to take variations in subdomains of the boundary, i.e., $H^1(\Gamma_c)$, and to claim the existence of the Gâteaux derivative for such configurations. For a variation $\widetilde{h}_c \in H^{1,1}(S_c) \cap L^2(0,T;H^1_{n0}(\Gamma_c))$, i.e., $\widetilde{h}_c \in L^2((0,T);H^1_{n0}(\Gamma_c))$ and $\widetilde{h}_{ct} \in L^2((0,T);L^2(\Gamma_c))$, of the control $\vec{g}$, the Gâteaux derivative of the Navier–Stokes system can be written in this following form:

$$\begin{cases} \langle \widetilde{w}_t, \vec{v}\rangle + \nu a(\widetilde{w},\vec{v}) + c(\widetilde{w};\vec{u},\vec{v}) + c(\vec{u};\widetilde{w},\vec{v}) + b(\vec{v},p_1) = 0 \quad \forall\, \vec{v} \in H^1_0(\Omega), \\ b(\widetilde{w},q) = 0 \quad \forall\, q \in L^2_0(\Omega), \\ (\widetilde{w},\vec{s})_\Gamma = (\widetilde{h}_c(t,\vec{x}),\vec{s})_{\Gamma_c} \quad \forall\, \vec{s} \in H^{-1/2}(\Gamma), \\ \widetilde{w}(0,\vec{x}) = 0, \quad \vec{x} \in \Omega. \end{cases} \tag{38}$$

Now, we can show that that the optimal solution must satisfy a first-order necessary condition. If $(\vec{u},\vec{g})$ is an optimal pair, then for every $\widetilde{h} \in H^{1,1}(S_c) \cap L^2(0,T;H^1_{n0}(\Gamma_c))$ and for every $\lambda \in \mathbb{R}$ we have, from the definition of an optimal solution,

$$\mathcal{J}(\vec{g} + \lambda\widetilde{h}) \ge \mathcal{J}(\vec{g}).$$

The above inequality implies

$$\frac{\mathcal{J}(\vec{g}+\lambda\widetilde{h}) - \mathcal{J}(\vec{g})}{\lambda} \ge 0 \quad \text{if} \quad \lambda \ge 0 \quad \text{and} \quad \frac{\mathcal{J}(\vec{g}+\lambda\widetilde{h}) - \mathcal{J}(\vec{g})}{\lambda} \le 0 \quad \text{if} \quad \lambda \le 0.$$

The limit must vanish when λ tends to zero and this leads to the following first-order necessary condition.

THEOREM 2.4. *If $(\vec{u},p,\vec{g})$ is an optimal pair for the problem in* (21), *then the Gâteaux derivative of $\mathcal{J}(\cdot,\cdot)$ vanishes at $(\vec{u},p,\vec{g})$.*

We would like to write the first-order necessary condition in a more explicit form. In order to do this we need this interesting preliminary result.

LEMMA 2.5. *Given $\Omega \in C^2$ and $\vec{u}_0 \in curl(H^2)(\Omega)$. Let $\widetilde{h}_c$ be given in $H^{1,1}(S_c) \cap L^2(0,T;H^1_{n0}(\Gamma_c))$ and let $\widetilde{w}(\widetilde{h}_c)$ be defined by* (38). *Then, for every $\widetilde{h}_2$ belonging to $L^2((0,T);H^1(\Omega))$, we have*

$$\int_0^T \int_\Omega \widetilde{h}_2 \widetilde{w}(\widetilde{h}_c)\, d\vec{x}dt = -\int_0^T \int_{\Gamma_c} \vec{\xi} \cdot \vec{w}\, d\vec{x}dt,$$

where $\vec{w}$ is the solution of the adjoint linearized problem

$$(39) \qquad \begin{cases} -(\vec{w}_t, \vec{v}) + \nu a(\vec{w}, \vec{v}) + c(\vec{v}; \vec{u}, \vec{w}) + c(\vec{u}; \vec{v}, \vec{w}) \\ \quad + b(\vec{v}, \sigma) = (\widetilde{h}_2, \vec{v}) \quad \forall\, \vec{v} \in H^1_0(\Omega), \\ b(\vec{w}, q) = 0 \quad \forall\, q \in L^2_0(\Omega), \\ \vec{w} = 0 \quad \forall\, \vec{x} \in \Gamma, \\ \vec{w}(T, \vec{x}) = 0 \quad \forall\, \vec{x} \in \Omega. \end{cases}$$

The function $\vec{\xi} = (\nu\gamma_1\vec{w} - \sigma\vec{n}) \in L^2((0,T);H^{-1/2}(\Gamma_c))$ is defined by

$$(40) \qquad \int_{\Gamma_c} \vec{\xi} \cdot \vec{v}\, d\vec{x} = -(\vec{w}_t, \vec{v}) + \nu a(\vec{w}, \vec{v}) + c(\vec{v}; \vec{u}, \vec{w}) + c(\vec{u}; \vec{v}, \vec{w}) \\ + b(\vec{v}, \sigma) - (\widetilde{h}_2, \vec{v}) \quad \forall\, \vec{v} \in H^1(\Omega).$$

Proof. We remark that $\widetilde{h}_2 \in L^2((0,T);H^1(\Omega))$; then, (39) has a solution $\vec{w} \in L^\infty((0,T);W(\Omega)) \cap L^2((0,T);V(\Omega))$ and $\sigma \in L^2((0,T);L^2_0(\Omega))$. The definition in (40) makes sense since the weak formulation of the adjoint equation tested against $H^1(\Omega)$ has solution for all $\widetilde{h}_2 \in (H^1)^*(\Omega)$; one can use the same techniques as in [21]. Hence, there exists a $\vec{\xi} \in H^{-1/2}(\Gamma_c)$, defined by (40). We need to evaluate the integral over time of $(\widetilde{w}, \widetilde{h}_2)$. The integral contains $\widetilde{h}_2$, for which we can use (40). We set $\vec{v} = \widetilde{w}$ in that equation and then we integrate by parts with respect to the time variable. We then obtain

$$\int_0^T (\widetilde{h}_2, \widetilde{w})\, dt = \int_0^T [-(\vec{w}_t, \widetilde{w}) + \nu a(\vec{w}, \widetilde{w}) c(\widetilde{w}; \vec{u}; \vec{w}) \\ + c(\vec{u}; \widetilde{w}, \vec{w}) + b(\widetilde{w}, \sigma)]\, dt - \int_0^T \int_{\Gamma_c} \vec{\xi} \cdot \widetilde{h}_c\, d\vec{x}dt \\ = \int_0^T [(\widetilde{w}_t, \vec{w}) + \nu a(\widetilde{w}, \vec{w}) + c(\widetilde{w}; \vec{u}; \vec{w}) + c(\vec{u}; \widetilde{w}, \vec{w})]\, dt - \int_0^T \int_{\Gamma_c} \vec{\xi} \cdot \widetilde{h}_c\, d\vec{x}dt.$$

The result then follows from the fact that the first term vanishes; it satisfies (39), the weak equation for the Gâteaux derivative, with $\vec{v} = \vec{w}$. □

In the next theorem we shall show that if the Gâteaux derivative vanishes, then $\vec{g}$ must be a solution of a differential equation.

THEOREM 2.6. *If $(\vec{u}, \vec{g})$ is an optimal pair for the problem in* (21), *then $\vec{g} \in H^{1,1}(S_c) \cap L^2(0,T;H^1_{n0}(\Gamma_c))$ with $\vec{g}(0, \vec{x}) = \gamma\vec{u}_0$ is solution of*

$$(41) \qquad \int_0^T \int_{\Gamma_c} \left[\vec{g} \cdot \widetilde{h} + \beta_1 \vec{g}_t \cdot \widetilde{h}_t + \beta_2 \partial_s \vec{g} \cdot \partial_s \widetilde{h} - \frac{1}{\beta}(\vec{\xi} \cdot \widetilde{h}) \right] d\vec{x}dt = 0$$

$\forall\, \widetilde{h} \in H^{1,1}(S_c) \cap L^2(0,T; H^1_{n0}(\Gamma_c))$ with $\widetilde{h} = 0$ at $t = 0$. The function $\vec{w} \in L^\infty((0,T); L^2(\Omega)) \cap L^2((0,T); V(\Omega))$ and is the solution of the adjoint linearized problem

$$(42)\qquad \begin{cases} -(\vec{w}_t, \vec{v}) + \nu\, a(\vec{w}, \vec{v}) + c(\vec{u}; \vec{v}, \vec{w}) + c(\vec{v}; \vec{u}, \vec{w}) \\ \quad + b(\vec{v}, \sigma) = \alpha(\vec{u} - \vec{U}, \vec{v}) \quad \forall\, \vec{w} \in H^1_0(\Omega), \\ b(\vec{w}, q) = 0 \quad \forall\, q \in L^2_0(\Omega), \\ \vec{w} = 0 \quad \forall\, \vec{x} \in \Gamma, \\ \vec{w}(T, \vec{x}) = 0 \quad \forall\, \vec{x} \in \Omega \end{cases}$$

and $\vec{\xi} = (\nu\gamma_1\vec{w} - \sigma\vec{n})$ on Γ_c is defined by

$$(43)\qquad \begin{aligned} \int_{\Gamma_c} \vec{\xi} \cdot \vec{v}\, d\vec{x} &= \int_{\Gamma_c} (\nu\gamma_1\vec{w} \cdot \gamma_0\vec{v} - \sigma\gamma_0\vec{v} \cdot \vec{n})\, d\vec{x} \\ &= -(\vec{w}_t, \vec{v}) + \nu\, a(\vec{w}, \vec{v}) + c(\vec{u}; \vec{v}, \vec{w}) \\ &\qquad + c(\vec{v}; \vec{u}, \vec{w}) + b(\vec{v}, \sigma) - \alpha(\vec{u} - \vec{U}, \vec{v}) \quad \forall\, \vec{v} \in H^1(\Omega). \end{aligned}$$

Proof. Let $(\vec{u}, \vec{g})$ be an optimal solution of the problem (21). We compute the Gâteaux derivative of the functional $\mathcal{J}(\vec{u}(\vec{g}), \vec{g})$ in the direction of $\widetilde{h}$ and then Lemma 2.5 completes the proof. We have

$$\begin{aligned} \frac{D\mathcal{J}(\vec{u}, \vec{g})}{D\vec{g}} \cdot \widetilde{h} &= \alpha \int_0^T \int_\Omega (\vec{u} - \vec{U}) \cdot \left(\frac{D\vec{u}}{D\vec{g}} \cdot \widetilde{h} \right) d\vec{x}dt \\ &\quad + \beta \int_0^T \int_{\Gamma_c} [\vec{g} \cdot \widetilde{h} + \beta_1 \vec{g}_t \cdot \widetilde{h}_t + \beta_2 \partial_s \vec{g} \cdot \partial_s \widetilde{h}]\, d\vec{x}dt. \end{aligned}$$

Now, by using Lemma 2.5, we can integrate by parts to obtain

$$\begin{aligned} \frac{D\mathcal{J}(\vec{u}, \vec{g})}{D\vec{g}} \cdot \widetilde{h} &= \alpha \int_0^T \int_\Omega (\vec{u} - \vec{U}) \cdot \vec{w}\, d\vec{x}dt \\ &\quad + \beta \int_0^T \int_{\Gamma_c} [\vec{g} \cdot \widetilde{h} + \beta_1 \vec{g}_t \cdot \widetilde{h}_t + \beta_2 \partial_s \vec{g} \cdot \partial_s \widetilde{h}]\, d\vec{x}dt \\ &= \int_0^T \int_{\Gamma_c} [\beta(\vec{g} \cdot \widetilde{h} + \beta_1 \vec{g}_t \cdot \widetilde{h}_t + \beta_2 \partial_s \vec{g} \cdot \partial_s \widetilde{h}) - (\vec{\xi} \cdot \widetilde{h})]\, d\vec{x}dt, \end{aligned}$$

where $\vec{w}$ is the solution of (42). Now, from Theorem 2.4, if $(\vec{u}, \vec{g})$ is a solution of the optimal control problem, the Gâteaux derivative must vanish. The regularity of $\vec{g}$ follows from the regularity properties shown for $\gamma_1\vec{w}$. □

Equation (41) provides the solution for the boundary control. Since $\widetilde{h} \in H^{1,1}(S_c) \cap L^2(0,T; H^1_{n0}(\Gamma_c))$ with $\widetilde{h}(0, \vec{x}) = 0$, we can take $\widetilde{h} = \psi(t)\vec{r}(\vec{x})$, where $\psi \in \mathcal{D}((0,T))$ with $\psi(0) = 0$ and $\vec{r}(\vec{x})$ in $H^1_{n0}(\Gamma_c)$. After integration by parts, we have

$$(44)\quad \beta_1(\vec{g}_t(T), \vec{r})\psi(T) + \int_0^T \psi(t) \left[(\vec{g}, \vec{r}) - \beta_1(\vec{g}_{tt}, \vec{r}) + \beta_2(\partial_s \vec{g}, \partial_s \vec{r}) - \frac{1}{\beta}(\vec{\xi}, \vec{r}) \right] dt = 0$$

$\forall\, \psi \in \mathcal{D}((0,T))$ with $\psi(0) = 0$ and $\forall \vec{r} \in H^1_{n0}(\Gamma_c)$. In this way, a necessary condition to satisfy (41) is to satisfy the differential equation

$$(45)\qquad (\vec{g}, \vec{r}) - \beta_1(\vec{g}_{tt}, \vec{r}) + \beta_2(\partial_s \vec{g}, \partial_s \vec{r}) - \frac{1}{\beta}(\vec{\xi}, \vec{r}) = 0 \quad \forall\, \vec{r} \in H^1_{n0}(\Gamma_c)$$

with $\vec{g}(0,\vec{x}) = \gamma_0\vec{u}_0(\vec{x})$ and $\vec{g}_t(T,\vec{x}) = 0$. The first of these boundary conditions is imposed on candidate minimizers in order to ensure the regularity of solutions; see (18). The second boundary condition is a result of the minimization process. Now we can use the space $H^1_n(\Gamma_c)$ for the test functions. Because of the orthogonality between $H^1_n(\Gamma_c)$ and $(H^1_n)^\perp(\Gamma_c)$, we can write a weak formulation of (45) with test functions in $H^1(\Gamma_c)$ by adding an arbitrary constant vector in the normal direction. We recall that $H^1(\Gamma_c) = H^1_n(\Gamma_c) \oplus (H^1_n(\Gamma_c))^\perp$ and thus

$$\vec{r} = \vec{r}_1 - \vec{n}\,\frac{\displaystyle\int_{\Gamma_c} \vec{r}_1 \cdot \vec{n}\, d\vec{x}}{\mu(\Gamma_c)},$$

where $\vec{r}_1 \in H^1(\Gamma_c)$. Now, the equation can be tested against $\vec{r}_1 \in H^1_0(\Gamma_c)$ in the following weak form:

$$(46)\quad (\vec{g},\vec{r}_1) - \beta_1(\vec{g}_{tt},\vec{r}_1) + \beta_2(\partial_s\vec{g},\partial_s\vec{r}_1) + k(t)(\vec{n},\vec{r}_1) = \frac{1}{\beta}(\vec{\xi},\vec{r}_1) \quad \forall\,\vec{r}_1 \in H^1_0(\Gamma_c),$$

where $k(t)$ is specified by the constraint

$$\int_\Gamma \vec{g}\cdot\vec{n}d\vec{x} = 0.$$

Finally, in order to obtain the solution of the optimal control problem, we have to solve the Navier–Stokes system

$$(47)\qquad \begin{cases} \langle \vec{u}_t,\vec{v}\rangle + \nu a(\vec{u},\vec{v}) + c(\vec{u};\vec{u},\vec{v}) + b(\vec{v},p) = 0 \quad \forall\,\vec{v}\in H^1_0(\Omega),\\ b(\vec{u},q) = 0 \quad \forall\,q \in L^2_0(\Omega),\\ (\vec{u},\vec{s})_\Gamma = (g(t,\vec{x}),\vec{s})_\Gamma \quad \forall\,\vec{s}\in H^{-1/2}(\Gamma),\\ \vec{u}(0,\vec{x}) = \vec{u}_0(\vec{x}) \in curl(H^2)(\Omega); \end{cases}$$

the adjoint system

$$(48)\qquad \begin{cases} -\langle \vec{w}_t,\vec{v}\rangle + \nu a(\vec{w},\vec{v}) + c(\vec{w};\vec{u},\vec{v}) + c(\vec{u};\vec{w},\vec{v})\\ \quad + b(\vec{v},\sigma) = \alpha(\vec{u}-\vec{U},\vec{v}) \quad \forall\,\vec{v}\in H^1_0(\Omega),\\ b(\vec{w},q) = 0 \quad \forall\,q\in L^2_0(\Omega),\\ \vec{w} = 0 \quad \forall\,\vec{x}\in\Gamma,\\ \vec{w}(T,\vec{x}) = 0 \quad \forall\,\vec{x}\in\Omega; \end{cases}$$

the boundary control equation

$$(49)\qquad \begin{cases} (\vec{g},\vec{r}) - \beta_1(\vec{g}_{tt},\vec{r}) + \beta_2(\partial_s\vec{g},\partial_s\vec{r}) + k(t)(\vec{n},\vec{r})\\ \quad = \dfrac{1}{\beta}[(\gamma_1\vec{w},\vec{r}) - (\vec{n}\sigma,\vec{r})] \quad \forall\,\vec{r}\in H^1_0(\Gamma_c),\\ \vec{g}(0,\vec{x}) = \gamma_0\vec{u}_0 \quad \forall\,\vec{x}\in\Gamma_c,\\ \vec{g}_t(T,\vec{x}) = 0 \quad \forall\,\vec{x}\in\Gamma_c,\\ \vec{g} = 0 \quad \forall\,\vec{x}\in\partial\Gamma_c; \end{cases}$$

and the compatibility condition for the boundary control

$$\int_\Gamma \vec{g} \cdot \vec{n} \, d\Gamma = 0. \tag{50}$$

Equation (50) is needed in order to calculate the variable $k(t)$. If the control is a tangential control, then the adjoint pressure and the term with $k(t)$ can be neglected.

3. Semidiscrete-in-time approximations.

3.1. Formulation of the semidiscrete-in-time optimal control problem.

Let $\sigma_N = \{t_n\}_{n=0}^N$ be a partition of $[0, T]$ into equal intervals $\Delta t = T/N$ with $t_0 = 0$ and $t_N = T$. For each fixed Δt (or N) and for every quantity $q(t, \vec{x})$, we associate the corresponding set $\{q^{(n)}(\vec{x})\}_{n=0}^N$ and a continuous piecewise linear function $q^N = q^N(t, \vec{x})$ such as $q^N(t_n, \vec{x}) = q^{(n)}(\vec{x})$ $\forall$ $n = 0, 1, \dots, N$. We will denote with boldface letter $\mathbf{q}$ the vector $(q^{(1)}, q^{(2)}, \dots, q^{(N)})$ of the discrete time components. Also, the space X^N will be denoted as $\mathbf{X}$. On this partition we define the discrete target velocity as $\vec{U}^{(n)}(\vec{x}) = \vec{U}(t_n, \vec{x})$ for $n = 0, 1, \dots, N$ when $\vec{U} \in U_{ad}$. Let Γ_c be part of the boundary on which we apply the boundary control $\vec{g}$ and

$$H_n^1(\Gamma_c) = \left\{ \vec{g} \in H^1(\Gamma_c) : \int_{\Gamma_c} \vec{g}^{(n)} \cdot \vec{n} \; d\vec{x} = 0 \right\} \quad \text{for } n = 1, 2, \dots, N,$$

and $H_{n0}^1(\Gamma) = H_0^1(\Gamma \setminus \Gamma_c) \cap H_n^1(\Gamma_c)$ denote the spaces of all the functions that are compatible with the divergence free motion of the fluid. We remark that the subspaces $H_{n0}^1(\Gamma)$ and $H_n^1(\Gamma)$ are closed subspaces of $H^1(\Gamma)$ and the space $H^1(\Gamma)$ can be decomposed in $H_n^1(\Gamma) \oplus (H_n^1)^\perp(\Gamma)$. We assume no-slip boundary conditions on the rest of the boundary $\Gamma \setminus \Gamma_c$. Hence, the component of the velocity $\vec{u}^{(n)}$ on the boundary is the canonical extension of $\vec{g}^{(n)}$ from $H_n^1(\Gamma_c)$ to $H^1(\Gamma)$. We recall that this extension is a continuous map. The state variables $\vec{u}^{(n)} \in H_0^1(\Omega)$ and $p^{(n)} \in L_0^2(\Omega)$ are constrained to satisfy the semidiscrete Navier–Stokes equations

$$\begin{cases} \frac{1}{\Delta t}(\vec{u}^{(n)} - \vec{u}^{(n-1)}, \vec{v}) + \nu a(\vec{u}^{(n)}, \vec{v}) \\ \quad + c(\vec{u}^{(n)}; \vec{u}^{(n)}, \vec{v}) + b(\vec{v}, p^{(n)}) = 0 \quad \forall \, \vec{v} \in H_0^1(\Omega), \\ b(\vec{u}^{(n)}, q) = 0 \quad \forall \, q \in L_0^2(\Omega), \\ (\vec{u}^{(n)}(\vec{x}), \vec{s})_\Gamma = (\vec{g}^{(n)}(\vec{x}), \vec{s})_{\Gamma_c} \quad \forall \, \vec{s} \in H^{-1/2}(\Gamma), \end{cases} \tag{51}$$

for $n = 1, 2, \dots, N$ with $\vec{u}^{(0)} = \vec{u}_0(\vec{x}) \in curl(H^2)(\Omega)$.

Optimization is achieved by means of the minimization of the discretized functional

$$\begin{aligned} \mathcal{J}^N(\vec{\mathbf{u}}, \vec{\mathbf{g}}) = {} & \frac{\alpha}{2} \sum_{n=1}^N \|\vec{u}^{(n)} - \vec{U}^{(n)}\|^2 \Delta t \\ & + \frac{\beta}{2} \sum_{n=1}^N \left[\|\vec{g}^{(n)}\|_{\Gamma_c}^2 \Delta t + \beta_1 \|\partial_s \vec{g}^{(n)}\|_{\Gamma_c}^2 \Delta t + \beta_2 \| \, \vec{g}^{(n)} - \vec{g}^{(n-1)} \, \|_{\Gamma_c}^2 \right]. \end{aligned} \tag{52}$$

Of course, if Δt tends to zero, this functional tends to the corresponding continuous functional (4).

The *admissibility set* A_{ad} is defined by

$$\begin{aligned} A_{ad} = \{ & (\vec{\mathbf{u}}, \mathbf{p}, \vec{\mathbf{g}}) \in \mathbf{H}^1(\Omega) \times \mathbf{L}_0^2(\Omega) \times \mathbf{H}_{n0}^1(\Gamma_c) \quad \textit{such that } (\vec{\mathbf{u}}, \mathbf{p}, \vec{\mathbf{g}}) \textit{ is} \\ & \textit{solution of } (51), \ \vec{g}^{(0)} = \gamma \vec{u}_0, \textit{ and the functional in } (52) \textit{ is bounded} \}. \end{aligned}$$

The formulation of the optimal control problem in the semidiscrete approximation is given as follows:

given $\Delta t = T/N$, $\vec{u}_0 \in curl(H^2)(\Omega)$, *and* $\vec{U} \in U_{ad}$, *then* $(\vec{\mathbf{u}}, \mathbf{p}, \vec{\mathbf{g}}) \in\ A_{ad}$ *is called an optimal solution if there exists* $\epsilon > 0$ *such that*

$$\mathcal{J}^N(\vec{\mathbf{u}}, \vec{\mathbf{g}}) \le \mathcal{J}^N(\widetilde{\mathbf{u}}, \widetilde{\mathbf{h}}) \quad \forall \widetilde{h} \in \mathbf{H}^1_{n0} \tag{53}$$

whenever $\|\vec{g}^{(n)} - \widetilde{h}^{(n)}\|_{\Gamma_c} \le \epsilon$, *with* $n = 1, 2, \dots, N$.

For the semidiscrete Navier–Stokes nonhomogeneous boundary problem, one can prove the following theorem [29].

Theorem 3.1. *Let* $\Delta t = T/N$ *and* $\vec{u}_0 \in curl(H^2)(\Omega)$. *Let* $\epsilon > 0$, $\vec{\mathbf{g}} \in \mathbf{H}^1_{n0}(\Omega)$ *such that* $\sum_{i=1}^N (\|\vec{g}^{(n)}\|_1^2 \Delta t + \|\vec{g}^{(n)} - \vec{g}^{(n-1)}\|^2) \le \epsilon$, *i.e.,* $\vec{g}^N$ *and* $\vec{g}'^N$ *are uniformly bounded by* ϵ *in* $L^2((0,T); H^1(\Gamma_c))$ *and in* $L^2((0,T); L^2(\Gamma_c))$, *respectively. Then, there exists a function* $\vec{\mathbf{u}} \in \mathbf{H}^1(\Omega)$ *that is a solution of the system*

$$\begin{cases} \frac{1}{\Delta t}(\vec{u}^{(n)} - \vec{u}^{(n-1)}, \vec{v}) + \nu a(\vec{u}^{(n)}, \vec{v}) + c(\vec{u}^{(n)}; \vec{u}^{(n)}, \vec{v}) \\ \qquad + b(\vec{v}, p^{(n)}) = 0 \quad \forall \vec{v} \in H^1_0(\Omega), \quad for \quad n = 1, \dots, N, \\ b(\vec{u}^{(n)}, q) = 0 \quad \forall q \in L^2_0(\Omega), \quad for \quad n = 1, \dots, N, \\ \vec{u}^{(n)}(\vec{x}) = \vec{g}^{(n)} \quad for\ \vec{x} \in \Gamma_c, \quad for \quad n = 1, \dots, N, \\ \vec{u}^{(n)}(\vec{x}) = 0 \quad for\ \vec{x} \in \Gamma \setminus \Gamma_c, \quad for \quad n = 1, \dots, N, \\ \vec{u}^{(0)} = \vec{u}_0(\vec{x}) \in curl(H^2)(\Omega), \end{cases} \tag{54}$$

with the following estimates:

$$\|\vec{u}^{(n)}\|_1^2 \le K, \quad n = 1, 2, \dots, N, \tag{55}$$

$$\sum_{n=1}^N \|\nabla \vec{u}^{(n)}\|^2 \Delta t \le K, \tag{56}$$

$$\sum_{n=1}^N \|\vec{u}^{(n)} - \vec{u}^{(n-1)}\|^2_{H^{-1}} \le K, \tag{57}$$

where the constant K *is independent of* Δt.

If $\vec{\mathbf{g}}$ and its time derivative are uniformly bounded, then the existence of solutions of the semidiscrete-in-time optimal control problem can be proved. This fact is an easy consequence of the definition of the optimal control problem and the boundedness of the functional.

Lemma 3.2. *Let* $\Delta t = T/N$, $\vec{u}_0 \in curl(H^2)(\Omega)$, *and* $\vec{U} \in U_{ad}$. *If* $(\vec{\mathbf{u}}, \vec{\mathbf{g}})$ *is the solution of the semidiscrete optimal control problem, then for all* β_1 *and* $\beta_2 > 0$ *there exists a constant* C *independent of* Δt *such that*

$$\sum_{n=1}^N \|\vec{g}^{(n)}\|^2_{1,\Gamma} \Delta t \le C, \tag{58}$$

$$\sum_{n=1}^N \|\vec{g}^{(n)} - \vec{g}^{(n-1)}\|^2_{\Gamma} \le C, \tag{59}$$

$$\sum_{n=1}^N \|\vec{u}^{(n)}\|^2_{\Omega} \Delta t \le C. \tag{60}$$

Hence, we have that $\vec{g}^N \in L^2((0,T);H^1(\Gamma))$, $\vec{g}'^N \in L^2((0,T);L^2(\Gamma))$, *and* $\vec{u}^N \in L^2((0,T);W(\Omega))\ \forall\ N$.

The proof can be obtained by using standard techniques and can be found in [29]. We can recall that if the norm of $\vec{g}^N \in L^2((0,T);H^1(\Gamma))$ and the norm of $\vec{g}'^N \in L^2((0,T);L^2(\Gamma))$ are uniformly bounded for all N, then $\vec{g}^N$ is uniformly bounded in $L^2((0,T);L^2(\Gamma))\ \forall\ N$. Now we can state and prove the existence of solutions for the optimal control problem in an open bounded domain Ω with boundary Γ in C^2.

THEOREM 3.3. *Given* $\Delta t = T/N$, $\vec{u}_0 \in curl(H^2)(\Omega)$, *and* $\vec{U} \in U_{ad}$, *there exists a solution* $(\vec{\mathbf{u}}, \mathbf{p}, \vec{\mathbf{g}}) \in A_{ad}$ *of* (51) *such that* $\vec{\mathbf{g}}$ *minimizes the cost functional.*

Proof. The proof proceeds as in the continuous case. Let $\Delta t = T/N$ and $\{\vec{\mathbf{g}}_k\}_{k=1}^{\infty}$ be a minimizing sequence in $\mathbf{H}^1_{n0}(\Gamma_c)$. Using Theorem 3.1 and the result in (58)–(59), we find that the corresponding sequence $\vec{\mathbf{u}}_k$ is uniformly bounded in $\mathbf{H}^1(\Omega)$. Now, we can proceed with a weakly convergent subsequence and show that this subsequence converges to the solution of the optimal control problem in the semidiscrete approximation. We can write

$$\begin{aligned} \vec{g}_k^{(n)} \to \vec{g}^{(n)} &\quad \text{in} \quad H^1_0(\Gamma_c) \quad \text{weakly}, \\ \vec{u}_k^{(n)} \to \widehat{u}^{(n)} &\quad \text{in} \quad H^1(\Omega) \quad \text{weakly} \end{aligned}$$

for $n = 1, 2, \ldots, N$. By using the fact that the injection of $H^1(\Omega)$ into $L^2(\Omega)$ is compact, the subsequence converges strongly. The lower semicontinuity of the functional in (52) allows the pair $(\vec{\mathbf{u}}, \vec{\mathbf{g}})$ to minimize the functional. Since we can pass to the limit in the linear and the nonlinear terms, the pair also satisfies the semidiscrete Navier–Stokes system (51). In fact, since $\vec{\mathbf{u}}_k$ converges to $\vec{\mathbf{u}}$ strongly in $\mathbf{L}^2(\Omega)$, then for any $\vec{\mathbf{z}} \in \mathcal{V}(\Omega)$, we have

$$\lim_{k\to\infty} c(\vec{\mathbf{u}}_k; \vec{\mathbf{u}}_k, \vec{\mathbf{z}}) = c(\vec{\mathbf{u}}; \vec{\mathbf{u}}, \vec{\mathbf{z}})\,.$$

Since $\mathcal{V}(\Omega)$ is dense in $\mathbf{H}^1_0(\Omega)$, this is still true for any $\vec{\mathbf{w}}$ in $\mathbf{H}^1_0(\Omega)$ by a continuity argument. This allows us to pass to the limit in the semidiscrete equations and complete the proof. □

3.2. First-order necessary condition. In this section, we derive the first-order necessary condition in a different way. Denote by $\mathbf{B}_1, \mathbf{B}_2$ the following sets:

$$\left\{ \begin{array}{l} \mathbf{B}_1 = \mathbf{H}^1(\Omega) \times \mathbf{L}^2_0(\Omega) \times \mathbf{H}^1_{n0}(\Gamma_c), \\ \mathbf{B}_2 = \mathbf{H}^{-1}(\Omega) \times \mathbf{L}^2_0(\Omega) \times \mathbf{H}^{1/2}(\Gamma_c). \end{array} \right. \tag{61}$$

We define the nonlinear map

$$M(\vec{\mathbf{u}}, \mathbf{p}, \vec{\mathbf{g}}) : \mathbf{B}_1 \to \mathbf{B}_2$$

as $M(\vec{\mathbf{u}}, \mathbf{p}, \vec{\mathbf{g}}) = (\vec{\mathbf{f}}, \mathbf{z}, \vec{\mathbf{b}})$ if and only if

$$\left\{ \begin{array}{l} \frac{1}{\Delta t}(\vec{u}^{(n)} - \vec{u}^{(n-1)}, \vec{v}) + \nu a(\vec{u}^{(n)}, \vec{v}) + c(\vec{u}^{(n)}; \vec{u}^{(n)}, \vec{v}) \\ \qquad + b(\vec{v}, p^{(n)}) = (\vec{f}^{(n)}, \vec{v}) \quad \forall\, \vec{v} \in H^1_0(\Omega), \quad \textit{for} \quad n = 1, \ldots, N, \\ b(\vec{u}^{(n)}, q) = (z^{(n)}, q) \quad \forall\, q \in L^2_0(\Omega), \quad \textit{for} \quad n = 1, \ldots, N, \\ (\vec{u}^{(n)}, \vec{s})_\Gamma - (\vec{g}^{(n)}, \vec{s})_{\Gamma_c} = (\vec{b}^{(n)}, \vec{s})_\Gamma \quad \forall\, \vec{s} \in H^{-1/2}(\Gamma), \quad \textit{for} \quad n = 1, \ldots, N, \\ \vec{u}^{(0)} = \vec{u}_0(\vec{x}) \in curl(H^2)(\Omega). \end{array} \right. \tag{62}$$

In the same manner, let $\widehat{\mathbf{g}}$ be an optimal solution and define

$$N(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}}):\mathbf{B}_1\to\mathbb{R}\times\mathbf{B}_2$$

as $N(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})=(a,\vec{\mathbf{f}},\mathbf{z},\vec{\mathbf{b}})$ if and only if

$$\begin{pmatrix} \mathcal{J}^N(\vec{\mathbf{u}},\vec{\mathbf{g}})-\mathcal{J}^N(\widehat{\mathbf{u}},\widehat{\mathbf{g}})\\ M(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})\end{pmatrix}=\begin{pmatrix}a\\(\vec{\mathbf{f}},\mathbf{z},\vec{\mathbf{b}})\end{pmatrix}. \tag{63}$$

Thus, the constraints can be expressed as $M(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})=(0,0,0)$ and the optimal control problem can be reformulated as follows:

find $(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}})$ *and* $a\leq 0$ *such that the equation* $N(\vec{\mathbf{u}},\mathbf{p},\widehat{\mathbf{g}})=(a,0,0,0)$
is satisfied for all $\vec{\mathbf{g}}$ *such that* $\|\vec{g}^{(n)}-\widehat{g}^{(n)}\|\leq\epsilon$ *for* $n=1,2,\dots,N$
for some $\epsilon>0$.

Since we are looking for local minimum points, it is natural to define the operators $M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$ and $N'(\vec{\mathbf{u}},\mathbf{p},\widehat{\mathbf{g}})$. Given a $(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$, we define the linear operator

$$M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}}):\mathbf{B}_1\to\mathbf{B}_2$$

as $M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})\cdot(\widetilde{\mathbf{w}},\mathbf{r},\widetilde{\mathbf{h}})=(\bar{\mathbf{f}},\bar{\mathbf{z}},\bar{\mathbf{b}})$ if and only if

$$\begin{cases} \frac{1}{\Delta t}(\widetilde{w}^{(n)}-\widetilde{w}^{(n-1)},\vec{v})+\nu a(\widetilde{w}^{(n)},\vec{v})+c(\widetilde{w}^{(n)};\vec{u}^{(n)},\vec{v})+c(\vec{u}^{(n)},\widetilde{w}^{(n)},\vec{v})\\ \quad +b(\vec{v},r^{(n)})=(\bar{f}^{(n)},\vec{v})\quad\forall\,\vec{v}\in H_0^1(\Omega),\quad \textit{for}\quad n=1,\dots,N,\\ b(\widetilde{w}^{(n)},q)=(\bar{z}^{(n)},q)\quad\forall q\in L_0^2(\Omega),\quad \textit{for}\quad n=1,\dots,N,\\ (\widetilde{w}^{(n)},\vec{s})_\Gamma-(\widetilde{h}^{(n)},\vec{s})_{\Gamma_c}=(\bar{b}^{(n)},\vec{s})_\Gamma\quad\forall\,\vec{s}\in H^{-1/2}(\Gamma),\quad \textit{for}\quad n=1,\dots,N,\\ \widetilde{w}^{(0)}=\vec{0}. \end{cases} \tag{64}$$

Let

$$N'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}}):\mathbf{B}_1\to\mathbb{R}\times\mathbf{B}_2$$

be defined as $N'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})\cdot(\widetilde{a},\widetilde{\mathbf{w}},\mathbf{r},\widetilde{\mathbf{h}})=(\bar{a},\bar{\mathbf{f}},\bar{\mathbf{z}},\bar{\mathbf{b}})$ if and only if

$$\begin{pmatrix} \mathcal{J}^{N'}(\vec{\mathbf{u}},\vec{\mathbf{g}}))\cdot(\widetilde{a},\widetilde{\mathbf{w}},\mathbf{r},\widetilde{\mathbf{h}}))\\ M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})\cdot(\widetilde{a},\widetilde{\mathbf{w}},\mathbf{r},\widetilde{\mathbf{h}}))\end{pmatrix}=\begin{pmatrix}\bar{a}\\(\bar{\mathbf{f}},\bar{\mathbf{z}},\bar{\mathbf{b}})\end{pmatrix}. \tag{65}$$

Now we have to prove that these operators are well defined, i.e., the equations for the Gâteaux derivatives are well posed and have solutions.

LEMMA 3.4. *Given* $\Delta t=T/N$, $\vec{u}_0\in$ *curl*$(H^2)(\Omega)$, *and* $\vec{\mathbf{u}}\in\mathbf{H}^1(\Omega)$. *Then, we have*

(i) *the operator* $M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$ *has closed range and is onto in* $\mathbf{B}_2$;

(ii) *the operator* $N'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$ *has closed range in* $\mathbb{R}\times\mathbf{B}_2$.

Proof. (i) We set

$$\begin{cases} \nu\widetilde{a}(\widetilde{w}^{(n)},\vec{v})=\nu a(\widetilde{w}^{(n)},\vec{v})+\frac{1}{\Delta t}(\widetilde{w}^{(n)},\vec{v})\quad\forall\vec{v}\in H_0^1(\Omega),\quad n=1,2,\dots,N,\\ (\widetilde{f}^{(n)},\vec{v})=(\vec{f}^{(n)},\vec{v})+\frac{1}{\Delta t}(\widetilde{w}^{(n-1)},\vec{v})\quad\forall\vec{v}\in H_0^1(\Omega),\quad n=1,2,\dots,N. \end{cases}$$

With this notation, the operator $M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$ can be written as

$$(66)\qquad \begin{cases} \nu\widetilde{a}(\widetilde{w}^{(n)},\vec{v})+c(\widetilde{w}^{(n)};\vec{u}^{\,(n)},\vec{v}) \\ \quad +\,c(\vec{u}^{\,(n)},\widetilde{w}^{(n)},\vec{v})+b(\vec{v},r^{(n)})=(\widetilde{f}^{(n)},\vec{v}) \quad \forall\,\vec{v}\in H_0^1(\Omega), \\ b(\widetilde{w}^{(n)},q)=(\bar{z}^{\,(n)},q) \quad \forall\,q\in L_0^2(\Omega), \\ (\widetilde{w}^{(n)},\vec{s})_\Gamma-(\widetilde{h}^{(n)},\vec{s})_{\Gamma_c}=(\bar{b}^{\,(n)},\vec{s})_\Gamma \quad \forall\,\vec{s}\in H^{-1/2}(\Gamma), \end{cases}$$

for $n=1,2,\dots,N$ with $\widetilde{w}^{(0)}=\vec{0}$. The function $\widetilde{f}$ is still in $H^{-1}(\Omega)$ and the range of M' is still the same if $Ran(M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}}))=\mathbf{B}_2$. This is a steady system and one can apply the standard techniques for the stationary case; see [21]. Let S be the Stokes operator

$$S=\begin{pmatrix} A & B^* \\ B & 0 \\ \gamma_0 & 0 \end{pmatrix}.$$

By the trace theorem, using the ellipticity of A and the inf-sup property, one can see [8], [17], [18] that the Stokes operator is an isomorphism from $H^1(\Omega)\times L_0^2(\Omega)$ $\to H^{-1}(\Omega)\times L_0^2(\Omega)\times H^{-1/2}(\Gamma)$. The operator $C(\vec{w}^{\,(n)})\vec{u}^{\,(n)}$ is continuous in $\vec{w}^{\,(n)}$ from $H^{1/2}(\Gamma)$ into $H^{-1}(\Omega)$ $\forall\ \vec{u}^{\,(n)}\in H^1(\Omega)$ and $n=1,2,\dots,N$ and thus compact from $H^1(\Omega)$ into $H^{-1}(\Omega)$. The operator $C(\vec{u}^{\,(n)})\vec{w}^{\,(n)}$ is continuous $\forall\ \vec{u}\in H^1(\Omega)$ and $\forall\ n=1,2,\dots,N$ from $H^1(\Omega)$ into $H^{-1/2}(\Gamma)$ and thus compact from $H^1(\Omega)$ into $H^{-1}(\Omega)$. The perturbation operator

$$M'^{(n)}(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})=S^{(n)}+\begin{pmatrix} C(\vec{u}^{\,(n)})\vec{w}^{\,(n)}+C(\vec{w}^{\,(n)})\vec{u}^{\,(n)} \\ 0 \\ 0 \end{pmatrix}$$

is a Fredholm operator $\forall\ n=1,2,\dots,N$, i.e., it has a closed range and a finite-dimensional kernel.

(ii) The operator $M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$ belongs to $\mathcal{L}(\mathbf{B}_1,Ran(M'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}}))$ and therefore its kernel is a closed subspace. We recall that a linear functional $\vec{f}$ on a Banach space can have either $Ran(\vec{f})=\{0\}$ or $Ran(\vec{f})=\{\mathbb{R}\}$. Now, $N'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$ acting on the kernel is either identically zero or onto $\mathbb{R}$. Let X,Y,Z be Banach spaces and $A:X\to Y$ and $B:X\to Z$ be linear continuous operators. If the range of B is closed in Z and the subspace $A\cdot\ker(B)$ is closed in Y, then, if we define $C:X\to Y\times Z$ by $Cx=(Ax,Bx)$, the range of C is closed in $Y\times Z$. Applying this result we prove that $Ran(N'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}}))$ is a closed set. □

The operator $N'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$ cannot be onto. If it were, by the implicit function theorem, we would have that there exists a solution, which is different from the optimal solution, that minimizes the functional for every small neighborhood of $(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}})$. This contradicts the hypothesis that $(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}})$ is an optimal solution. Therefore the optimality condition implies the following theorem.

THEOREM 3.5. *Given $\Delta t=T/N$ and $\vec{u}_0\in curl(H^2)(\Omega)$. If $(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}})\in(\mathbf{H}^1(\Omega)\times\mathbf{L}_0^2(\Omega)\times\mathbf{H}_{n0}^1(\Gamma_c))$ is a solution of the semidiscrete optimal control problem, then the operator $N'(\vec{\mathbf{u}},\mathbf{p},\vec{\mathbf{g}})$ is not onto in $\mathbb{R}\times\mathbf{B_2}$.*

In the next theorem, we write the first-order necessary condition and characterize the optimal control solution as a solution of the corresponding Euler system of equations.

THEOREM 3.6. *Given $\Delta t = T/N$ and $\vec{u}_0 \in curl(H^2)(\Omega)$. If $(\widehat{\mathbf{u}}, \mathbf{p}, \widehat{\mathbf{g}}) \in \mathbf{H}^1(\Omega) \times \mathbf{L}_0^2(\Omega) \times \mathbf{H}_{n0}^1(\Gamma_c)$ is an optimal solution, i.e., the operator $N'(\vec{\mathbf{u}}, \mathbf{p}, \vec{\mathbf{g}})$ is not onto, then there exists a nonzero Lagrangian multiplier $\vec{\mathbf{w}}, \boldsymbol{\sigma}, \vec{\boldsymbol{\xi}} \in \mathbf{H}_0^1(\Omega) \times \mathbf{L}_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$ satisfying the Euler equations*

$$\mathcal{J}^{N'}(\widehat{\mathbf{u}}, \widehat{\mathbf{g}}) \cdot (\widetilde{\mathbf{w}}, \mathbf{r}, \widetilde{\mathbf{h}}) + \Big\langle (\vec{\mathbf{w}}, \boldsymbol{\sigma}, \vec{\boldsymbol{\xi}}), M'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}) \cdot (\widetilde{\mathbf{w}}, \mathbf{r}, \widetilde{\mathbf{h}}) \Big\rangle = 0$$
$$\forall\, (\widetilde{\mathbf{w}}, \mathbf{r}, \widetilde{\mathbf{h}}) \in \mathbf{H}^1(\Omega) \times \mathbf{L}_0^2(\Omega) \times \mathbf{H}_{n0}^1(\Gamma_c), \tag{67}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $\mathbb{R} \times \mathbf{B}_2$ and $\mathbb{R} \times \mathbf{B}_2^*$.

Proof. From Lemma 3.4, the range of $N'(\vec{\mathbf{u}}, \mathbf{p}, \vec{\mathbf{g}})$ is a closed set and, from Theorem 3.5, this range is a closed proper subspace of $\mathbb{R} \times \mathbf{B}_2$. The Hahn–Banach theorem then implies that there exists a nonzero element of $\mathbb{R} \times \mathbf{B}_2^* = \mathbb{R} \times \mathbf{H_0^1}(\Omega) \times \mathbf{L_0^2}(\Omega) \times \mathbf{H^{1/2}}(\Gamma)$ that annihilates the range of $N'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}})$. One can find an $(\widehat{a}, \vec{\mathbf{w}}, \boldsymbol{\sigma}, \vec{\boldsymbol{\xi}}) \in \mathbb{R} \times \mathbf{B_2}^*$ such that

$$\Big\langle (\widetilde{a}, \widetilde{\mathbf{f}}, \widetilde{\mathbf{z}}, \widetilde{\mathbf{b}}), (\widehat{a}, \vec{\mathbf{w}}, \boldsymbol{\sigma}, \vec{\boldsymbol{\xi}}) \Big\rangle = 0 \quad \forall\, (\widetilde{a}, \widetilde{\mathbf{f}}, \widetilde{\mathbf{z}}, \widetilde{\mathbf{b}}) \in Ran(N'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}})),$$

where $\widehat{a}$ is different from zero as this solution is nontrivial, i.e., $M'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}})$ is onto. If we set $\widehat{a} = 1$, we have (67). □

3.3. The optimality system. From the first-order necessary condition, we can further characterize the optimal control as a solution of an differential equation.

THEOREM 3.7. *Given $\Delta t = T/N$ and $\vec{u}_0 \in curl(H^2)(\Omega)$. Let $(\vec{\mathbf{u}}, \vec{\mathbf{p}}, \vec{\mathbf{g}}) \in (\mathbf{H}^1(\Omega) \times \mathbf{L}_0^2(\Omega) \times \mathbf{H}_{n0}^1(\Gamma_c))$ denote an optimal control solution. Then, the control $\vec{g}^{\,(n)}$ satisfying the compatibility conditions*

$$\vec{g}^{\,(0)} = \gamma_0 \vec{u}_0, \qquad \int_\Gamma \vec{g}^{\,(n)} \cdot \vec{n}\; d\vec{x} = 0 \quad \textit{for}\, n = 0, 1, \ldots, N$$

is solution of the system

$$\int_{\Gamma_c} \left[\vec{g}^{\,(n)} \cdot \widetilde{h} - \frac{\beta_1}{\Delta t^2} (\vec{g}^{\,(n+1)} - 2\vec{g}^{\,(n)} + \vec{g}^{\,(n-1)}) \cdot \widetilde{h} + \beta_2 \partial_s \vec{g}^{\,(n)} \cdot \partial_s \widetilde{h} + k^{(n)} \vec{n} \cdot \widetilde{h} \right] d\vec{x}$$
$$= \frac{1}{\beta} \int_{\Gamma_c} (\vec{\xi}^{(n)} \cdot \widetilde{h})\, d\vec{x} \qquad \forall \widetilde{h} \in H_0^1(\Gamma_c),$$

for $n = 1, \ldots, N-1$ ($\vec{g}^{\,N} = \vec{g}^{\,N-1}$). The function $\vec{\xi} \in \mathbf{H}^{-1/2}(\Gamma_c)$ is defined by

$$\int_{\Gamma_c} \vec{\xi}^{(n-1)} \cdot \widetilde{v}\, d\vec{x} = \int_{\Gamma_c} (\gamma_1 \vec{w}^{\,(n-1)} - \sigma^{(n-1)} \vec{n}) \cdot \widetilde{v}\, d\vec{x} \tag{68}$$
$$= -\alpha(\vec{u}^{\,(n)} - \vec{U}^{\,(n)}, \widetilde{v}) - \frac{1}{\Delta t} (\vec{w}^{\,(n)} - \vec{w}^{\,(n-1)}, \widetilde{v})_\Omega + \nu a(\vec{w}^{\,(n-1)}, \widetilde{v})$$
$$+ c(\widetilde{v}; \vec{u}^{(n)}, \vec{w}^{\,(n-1)}) + c(\vec{u}^{\,(n)}; \widetilde{v}, \vec{w}^{\,(n-1)}) + b(\widetilde{v}, \sigma^{(n-1)})$$
$$\forall \widetilde{v} \in H^1(\Omega), \quad \textit{for} \quad n = 1, \ldots, N,$$

where $\vec{\mathbf{w}}$ *and* $\boldsymbol{\sigma}$ *satisfies*

$$
(69)\qquad
\begin{cases}
-\frac{1}{\Delta t}(\vec{w}^{\,(n)} - \vec{w}^{\,(n-1)}, \vec{v}) + \nu a(\vec{w}^{\,(n-1)}, \vec{v}) + c(\vec{w}^{\,(n-1)}; \vec{u}^{\,(n)}, \vec{v}) \\
\qquad + c(\vec{u}^{\,(n)}; \vec{w}^{\,(n-1)}, \vec{v}) + b(\vec{v}, \sigma^{(n-1)}) = \alpha(\vec{u}^{\,(n)} - \vec{U}^{\,(n)}, \vec{v}) \\
\qquad \forall\, \vec{v} \in H^1_0(\Omega), \quad \textit{for} \quad n = 1, \dots, N, \\
b(\vec{w}^{\,(n-1)}, q) = 0 \quad \forall\, q \in L^2_0(\Omega), \quad \textit{for} \quad n = 1, \dots, N, \\
(\vec{w}^{\,(n-1)}, \vec{s}\,) = 0 \quad \forall\, \vec{s} \in H^{-1/2}(\Gamma), \quad \textit{for} \quad n = 1, \dots, N, \\
\vec{w}^{\,(N)} = 0.
\end{cases}
$$

Proof. Given $\widetilde{\mathbf{h}} \in \mathbf{H}^1_{n0}(\Gamma_c)$, the first-order necessary condition (67) can be written as

$$
\begin{aligned}
&\alpha \sum_{n=1}^{N} ((\vec{u}^{\,(n)} - \vec{U}^{\,(n)})\widetilde{w}^{(n)}, 1)\Delta t \\
&+ \beta \sum_{n=1}^{N} \int_{\Gamma_c} \left[\vec{g}^{\,(n)} \cdot \widetilde{h}^{(n)} + \frac{\beta_1}{\Delta t^2}(\vec{g}^{\,(n)} - \vec{g}^{\,(n-1)})(\widetilde{h}^{(n)} - \widetilde{h}^{(n-1)}) + \beta_2 \partial_s \vec{g}^{\,(n)} \cdot \partial_s \widetilde{h}^{(n)} \right] d\vec{x}\Delta t \\
&+ \sum_{n=1}^{N} (\bar{f}^{\,(n)}, \vec{w}^{\,(n)})_\Omega \Delta t + \sum_{n=1}^{N} (\bar{z}^{\,(n)}, \sigma^{(n)})_\Omega \Delta t + \sum_{n=1}^{N} (\bar{b}^{\,(n)}, \vec{\xi}^{(n)})_\Gamma \Delta t = 0
\end{aligned}
$$

$\forall\ (\bar{\mathbf{f}}, \bar{\mathbf{z}}, \bar{\mathbf{b}}) \in \mathbf{H}^{-1}(\Omega) \times \mathbf{L}^2_0(\Omega) \times \mathbf{H}^{-1/2}(\Gamma)$. We have

$$
\begin{aligned}
&\alpha \sum_{n=1}^{N} ((\vec{u}^{\,(n)} - \vec{U}^{\,(n)})\widetilde{w}^{(n)}, 1)\Delta t \\
&+ \beta \sum_{n=1}^{N} \int_{\Gamma_c} \left[\vec{g}^{\,(n)} \cdot \widetilde{h}^{(n)} + \frac{\beta_1}{\Delta t^2}(\vec{g}^{\,(n)} - \vec{g}^{\,(n-1)})(\widetilde{h}^{(n)} - \widetilde{h}^{(n-1)}) + \beta_2 \partial_s \vec{g}^{\,(n)} \cdot \partial_s \widetilde{h}^{(n)} \right] d\vec{x}\Delta t \\
&- \sum_{n=1}^{N} \left[\frac{1}{\Delta t}(\widetilde{w}^{(n)} - \widetilde{w}^{(n-1)}, \vec{w}^{\,(n)}) + \nu a(\widetilde{w}^{(n)}, \vec{w}^{\,(n)}) + c(\widetilde{w}^{(n)}; \vec{u}^{\,(n)}, \vec{w}^{\,(n)}) \right. \\
&\qquad \left. +\ c(\vec{u}^{\,(n)}; \widetilde{w}^{(n)}, \vec{w}^{\,(n)}) + b(\vec{w}^{\,(n)}, r^{(n)}) \right] \Delta t + \sum_{n=1}^{N} b(\widetilde{w}^{(n)}, \sigma^{(n)})\Delta t \\
&+ \sum_{n=1}^{N} [(\vec{w}^{\,(n)}, \vec{\xi}^{(n)})_\Gamma - (\widetilde{h}^{(n)}, \vec{\xi}^{(n)})_{\Gamma_c}]\Delta t = 0
\end{aligned}
$$

$\forall\ (\widetilde{\mathbf{w}}, \vec{\mathbf{r}}, \widetilde{\mathbf{h}}) \in \mathbf{H}^1(\Omega) \times \mathbf{L}^2_0(\Omega) \times \mathbf{H}^1_{n0}(\Gamma)$. After integration by parts, we define the quantity $\vec{\boldsymbol{\xi}}$, as in (68), with $\vec{\mathbf{w}}$ satisfying the adjoint equations in (69). Thus, we write

$$
\begin{aligned}
&\frac{\beta_1}{\Delta t} \int_{\Gamma_c} \widetilde{h}^{(N)}(\vec{g}^{\,(N)} - \vec{g}^{\,(N-1)})\, d\vec{x}\Delta t + \sum_{n=1}^{N-1} \int_{\Gamma_c} [\vec{g}^{\,(n)} \cdot \widetilde{h}^{(n)} + \beta_2 \partial_s \vec{g}^{\,(n)} \cdot \partial_s \widetilde{h}^{(n)}]\, d\vec{x}\Delta t \\
&\qquad + \sum_{n=1}^{N-1} \int_{\Gamma_c} \left[-\frac{\beta_1}{\Delta t^2}(\vec{g}^{\,(n+1)} - 2\vec{g}^{\,(n)} + \vec{g}^{\,(n-1)}) - \frac{1}{\beta}\vec{\xi}^{(n)} \cdot \widetilde{h}^{(n)} \right] d\vec{x}\Delta t = 0.
\end{aligned}
$$

As the variation $\widetilde{h}$ is independent in the space $\mathbf{H}^1_{n0}(\Gamma_c)$, we have

$$\int_{\Gamma_c}\left[\vec{g}^{\,(n)}\cdot\widetilde{h}-\frac{\beta_1}{\Delta t^2}(\vec{g}^{\,(n+1)}-2\vec{g}^{\,(n)}+\vec{g}^{\,(n-1)})\cdot\widetilde{h}+\beta_2\partial_s\vec{g}^{\,(n)}\cdot\partial_s\widetilde{h}-\frac{1}{\beta}\vec{\xi}^{(n)}\cdot\widetilde{h}\right]d\vec{x}=0$$

$$\forall\widetilde{h}\in H^1_{n0}(\Gamma_c),\qquad n=1,\ldots,N-1, \tag{70}$$

with $\vec{g}^{\,(N)}=\vec{g}^{\,(N-1)}$. We recall that $H^1(\Gamma_c)=H^1_n(\Gamma_c)\oplus(H^1_n)^{\perp}(\Gamma_c)$ and thus

$$\widetilde{h}=\widetilde{h}_1-\vec{n}\frac{\int_{\Gamma_c}\widetilde{h}_1\cdot\vec{n}\,d\vec{x}}{\mu(\Gamma_c)},$$

where $\widetilde{h}_1\in H^1(\Gamma_c)$. We can write a weak formulation of (70) in the space $H^1_n(\Gamma_c)$ and then, by using the above decomposition, write a new formulation in $H^1(\Gamma_c)$. In this new formulation a constant vector in the normal direction appears. The weak formulation in $H^1_{n0}(\Gamma_c)$ reads

$$\begin{aligned}\int_{\Gamma_c}\Big[\vec{g}_1^{(n)}\cdot\widetilde{h}_1-\frac{\beta_1}{\Delta t^2}(\vec{g}_1^{(n+1)}-2\vec{g}_1^{(n)}+\vec{g}_1^{(n-1)})\cdot\widetilde{h}_1+\beta_2\partial_s\vec{g}_1^{(n)}\cdot\partial_s\widetilde{h}_1\\ +k^{(n)}\vec{n}\cdot\widetilde{h}_1\Big]\,d\vec{x}=\frac{1}{\beta}\int_{\Gamma_c}(\vec{\xi}^{(n)}\cdot\widetilde{h}_1)\,d\vec{x}\qquad\forall\vec{h}_1\in H^1_0(\Gamma_c),\quad \textit{for}\quad n=1,\ldots,N-1.\end{aligned}$$

The constant $k^{(n)}$ can be calculated by using the constraint

$$\int_{\Gamma}\vec{g}^{\,(n)}\cdot\vec{n}\;d\vec{x}=0.\qquad\Box$$

Now, in order to obtain the solution of the semidiscrete-in-time optimal control problem, we have to solve the semidiscrete Navier–Stokes system

$$\begin{cases}\frac{1}{\Delta t}(\vec{u}^{\,(n)}-\vec{u}^{\,(n-1)},\vec{v})+\nu a(\vec{u}^{\,(n)},\vec{v})+c(\vec{u}^{\,(n)};\vec{u}^{\,(n)},\vec{v})\\ \quad+b(\vec{v},p^{(n)})=0\quad\forall\,\vec{v}\in H^1_0(\Omega),\quad\text{for}\quad n=1,\ldots,N,\\ b(\vec{u}^{\,(n)},q)=0\quad\forall\,q\in L^2_0(\Omega),\quad\text{for}\quad n=1,\ldots,N,\\ (\vec{u}^{\,(n)},\vec{s})_{\Gamma}=(\vec{g}^{\,(n)},\vec{s})_{\Gamma_c}\quad\forall\,\vec{s}\in H^{-1/2}(\Gamma),\quad\textit{for}\quad n=1,\ldots,N,\\ \vec{w}^{\,(0)}=\vec{u}_0\in curl(H^2)(\Omega),\end{cases} \tag{71}$$

the semidiscrete adjoint system

$$\begin{cases}-\frac{1}{\Delta t}(\vec{w}^{\,(n)}-\vec{w}^{\,(n-1)},\vec{v})+\nu a(\vec{w}^{\,(n-1)},\vec{v})+c(\vec{u}^{\,(n)};\vec{v},\vec{w}^{\,(n-1)})\\ \quad+c(\vec{v};\vec{u}^{\,(n)},\vec{w}^{\,(n-1)})+b(\vec{v},\sigma^{(n-1)})=\alpha(\vec{u}^{\,(n)}-\vec{U}^{\,(n)},\vec{v})\\ \quad\forall\,\vec{v}\in H^1_0(\Omega),\quad\text{for}\quad n=1,\ldots,N,\\ b(\vec{w}^{\,(n)},q)=0\quad\forall\,q\in L^2_0(\Omega),\quad\text{for}\quad n=1,\ldots,N,\\ \vec{w}^{\,(n-1)}=0\quad\text{on}\quad\Gamma,\quad\textit{for}\quad n=1,\ldots,N,\\ \vec{w}^{\,(N)}=0\quad\text{in}\quad\Omega,\end{cases} \tag{72}$$

and the optimality condition

$$
(73)\quad \begin{cases} -\frac{\beta_1}{\Delta t^2}(\vec{g}^{\,(n+1)} - 2\vec{g}^{\,(n)} + \vec{g}^{\,(n-1)}, \widetilde{h}) + \beta_2(\partial_s \vec{g}^{\,(n)}, \partial_s \widetilde{h}) + (\vec{g}^{\,(n)}, \widetilde{h}) + k^{(n)}(\vec{n}, \widetilde{h}) \\ \quad = \frac{1}{\beta}(\gamma_1 \vec{\omega}^{\,(n)} - \vec{n}\sigma^{(n)}, \widetilde{h}) \quad \forall \widetilde{h} \in H^1_0(\Gamma_c), \quad \text{for} \quad n = 1, \dots, N-1, \\ (\vec{g}^{\,(n)}, \vec{n})_\Gamma = 0 \quad \text{for} \quad n = 1, \dots, N-1, \\ \vec{g}^{\,(n)} = 0 \quad \text{on} \quad \Gamma \setminus \Gamma_c \quad \text{for} \quad n = 1, \dots, N-1, \\ \vec{g}^{\,(0)} = \gamma_0 \vec{u}_0 \quad \text{on} \quad \Gamma_c, \\ \vec{g}^{(N)} = \vec{g}^{\,(N-1)} \quad \text{on} \quad \Gamma_c. \end{cases}
$$

4. Fully discrete space-time approximations.

4.1. Assumptions on the finite element spaces.

We consider only conforming finite element approximations. Let $X^h \subset H^1(\Omega)$ and $S^h \subset L^2(\Omega)$ be two families of finite dimensional subspaces parameterized by h that tends to zero. We also denote $X^h_0 = X^h \cap H^1_0(\Omega)$ and $S^h_0 = S^h \cap L^2_0(\Omega)$. We make the following assumptions on X^h and S^h

(a) *The approximation hypotheses:* there exists an integer l and a constant C, independent of h, $\vec{u}$, and p, such that for $1 \le k \le l$ we have

$$
(74)\qquad \inf_{\vec{u}_h \in X^h} \|\vec{u}_h - \vec{u}\|_1 \le Ch^k \|\vec{u}\|_{k+1} \quad \forall\, \vec{u} \in H^{k+1}(\Omega) \cap H^1_0(\Omega),
$$

$$
(75)\qquad \inf_{p_h \in S^h} \|p - p_h\| \le Ch^k \|p\|_k \qquad \forall\, p \in H^k(\Omega) \cap L^2_0(\Omega).
$$

(b) *The inf-sup condition or LBB condition:* there exists a constant C', independent of h, such that

$$
(76)\qquad \inf_{0 \ne q_h \in S^h} \ \sup_{0 \ne \vec{u}_h \in X^h} \frac{\int_\Omega q_h \operatorname{div} \vec{u}_h \, dx}{\|\vec{u}_h\|_1 \|q_h\|} \ge C' > 0.
$$

This condition assures the stability of the discrete Navier–Stokes solutions.

To preserve the antisymmetry of the trilinear form $c(\vec{u}; \vec{v}, \vec{w})$ on the finite element spaces, we introduce the modified trilinear form (see [32])

$$
\widetilde{c}(\vec{u}; \vec{v}, \vec{w}) = \frac{1}{2}\{c(\vec{u}; \vec{v}, \vec{w}) - c(\vec{u}; \vec{w}, \vec{v})\} \qquad \forall\, \vec{u}, \vec{v}, \vec{w} \in H^1(\Omega),
$$

from which we have

$$
(77)\qquad \begin{cases} \widetilde{c}(\vec{u}; \vec{v}, \vec{w}) = -\widetilde{c}(\vec{u}; \vec{w}, \vec{v}) \quad \forall\, \vec{u}, \vec{v}, \vec{w} \in H^1(\Omega), \\ \widetilde{c}(\vec{u}; \vec{v}, \vec{v}) = 0 \quad \forall\, \vec{u}, \vec{v}, \vec{w} \in H^1(\Omega) \end{cases}
$$

and

$$
c(\vec{u}; \vec{v}, \vec{w}) = \widetilde{c}(\vec{u}; \vec{v}, \vec{w}) \qquad \forall\, \vec{u} \in H^1_0(\Omega) \cap W(\Omega),\ \forall\, \vec{v} \in H^1(\Omega),\ \forall \vec{w} \in H^1_0(\Omega).
$$

In the framework of the conforming finite element approximation and only in the two-dimensional case (see [32]), we have

$$
(78)\qquad \begin{cases} |\widetilde{c}(\vec{u}; \vec{v}, \vec{w})| \le K_1 \|\vec{u}\|_1 \|\vec{v}\|_1 \|\vec{w}\|_1, \\ |\widetilde{c}(\vec{u}; \vec{v}, \vec{w})| \le K_2 \|\vec{u}\|^{\frac{1}{2}} \|\nabla \vec{u}\|^{\frac{1}{2}} \|\nabla \vec{v}\| \|\vec{w}\|^{\frac{1}{2}} \|\nabla \vec{w}\|^{\frac{1}{2}} \end{cases}
$$

$\forall\ \vec{u}, \vec{v}, \vec{w} \in H^1(\Omega)$.

Next, let $P^h = X^h|_\Gamma$, i.e., P_h consists of the restriction, to the boundary Γ, of functions belonging to X^h. For all choices of conforming finite element space X^h we then have that $P^h \subset H^{-1/2}(\Gamma)$. For the subspaces $P^h = X^h|_\Gamma$, we assume the approximation property: there exists an integer l and a constant C, independent of $h, \vec{s}$ such that for $1 \leq k \leq l$ we have

$$\inf_{\vec{s}_h \in p_h} \|\vec{s}_h - \vec{s}\|_{-1/2,\Gamma} \leq Ch^k \|\vec{u}\|_{k-1/2} \quad \forall \vec{s} \in H^{k-1/2}(\Gamma). \tag{79}$$

Now, let $Q^h = X^h|_{\Gamma_c}$, i.e., Q^h consists of the restriction, to the boundary segment Γ_c, of the functions belonging to X^h. For all choices of conforming finite element spaces X^h, we have that $Q^h \subset H^1(\Gamma_c)$. We define $Q_0^h = Q^h \cap H^1_{n0}(\Gamma_c)$. If the same type of polynomials are used in $Q_0^h = Q^h \cap (\Gamma_c)$ we have the following.

(c) *boundary approximating property:* there exists an integer k and a constant C, independent of $h, \vec{s}$ such that for $1 \leq m \leq k$ we have

$$\inf_{\vec{s}_h \in Q_0^h} \|\vec{s}_h - \vec{s}\|_{s,\Gamma_c} \leq Ch^{m-s+1/2} \|\vec{s}\|_{m+1/2} \quad \forall \vec{s} \in H^1_{n0}(\Gamma_c), \quad 0 \leq s \leq 1. \tag{80}$$

See [3] and [9] for details concerning the approximation on the boundary.

4.2. Formulation of the fully discrete optimal control approximation. Let $\sigma_N = \{t_n\}_{n=0}^N$ be a partition of $[0,T]$ in equal intervals $\Delta t = T/N$ with $t_0 = 0$ and $t_N = T$. For each fixed Δt (or N) and for every quantity $q(t,\vec{x})$, we associate the corresponding set $\{q_h^{(n)}\}_{n=1}^N$. We will denote the vector $(q_h^{(1)}, q_h^{(2)}, \ldots, q_h^{(N)})$ with bold-faced letter $\mathbf{q}_h$ and the space Y^N as $\mathbf{Y}$. The continuous linear function $\vec{q}_h^N(t,\vec{x})$ is defined by $\vec{q}_h^N(t_n,\vec{x}) = q_h(t_n,\vec{x})$ $\forall$ $n = 0,1,2\ldots,N$.

Given $\Delta t = T/N$, $\vec{\mathbf{g}} \in \mathbf{H}^{1/2}(\Gamma)$, and $\vec{u}_0 \in curl(H^2)(\Omega)$, then $(\vec{\mathbf{u}}_h, \mathbf{p}_h)$ is called a *generalized solution* of the fully discrete time-space approximate Navier–Stokes equations if $\vec{u}_h^{(n)} \in X^h$, $p_h^{(n)} \in S_0^h$ and $(\vec{u}_h^{(n)}, p_h^{(n)})$ satisfies the following system of equations:

$$\begin{cases} \frac{1}{\Delta t}(\vec{u}_h^{(n)} - \vec{u}_h^{(n-1)}, \vec{v}_h) + \nu a(\vec{u}_h^{(n)}, \vec{v}_h) \\ \qquad + \tilde{c}(\vec{u}_h^{(n)}; \vec{u}_h^{(n)}, \vec{v}_h) + b(\vec{v}_h, p^{(n)}) = 0 \quad \forall\, \vec{v}_h \in X_0^h, \\ b(\vec{v}_h^{(n)}, q_h) = 0 \quad \forall\, q_h \in S_0^h, \\ (\vec{u}_h^{(n)}, \vec{s})_\Gamma = (\vec{g}_h^{(n)}, \vec{s})_{\Gamma_c} \quad \forall\, \vec{s} \in Q_0^h, \end{cases} \tag{81}$$

for $n = 1, 2, \ldots, N$, with initial velocity $\vec{u}_h^{(0)} = \pi^h \vec{u}_0(\vec{x})$.

The formulation of the problem in the fully discrete approximation becomes the following:

> *given* $\Delta t = T/N$, $\vec{u}_0 \in curl(H^2)(\Omega)$, *and* $\vec{U} \in U_{ad}$, *find* $(\vec{\mathbf{u}}_h, \mathbf{p}_h, \vec{\mathbf{g}}_h)$ *in* $\mathbf{X}^h \times \mathbf{S}_0^h \times \mathbf{Q}_0^h$ *such that* $(\vec{u}_h^{(n)}, p_h^{(n)}, \vec{g}_h^{(n)})$ *is the solution of* (81) *and minimizes the cost function*

$$\begin{aligned} \mathcal{J}_h^N(\vec{\mathbf{u}}_h, \vec{\mathbf{g}}_h) &= \frac{\alpha}{2} \sum_{n=1}^N \|\vec{u}_h^{(n)} - \vec{U}^{(n)}\|^2 \Delta t \\ &+ \frac{\beta}{2} \sum_{n=1}^N [\|\vec{g}_h^{(n)}\|^2_{\Gamma_c} \Delta t + \beta_1 \|\partial_s \vec{g}_h^{(n)}\|^2_{\Gamma_c} \Delta t + \beta_2 \|(\vec{g}_h^{(n)} - \vec{g}_h^{(n-1)})\|^2_{\Gamma_c}], \end{aligned} \tag{82}$$

with $\vec g^{(0)} = \pi_h \gamma \vec u_0$.

In the above definition, the operator π_h approximates the trace of a function in the corresponding finite element space. In the framework of conforming finite elements, the existence can be proved by using the same standard techniques. We state the theorem for completeness.

THEOREM 4.1. *Given* $\Delta t = T/N$, $\vec u_0 \in curl(H^2)(\Omega)$, *and* $\vec U \in U_{ad}$, *there exists a solution* $(\vec{\mathbf{u}}_h, \mathbf{p}_h, \vec{\mathbf{g}}_h)$ *in* $\mathbf{X}^h \times \mathbf{S}_0^h \times \mathbf{Q}_0^h$ *of the fully discrete optimal control problem.*

4.3. First-order necessary condition and the optimality system. We can derive the first-order necessary condition, the Euler equation, and the final characterization for the optimal control. All these results can be obtained proceeding in a manner similar to the semidiscrete case. For conforming finite elements we can state the following theorem.

THEOREM 4.2. *Given* $\Delta t = T/N$ *and* $\vec u_0 \in curl(H^2)(\Omega)$. *Let* $(\vec{\mathbf{u}}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{g}}_h) \in \mathbf{X}^h \times \mathbf{S}_0^h \times \mathbf{Q}_0^h$ *denote an optimal control solution of the discrete optimal control problem. Then, the control* $\vec{\mathbf{g}}_h$ *satisfies the following system:*

$$\int_{\Gamma_c} \Bigg[\vec g_h^{(n)} \cdot \widetilde{\vec r}_h - \frac{\beta_1}{\Delta t^2}(\vec g_h^{(n+1)} - 2\vec g_h^{(n)} + \vec g_h^{(n-1)}) \cdot \widetilde{\vec r}_h + \beta_2 \partial_s \vec g_h^{(n)} \cdot \partial_s \widetilde{\vec r}_h$$
$$+ k^{(n)} \vec n \cdot \widetilde{\vec r}_h - \frac{1}{\beta}(\vec \xi_h^{(n)} \cdot \widetilde{\vec r}_h) \Bigg] \, d\vec x dt = 0 \quad for \quad n = 1, \ldots, N-1,$$

with

$$\int_{\Gamma} \vec g_h^{(n)} \cdot \vec n \, d\vec x = 0 \quad for \quad n = 1, \ldots, N-1,$$
$$\vec g^{(0)} = \gamma_0 \vec u_0 \quad and \quad \vec g^{(N)} = \vec g^{(N-1)} \quad on \quad \Gamma_c,$$

where $\vec{\boldsymbol{\xi}}_h \in \mathbf{P}^h(\Gamma)$ *is defined by*

$$(\vec \xi_h^{(n-1)}, \widetilde{\vec v}_h)_{\Gamma_c} = \left(\frac{\partial \vec w_h^{(n-1)}}{\partial n} - \sigma_h^{(n-1)} \vec n, \widetilde{\vec v}_h \right)_{\Gamma_c}$$
$$= -\alpha(\vec u_h^{(n)} - \vec U^{(n)}, \vec v_h)_\Omega - \frac{1}{\Delta t}(\vec w_h^{(n)} - \vec w_h^{(n-1)}, \widetilde{\vec v}_h)_\Omega$$
$$+ \nu a(\vec w_h^{(n-1)}, \widetilde{\vec v}_h) + \widetilde c(\widetilde{\vec v}_h; \vec u_h^{(n)}, \vec w_h^{(n-1)}) + \widetilde c(\vec u_h^{(n)}; \widetilde{\vec v}_h, \vec w_h^{(n-1)})$$
$$+ b(\widetilde{\vec v}_h, \sigma_h^{(n-1)}) \quad \forall \widetilde{\vec v}_h \in X^h(\Omega), \quad for \quad n = 1, \ldots, N,$$

and $\vec{\mathbf{w}}_h$ *and* $\boldsymbol{\sigma}_h$ *satisfy*

$$\begin{cases} -\frac{1}{\Delta t}(\vec w_h^{(n)} - \vec w_h^{(n-1)}, \vec v_h) + \nu a(\vec w_h^{(n-1)}, \vec v_h) + \widetilde c(\vec v_h; \vec u_h^{(n)}, \vec w_h^{(n-1)}) + \widetilde c(\vec u_h^{(n)}, \vec v_h; \vec w_h^{(n-1)}) \\ \qquad + b(\vec v_h, \sigma_h^{(n-1)}) = \alpha(\vec u_h^{(n)} - \vec U^{(n)}, \vec v_h) \quad \forall \vec v_h \in X_0^h(\Omega), \quad for \quad n = 1, \ldots, N, \\ b(\vec w_h^{(n-1)}, q_h) = 0 \quad \forall q \in S_0^h(\Omega) \quad for \quad n = 1, \ldots, N, \\ \vec w_h^{(n-1)} = 0 \quad on \quad \Gamma \quad for \quad n = 1, \ldots, N, \\ \vec w_h^{(N)} = 0. \end{cases}$$

We remark that an optimal solution is a solution of the above system but among the solutions of this system there may be solutions that are not optimal.

5. Implementation of fully discrete space-time approximations. In our numerical tests we consider only tangential control. This enables us to dispense with the unknown function $k^{(n)}$ in the optimality system. The equations for tangential control consist of

(a) the Navier–Stokes system

$$(83)\qquad \begin{cases} \frac{1}{\Delta t}(\vec{u}_h^{(n)} - \vec{u}_h^{(n-1)}, \vec{v}_h) + \nu a(\vec{u}_h^{(n)}, \vec{v}_h) + \widetilde{c}(\vec{u}_h^{(n)}; \vec{u}_h^{(n)}, \vec{v}_h) \\ \qquad +b(\vec{v}_h, p_h^{(n)}) = 0 \quad \forall \vec{v}_h \in X_0^h(\Omega), \quad \text{for} \quad n = 1, \dots, N, \\ b(\vec{u}_h^{(n)}, q_h) = 0 \quad \forall q_h \in S_0^h(\Omega), \quad \text{for} \quad n = 1, \dots, N, \\ (\vec{u}_h^{(n)}, \vec{s}_h)_\Gamma = (\vec{g}_h^{(n)}, \vec{s}_h)_{\Gamma_c} \quad \forall \vec{s}_h \in P^h(\Gamma), \quad \text{for} \quad n = 1, \dots, N, \\ \vec{w}_h^{(0)} = \pi^h \vec{u}_0 \quad \text{in} \quad \Omega; \end{cases}$$

(b) the adjoint system

$$(84)\qquad \begin{cases} -\frac{1}{\Delta t}(\vec{w}_h^{(n)} - \vec{w}_h^{(n-1)}, \vec{v}_h) + \nu a(\vec{w}_h^{(n-1)}, \vec{v}_h) + \widetilde{c}(\vec{u}_h^{(n)}; \vec{v}_h, \vec{w}_h^{(n-1)}) \\ \qquad +\widetilde{c}(\vec{v}_h; \vec{u}_h^{(n)}, \vec{w}_h^{(n-1)}) + b(\vec{v}_h, \sigma_h^{(n-1)}) = \alpha(\vec{u}_h^{(n)} - \vec{U}^{(n)}, \vec{v}_h) \\ \qquad\qquad \forall \vec{v}_h \in X_0^h(\Omega), \quad \text{for} \quad n = 1, \dots, N, \\ b(\vec{w}_h^{(n-1)}, q_h) = 0 \quad \forall q_h \in S_0^h(\Omega), \quad \text{for} \quad n = 1, \dots, N, \\ \vec{w}_h^{(n-1)} = 0 \quad \text{on} \quad \Gamma, \quad \text{for} \quad n = 1, \dots, N, \\ \vec{w}_h^{(N)} = 0 \quad \text{in} \quad \Omega; \end{cases}$$

(c) the control equation

$$(85)\qquad \begin{cases} -\frac{\beta_1}{\Delta t^2}(\vec{\lambda}_h^{(n+1)} - 2\vec{\lambda}_h^{(n)} + \vec{\lambda}_h^{(n-1)}, \vec{r}_h) + \beta_2(\partial_s \vec{\lambda}_h^{(n)}, \partial_s \vec{r}_h) + (\vec{\lambda}_h^{(n)}, \vec{r}_h) \\ \qquad = \frac{1}{\beta}(\frac{\partial \vec{w}^{(n)}}{\partial n}, \vec{r}_h) \quad \forall \vec{r}_h \in Q_0^h(\Gamma_c), \quad \text{for} \quad n = 1, \dots, N-1, \\ \vec{\lambda}_h^{(n)} = 0 \quad \text{on} \quad \Gamma \setminus \Gamma_c, \quad \text{for} \quad n = 1, \dots, N-1, \\ \vec{\lambda}_h^{(0)} = \pi_h \gamma_0 \vec{u}_0 \quad \text{on} \quad \Gamma_c, \\ \vec{\lambda}_h^{(N)} = \vec{\lambda}_h^{(N-1)} \quad \text{on} \quad \Gamma_c; \end{cases}$$

(d) the compatibility boundary equation

$$(86)\qquad \vec{\lambda}^{(n)} = \vec{g}^{(n)} \quad \text{for} \quad n = 1, \dots, N,$$

where it is understood that $\vec{g}^{(n)}$ is a tangential vector.

In order to solve this numerical system, let us consider the gradient method for the optimal control problem. We have to split the system into three parts in order to apply the algorithm: the Navier–Stokes system (83), the adjoint system (84), and the control equation (85). In the gradient algorithm, we satisfy the relation in (86) only when convergence is achieved. Let $\mathcal{J}_h^N(k) = \mathcal{J}_h^N(\vec{\mathbf{u}}_h(k), \vec{\mathbf{g}}_h(k))$ and τ be the tolerance required for the convergence of the functional.

The *gradient algorithm* we study proceeds as follows:

```
(a) initialization:
    (i) given $\vec{\mathbf{g}}_h(0), \tau$, and $\epsilon = 1$;
    (ii) solve (83) with $\vec{\mathbf{g}}_h(0)$ for $\vec{\mathbf{u}}_h(0)$;
    (iii) evaluate $\mathcal{J}_h^N(0)$;
(b) main loop:
    (iv) solve (84) with $\vec{\mathbf{u}}_h(k-1)$ for $\vec{\mathbf{w}}_h(k)$;
    (v) solve (85) with $\vec{\mathbf{w}}_h(k)$ for $\vec{\boldsymbol{\lambda}}_h(k)$;
    (vi) set $\vec{g}_h^{(n)}(k) = \vec{g}_h^{(n)}(k-1) - \epsilon(\vec{g}_h^{(n)}(k-1) - \vec{\lambda}_h^{(n)}(k))$
    (vii) solve (83) for $\vec{\mathbf{u}}_h(k)$;
    (viii) evaluate $\mathcal{J}_h^N(k)$;
    (ix) if $\mathcal{J}_h^N(k) \le \mathcal{J}_h^N(k-1)$, set $\epsilon = 1.5\epsilon$ and go to (iv);
   if $\mathcal{J}_h^N(k) > \mathcal{J}_h^N(k-1)$, set $\epsilon = 0.5\epsilon$ and go to (vi).
```

The algorithm stops when $|\mathcal{J}_h^N(k) - \mathcal{J}_h^N(k-1)|/\mathcal{J}_h^N(k) \le \tau$. We can show that this gradient algorithm converges to a solution of the discrete optimality system.

THEOREM 5.1. *Let* $(\vec{\mathbf{u}}_h(k), \vec{\mathbf{w}}_h(k), \mathbf{p}_h(k), \boldsymbol{\sigma}_h(k), \vec{\mathbf{g}}_h(k))$ *be the kth iterate of the gradient algorithm and let* $(\vec{\mathbf{u}}_h, \vec{\mathbf{w}}_h, \mathbf{p}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{g}}_h)$ *denote a solution of the fully discrete optimality system* (83)–(85). *Then, for* Δt *sufficiently small, there exists a ball* $\mathbf{B} \in \mathbf{Q}_0^h$*, whose radius depends on the ratio* α/β*, such that, if* $\vec{\mathbf{g}}_h(0) \in \mathbf{B}$*, the solution of the gradient algorithm converges to* $(\vec{\mathbf{u}}_h, \vec{\mathbf{w}}_h, \mathbf{p}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{g}}_h)$ *as* $k \to \infty$.

Proof. The theorem follows if we make use of this classical result on the gradient algorithm. Let X be a Hilbert space with norm $\|\cdot\|$ and let $\mathcal{J}(\cdot)$ be a real-valued functional on X. Suppose that $\mathcal{J}(\cdot)$ is of class C^2; suppose that $\widehat{x}$ is a local minimizer of $\mathcal{J}(\cdot)$; suppose that there exists a ball B of X, centered at $\widehat{x}$, such that there exist two real numbers c_1 and c_2 such that $\forall\ \widetilde{g} \in B$ and $\forall\ \delta x_1, \delta x_2 \in X$

$$(87) \quad \mathcal{J}''(\widetilde{x})(\delta x_1, \delta x_2) \le c_1 \|\delta x_1\| \|\delta x_2\| \qquad \text{and} \qquad c_2 \|\delta x_1\|^2 \le \mathcal{J}''(\widetilde{x})(\delta x_1, \delta x_1),$$

where $\mathcal{J}''(\widetilde{x})(\delta x_1, \delta x_2)$ is the bilinear form associated with the second derivatives of $\mathcal{J}(\cdot)$; and suppose that ρ_k is chosen so that

$$(88) \qquad 0 < a \le \rho_k \le b < \frac{2c_2}{c_1} \qquad \forall\ k$$

for some positive numbers a and b. Then, the iterates of the gradient algorithm

$$x(k+1) = x(k) - \rho_k \nabla \mathcal{J}(x(k)), \quad k = 0, 1, 2 \ldots,$$

converge to $\widehat{x}$ for any initial iterate $x(0) \in B$. For details and proof see, e.g., [10] or [20].

In order to bound the second variation of the functional, we follows the technique used in [22, Thm. 5.21]. Here the norm $\|\vec{\mathbf{q}}_h\|_i$ is defined by $\|\vec{\mathbf{q}}_h\|_i^2 = \sum_{n=1}^N \Delta t \|\vec{q}_h^{(n)}\|_i^2$ and $\|\vec{\mathbf{q}}_h\|_{1,1,\Gamma_c}$ by $\|\vec{\mathbf{q}}_h\|_{1,1,\Gamma_c} = \|\vec{\mathbf{q}}_h\|_{1,\Gamma_c} + \sum_{n=1}^N \|\vec{q}_h^{(n)} - \vec{q}_h^{(n-1)}\|_{\Gamma_c}$. For each $\widetilde{\mathbf{g}}_h \in \mathbf{Q}_0^h$, the nonvanishing terms of the second variation of $\mathcal{J}_h^N(\widetilde{\mathbf{u}}_h(\widetilde{\mathbf{g}}_h), \widetilde{\mathbf{g}}_h)$ are given by

$$\frac{D^2 \mathcal{J}_h^N(\vec{\mathbf{u}}_h, \vec{\mathbf{g}}_h)}{D\vec{\mathbf{g}}_{1h} D\vec{\mathbf{g}}_{2h}} \cdot \delta\vec{\mathbf{g}}_{1h} \cdot \delta\vec{\mathbf{g}}_{2h}$$
$$= \alpha \sum_{n=1}^N \int_\Omega \left(\widetilde{\vec{w}}_{1h}^{(n)} \cdot \widetilde{\vec{w}}_{2h}^{(n)} + (\vec{u}_h^{(n)} - \vec{U}_h^{(n)}) \cdot \widetilde{\vec{z}}_h^{(n)} \right) d\vec{x}\, \Delta t$$

$$
\begin{aligned}
&+\beta\sum_{n=1}^{N}\int_{\Gamma_c}\Big[\vec{g}_{1h}^{(n)}\delta\vec{g}_{2h}^{(n)}+\beta_2\partial_s\delta\vec{g}_{1h}^{(n)}\partial_s\delta\vec{g}_{2h}^{(n)}\\
&\qquad+\frac{\beta_1}{\Delta t^2}(\delta\vec{g}_{1h}^{(n)}-\delta\vec{g}_{1h}^{(n-1)})(\delta\vec{g}_{2h}^{(n)}-\delta\vec{g}_{2h}^{(n-1)})\Big]d\vec{x}\,\Delta t,
\end{aligned}
$$

where the first variations $\widetilde{w}_{1h}^{(n)}\in X^h$ and $\widetilde{w}_{2h}^{(n)}\in X^h$ are solutions of $\widetilde{w}_{ih}^{(0)}=\vec{0}$ and

$$
(89)\qquad\left\{
\begin{aligned}
&\frac{1}{\Delta t}(\widetilde{w}_{ih}^{(n)}-\widetilde{w}_{ih}^{(n-1)},\vec{v}_h)+\nu a(\widetilde{w}_{ih}^{(n)},\vec{v}_h)+\widetilde{c}(\widetilde{w}_{ih}^{(n)};\vec{u}_h^{(n)},\vec{v}_h)\\
&\qquad+\widetilde{c}(\vec{u}_h^{(n)};\widetilde{w}_{ih}^{(n)},\vec{v}_h)+b(\vec{v}_h,\widetilde{r}_{ih}^{(n)})=(\delta\vec{g}_{ih}^{(n)},\vec{v}_h)\quad\forall\,\vec{v}_h\in X_0^h(\Omega),\\
&b(\widetilde{w}_{ih}^{(n)},q_h)=0\quad\forall\,q_h\in S_0^h(\Omega),\\
&(\widetilde{w}_{ih}^{(n)},\vec{s})_\Gamma=(\delta\vec{g}_{ih}^{(n)},\vec{s})_{\Gamma_c}\quad\forall\vec{s}\in H^{-1/2}(\Gamma),
\end{aligned}
\right.
$$

for $n=1,2,\ldots,N$ and for $i=1,2$, respectively, and the second variation $\widetilde{z}_h^{(n)}\in X_0^h$ is the solution of $\widetilde{z}_h^{(0)}=\vec{0}$ and

$$
(90)\left\{
\begin{aligned}
&\frac{1}{\Delta t}(\widetilde{z}_h^{(n)}-\widetilde{z}_h^{(n-1)},\vec{v}_h)+\nu a(\widetilde{z}_h^{(n)},\vec{v}_h)+\widetilde{c}(\widetilde{z}_h^{(n)};\widetilde{u}_h^{(n)},\vec{v}_h)+\widetilde{c}(\widetilde{u}_h^{(n)};\widetilde{z}_h^{(n)},\vec{v}_h)\\
&\qquad+b(\vec{v}_h,\widetilde{s}_h^{(n)})=-\widetilde{c}(\widetilde{w}_{1h}^{(n)};\widetilde{w}_{2h}^{(n)},\vec{v}_h)-\widetilde{c}(\widetilde{w}_{2h}^{(n)};\widetilde{w}_{1h}^{(n)},\vec{v}_h)\quad\forall\,\vec{v}_h\in X_0^h(\Omega),\\
&b(\widetilde{z}_h^{(n)},q_h)=0\quad\forall\,q_h\in S_0^h(\Omega),
\end{aligned}
\right.
$$

for $n=1,2,\ldots,N$.

Let $\vec{w}_h^{(n)}\in X_0^h$ be the solution of the adjoint system $\widetilde{w}_h^{(N)}=0$ and

$$
(91)\qquad\left\{
\begin{aligned}
&-\frac{1}{\Delta t}(\vec{w}_h^{(n)}-\vec{w}_h^{(n-1)},\vec{v}_h)+\nu a(\vec{v}_h,\vec{w}_h^{(n-1)})+\widetilde{c}(\vec{v}_h;\vec{u}_h^{(n)},\vec{w}_h^{(n-1)})\\
&\quad+\widetilde{c}(\vec{u}_h^{(n)};\vec{v}_h,\vec{w}_h^{(n-1)})+b(\vec{v}_h,r_h^{(n-1)})\\
&\quad=\ \alpha(\vec{u}_h^{(n)}-\vec{U}^{(n)},\vec{v}_h)\quad\forall\vec{v}_h\in X_0^h(\Omega),\\
&b(\vec{w}_h^{(n-1)},q_h)=0\quad\forall q_h\in S_0^h(\Omega),
\end{aligned}
\right.
$$

for $n=1,2,\ldots,N$.

By following arguments similar to those used in [22], we have

$$
\begin{aligned}
&\alpha\Delta t\sum_{n=1}^{N}\int_\Omega(\vec{u}^{(n)}-\vec{U}^{(n)})\cdot\widetilde{z}_h^{(n)}\\
&\qquad\qquad=\ -\Delta t\sum_{i=1}^{N}\Big(\widetilde{c}(\widetilde{w}_{1h}^{(n)};\widetilde{w}_{2h}^{(n)},\vec{w}_h^{(n-1)})+\widetilde{c}(\widetilde{w}_{2h}^{(n)};\widetilde{w}_{1h}^{(n)},\vec{w}_h^{(n-1)})\Big)
\end{aligned}
$$

and the estimates

$$\|\vec{\mathbf{w}}_h\|_1\le\alpha f_2(\|\vec{\mathbf{g}}_h-\widetilde{\mathbf{g}}_h\|_{1,\Gamma_c}),$$

$$\|\widetilde{\mathbf{w}}_h\|_1\le f_1(\|\vec{\mathbf{g}}_h-\widetilde{\mathbf{g}}_h\|_{1,\Gamma_c})\|\delta\vec{\mathbf{g}}_h\|_{1,1,\Gamma_c},$$

where $f_1(\cdot)$ and $f_1(\cdot)$ are continuous functions. Therefore, for some constants $C_1,C_2>0$, we have

$$
\begin{aligned}
&|D^2\mathcal{J}_h^N(\widetilde{\mathbf{u}}_h(\widetilde{\mathbf{g}}_h),\widetilde{\mathbf{g}}_h)(\delta\vec{\mathbf{g}}_{1h},\delta\vec{\mathbf{g}}_{2h})|\\
&\le\ \Big(3\beta\beta_{max}+\big(\alpha+\alpha C_1f_2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|_{1,\Gamma_c})\big)f_1^2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|_{1,\Gamma_c})\Big)\|\delta\vec{\mathbf{g}}_{1h}\|_{1,1,\Gamma_c}\|\delta\vec{\mathbf{g}}_{2h}\|_{1,1,\Gamma_c}
\end{aligned}
$$

and

$$
\begin{aligned}
&|D^2\mathcal{J}_h^N(\widetilde{\mathbf{u}}_h(\widetilde{\mathbf{g}}_h),\widetilde{\mathbf{g}}_h)(\delta\vec{\mathbf{g}}_{1h},\delta\vec{\mathbf{g}}_{1h})|\\
&\qquad\geq \Big(\beta\beta_{min} - C_2\alpha f_2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|_{1,\Gamma_c})f_1^2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|_{1,\Gamma_c})\Big)\|\delta\vec{\mathbf{g}}_{1h}\|^2_{1,1,\Gamma_c},
\end{aligned}
$$

where $\beta_{min}=\min\{1,\beta_1,\beta_2\}$ and $\beta_{max}=\max\{1,\beta_1,\beta_2\}$.

Now, assume that $\|\widehat{\mathbf{g}}_h-\widetilde{\mathbf{g}}_h\|_{1,\Gamma_c}\leq\xi$ so that consequently $f_1(\|\widehat{\mathbf{g}}_h-\widetilde{\mathbf{g}}_h\|_{1,\Gamma_c})\leq\xi_1$ and $\alpha f_2(\|\widehat{\mathbf{g}}_h-\widetilde{\mathbf{g}}_h\|_{1,\Gamma_c})\leq\xi_2$ for some $\xi_1,\xi_2\leq\infty$ and small β. Then, we may choose c_1 and c_2 such that

$$
c_1 = 3\beta\beta_{max} + (\alpha C_1\xi_2)\xi_1^2 \qquad\text{and}\qquad c_2=\beta\beta_{min}-\alpha C_2\xi_2\xi_1^2.
$$

It follows that, for small enough values of the ratio α/β, there exists a constant c_1 such that

$$
\frac{D^2\mathcal{J}_h^N(\vec{\mathbf{g}}_h)}{D\vec{\mathbf{g}}_{1h}D\vec{\mathbf{g}}_{2h}}\cdot\delta\vec{\mathbf{g}}_{1h}\cdot\delta\vec{\mathbf{g}}_{2h}\leq c_1\|\delta\vec{g}_{1h}^{(n)}\|^2_{1,1,\Gamma_c}\|\delta\vec{g}_{2h}^{(n)}\|^2_{1,1,\Gamma_c}
$$

and there exists a constant $c_2>0$ such that

$$
\begin{aligned}
&\frac{D^2\mathcal{J}_h^N(\vec{\mathbf{g}}_h)}{D\vec{\mathbf{g}}_{1h}D\vec{\mathbf{g}}_{2h}}\delta\vec{\mathbf{g}}_{1h}\delta\vec{\mathbf{g}}_{1h}=\alpha\sum_{n=1}^{N}\|\vec{w}_{h1}^{(n)}\|^2\Delta t\\
&+\beta\sum_{n=1}^{N}[\|\delta\vec{g}_{1h}^{(n)}\|^2\Delta t+\beta_2\|\partial_s\delta\vec{g}_{1h}^{(n)}\|^2\Delta t+\beta_1\|\vec{g}_{1h}^{(n)}-\vec{g}_{1h}^{(n-1)}\|^2]\geq c_2\|\delta\vec{\mathbf{g}}_{h2}\|^2_{1,1,\Gamma_c}. \qquad\square
\end{aligned}
$$

6. Computational examples.

6.1. Test 1. We consider a unit square domain $(0,1)\times(0,1)\subset\mathbb{R}^2$. We assume that the time interval $[0,T]$ is divided into N equal intervals ($\Delta t=T/N$). The finite element spaces are chosen to be the Taylor–Hood finite element pair with respect to a rectangular mesh, i.e., continuous piecewise biquadratic for the velocity and continuous piecewise bilinear for the pressure. The same polynomials are used for the restriction on the boundary. The mesh size is denoted by h and calculations with varying mesh sizes have been performed. In this first test, the control is the tangential velocity on the boundary and the target flow velocity is defined by

$$
\begin{aligned}
&\phi(t,z)=(1-\cos(2\pi tz))\times(1-z)^2,\\
&U(x,y)=10\frac{d}{dy}\left(\phi(.4,x)\phi(.4,y)\right),\qquad V(x,y)=-10\frac{d}{dx}\left(\phi(.4,x)\phi(.4,y)\right);
\end{aligned}
$$

note that the target velocity field vanishes on the boundary Γ.

Velocity tracking evolution. The initial velocity is chosen to be

$$
u_0(x,y)=-5U(x,y),\qquad v_0(x,y)=-5V(x,y), \tag{92}
$$

i.e., it is a high energy flow rotating in an opposite direction with respect to the target flow. Note that the initial velocity field vanishes on the boundary so that, by the compatibility condition (18), the control $\vec{g}$ vanishes at $t=0$.

In this test, $\Delta t = 0.05$, $1/\nu = 300$, α has been set to 1, β to 0.0001, and $\beta_1 = \beta_2$ to 0.1. The control is the right-side tangential boundary velocity. This evolution is described in Figure 1 with the controlled fluid depicted on the left and the desired flow on the right. All the pictures are normalized by the maximum values. At the beginning, the effect of control is felt in a very limited area near the control boundary because the high energy initial flow prevails in the central part of the domain. Then, the control gains force and progressively reaches the whole domain and finally, the controlled flow reaches the optimal approximation and keeps this stationary configuration. Figure 2 shows the error $\|\vec{u} - \vec{U}\|$ between the controlled flow $\vec{u}$ and the target flow $\vec{U}$. As we can see, the error rapidly goes to a constant value which represents the optimal steady approximation. Figures 3 and 4, respectively, show the corresponding values of the norm of the control $\vec{g}$ and its time derivative as functions of time. The control vanishes at $t = 0$ due to the compatibility condition (18), but then works hard for a short time in order to steer the controlled flow to the desired one; it then decreases and remains flat. Near $t = T$, it is not necessary to drive the flow and a decreasing boundary velocity minimizes the functional. This small change does not affect the norm error in Figure 2 due to the scale of the graph but we can see a small improvement in the match at $t = 4$. We remark that the optimal stationary controlled flow is very different from the target flow. One-sided control is not enough and the result can be considered poor. Smaller values of β cannot help and for a better control we need to extend the extent of the boundary on which control is applied.

Velocity tracking with different values of $\boldsymbol{\beta}$, $\boldsymbol{\beta_1}$, and $\boldsymbol{\beta_2}$. We want to study the effect of changing the form of the functional by changing the parameters β and β_1 ($\beta_2 = \beta_1$). The initial velocity field is set to zero.

In Figure 5, we show the error $\|\vec{u} - \vec{U}\|$ between the controlled flow $\vec{u}$ and the target flow $\vec{U}$ for different values of $\beta_1 = \beta_2$. We have $\beta_1 = \beta_2$ equal to 0.01 (a), 0.1 (b), and 1 (c). The value of β in this computation is held constant at 0.0001. The time step Δt is again 0.05 and $h = 1/16$. We note that changing β_1 has little effect on the optimal solution. Of course, the reduction in the error is a little quicker when β_1 is smaller. The norm of $\vec{g}$ and its derivative are shown in Figures 6 and 7. For low values of β_1, the control is allowed to move quicker and the sensibility of the system is greater. We note that there is a small change in control magnitude but the optimal stationary flow is still approximately the same. Differences can be noted in the nonsteady part of the evolution flow.

For different values of β we do obtain different optimal controls. In Figure 8, we see the error $\|\vec{u} - \vec{U}\|$ between the controlled flow $\vec{u}$ and the target flow $\vec{U}$ for β equal to 0.0001 (a), 0.001 (b), and 0.01 (c). For small values of β, the optimal control solution is limited by the magnitude of the control and its derivatives. For values of $\beta > 0.005$, the control is poor. Generally, a good control involves small values of β (around 0.001). In Figures 9 and 10, we show, respectively, the norm of the control $\vec{g}$ and its time derivative for the corresponding values of β. As expected, the control $\vec{g}$ approaches zero for low values of β.

Different number of controlled sides. Now we examine the effect of applying control on more sides of the flow domain. We try to repeat the evolution problem changing the number of sides. In Figures 11 to 14, we have the norm of the error, the control norm, the derivative norm in time and in space, respectively, for applying control on different numbers of sides. One-sided control is (a), two-sided control is (b), three-sided control is (c), and finally the control on the whole boundary is (d). Of course the match is improving with increasing number of controlled sides. As we can see in Figure 12, the control behaves better and the maximum strength required decreases when the number of controlled sides increases. A picture of the stationary match can be found in Figure 15. The improvement is evident. The tracking time evolution for the four-sided case is reported in Figure 16.

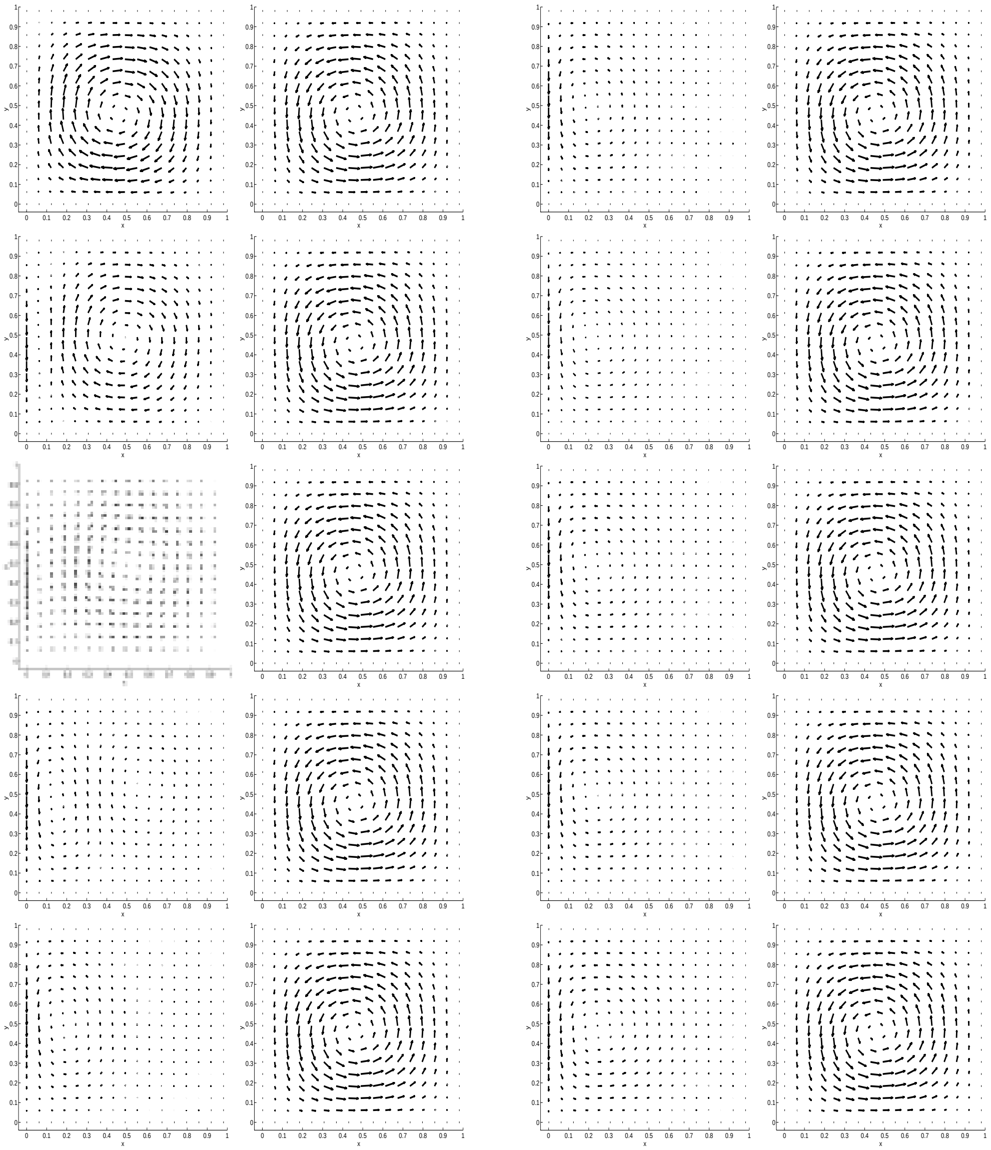


FIG. 1. *Test* 1. *Controlled (left) and target (right) flows at* $t = 0, 0.1, 0.2, 0.4, 0.6$ *(left pair of columns) and* $t = 0.8, 1, 2, 3, 4$ *(right pair of columns) for one-sided control.*

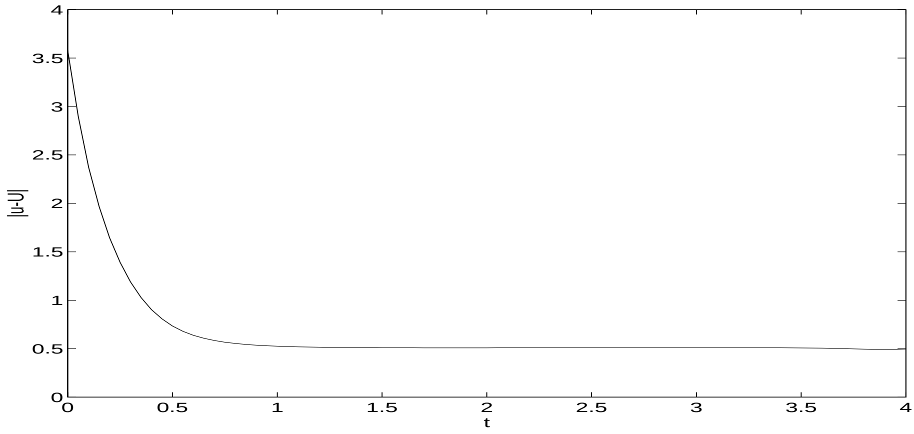


FIG. 2. *Test* 1. *Error* $\|\vec{u} - \vec{U}\|$.

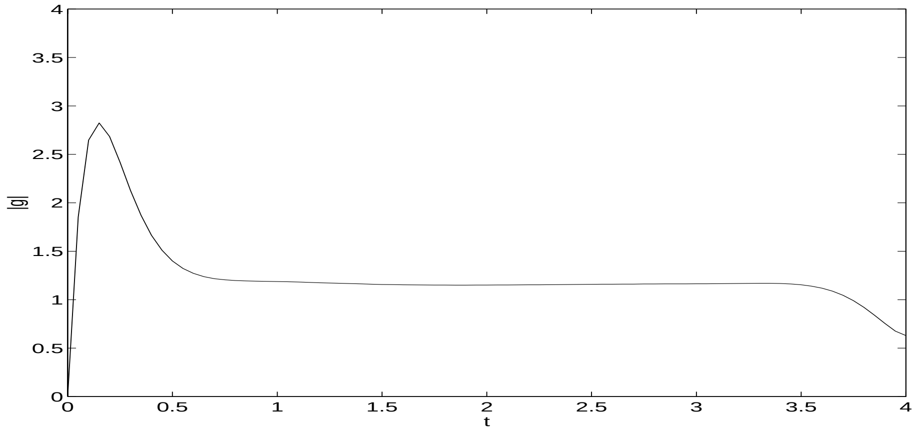


FIG. 3. *Test* 1. *Control norm* $\|g\|$.

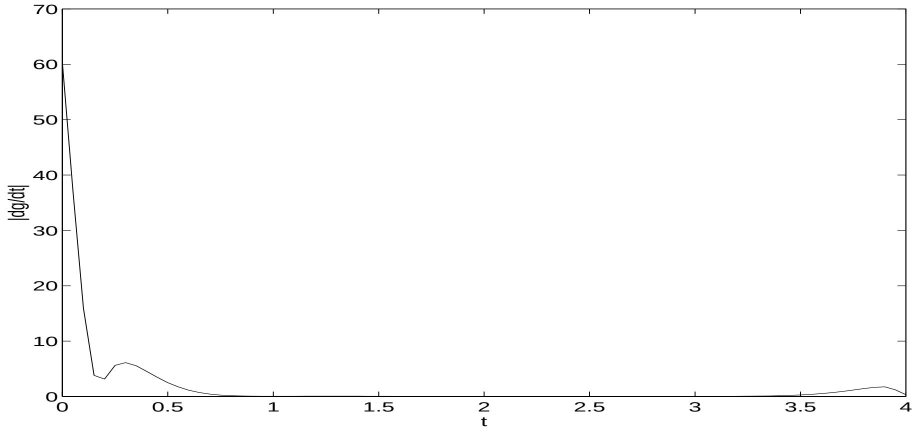


FIG. 4. *Test* 1. *Control norm* $\|\vec{g}_t\|$.

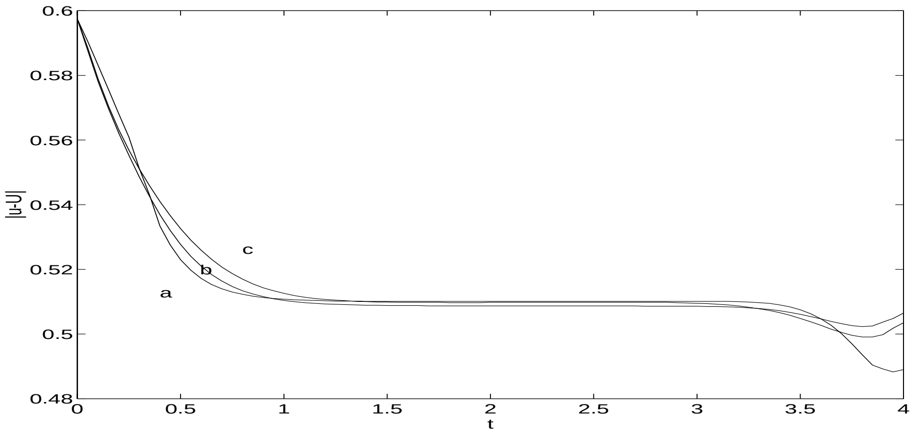


FIG. 5. *Test* 1. *Error for different* β_1.

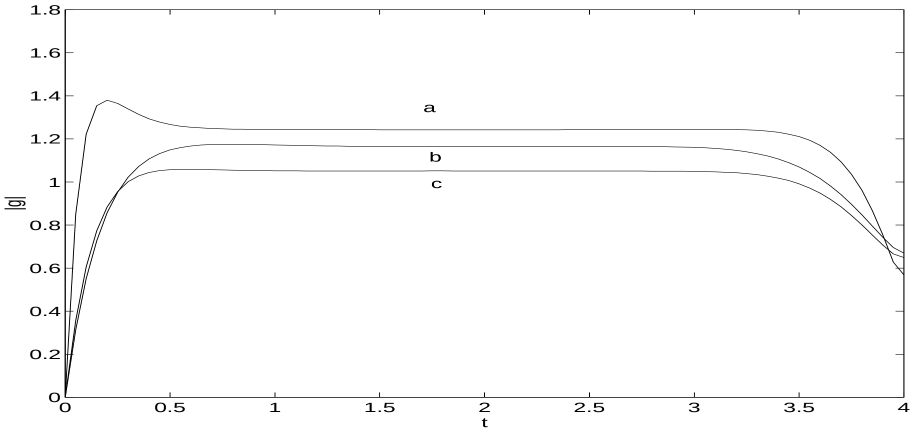


FIG. 6. *Test* 1. *Control norm* $\|g\|$ *for different* β_1.

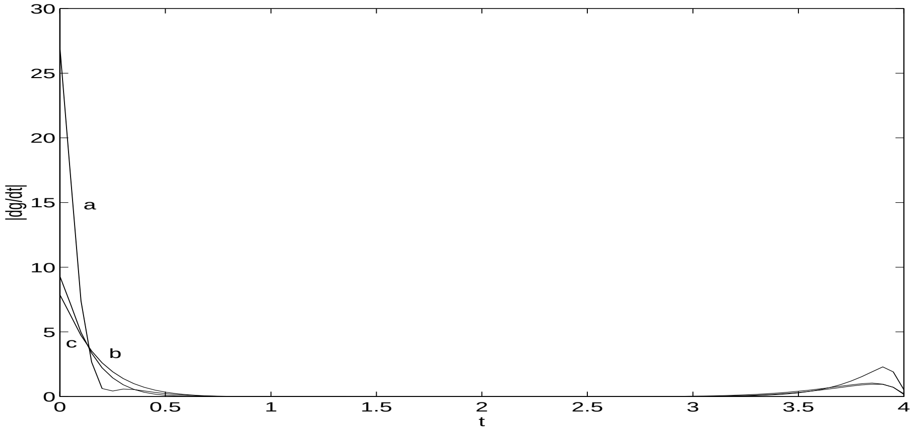


FIG. 7. *Test* 1. *Control norm* $\|g_t\|$ *for different* β_1.

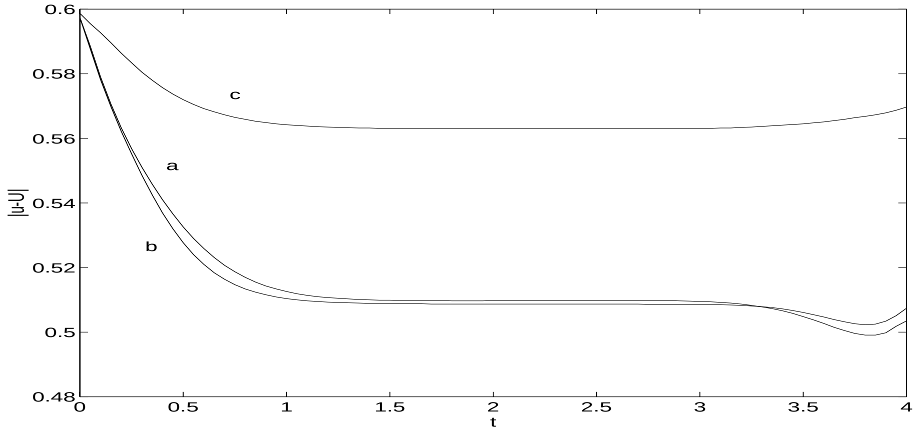


FIG. 8. *Test* 1. *Error* $\|\vec{u} - \vec{U}\|$ *for different* β.

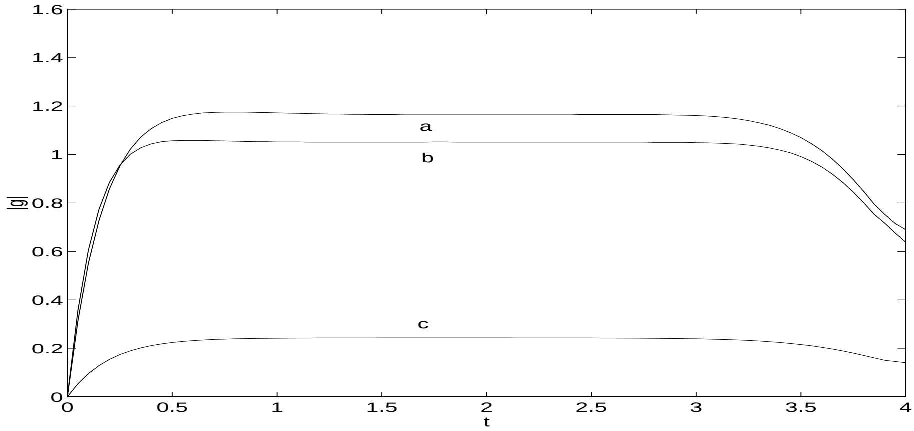


FIG. 9. *Test* 1. *Control norm* $\|g\|$ *for different* β.

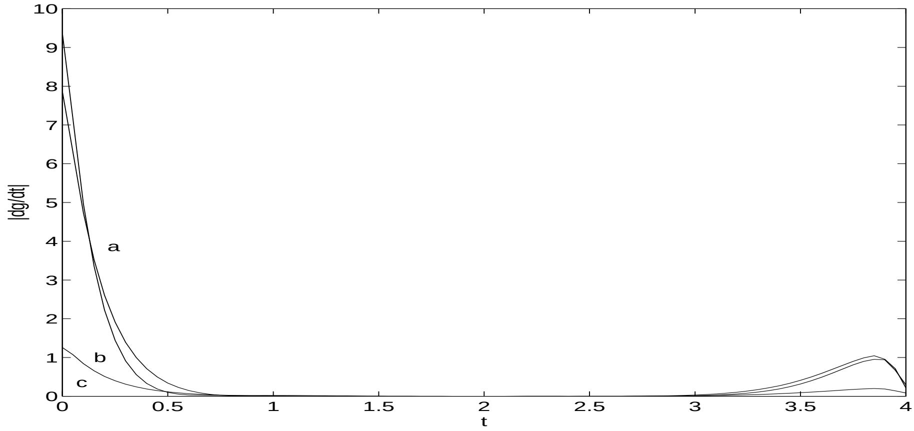


FIG. 10. *Test* 1. *Control norm* $\|g_t\|$ *for different* β.

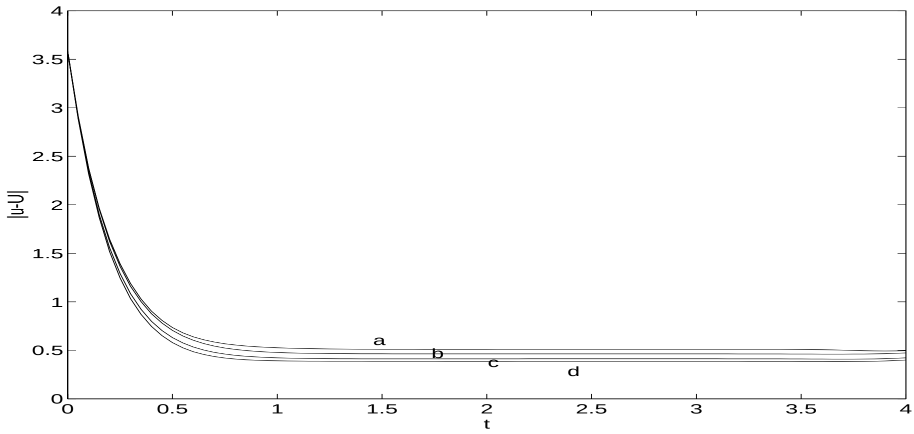


FIG. 11. *Test* 1. *Error* $\|\vec{u}-\vec{U}\|$ *for different number of controlled sides.*

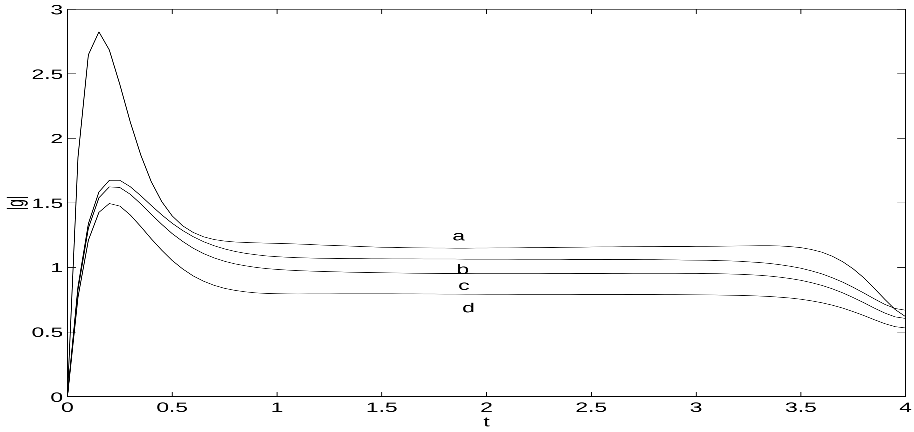


FIG. 12. *Test* 1. *Control norm* $\|g\|$ *for different number of controlled sides.*

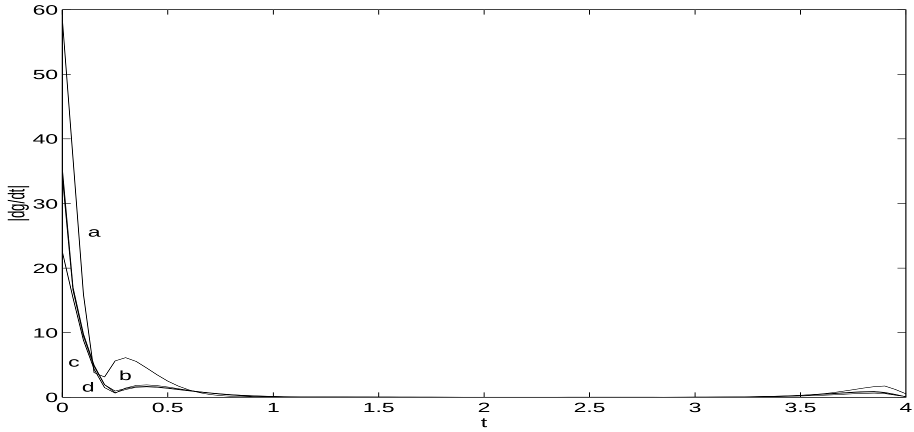


FIG. 13. *Test* 1. *Control norm* $\|\vec{g}_t\|$ *for different number of controlled sides.*

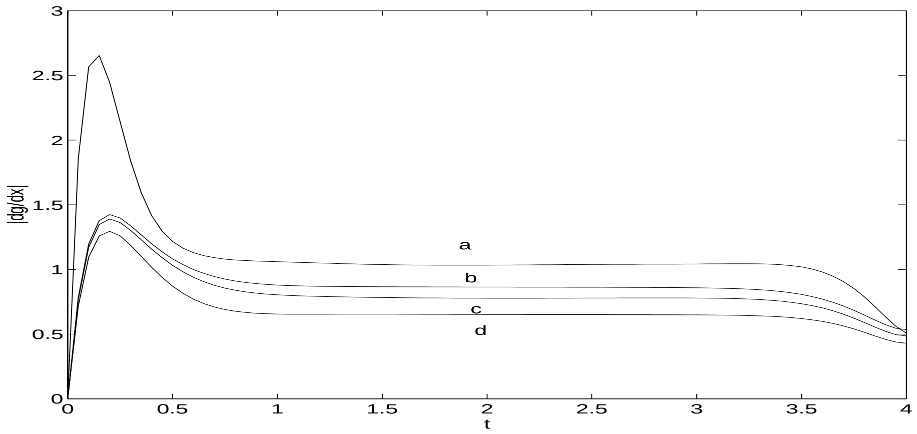


FIG. 14. *Test* 1. *Control norm* $\|g_x\|$ *for different number of controlled sides.*

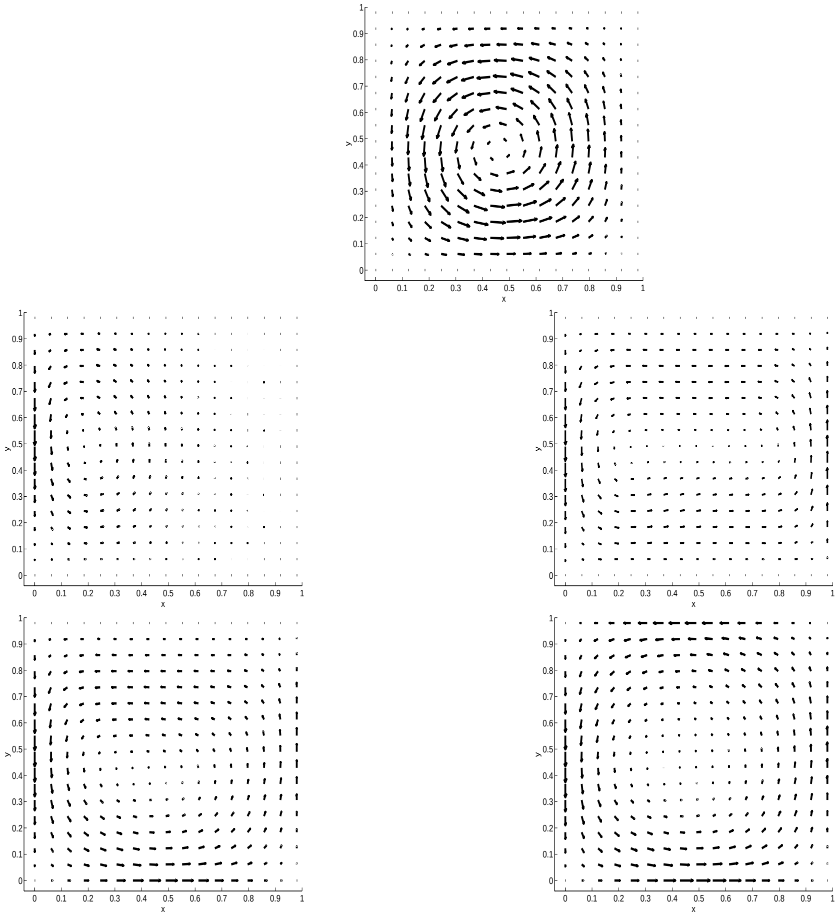


FIG. 15. *Test* 1. *Desired flow (top), stationary one-sided (middle left), two-sided (middle right), three-sided (bottom left), and four-sided (bottom right) controlled flows.*

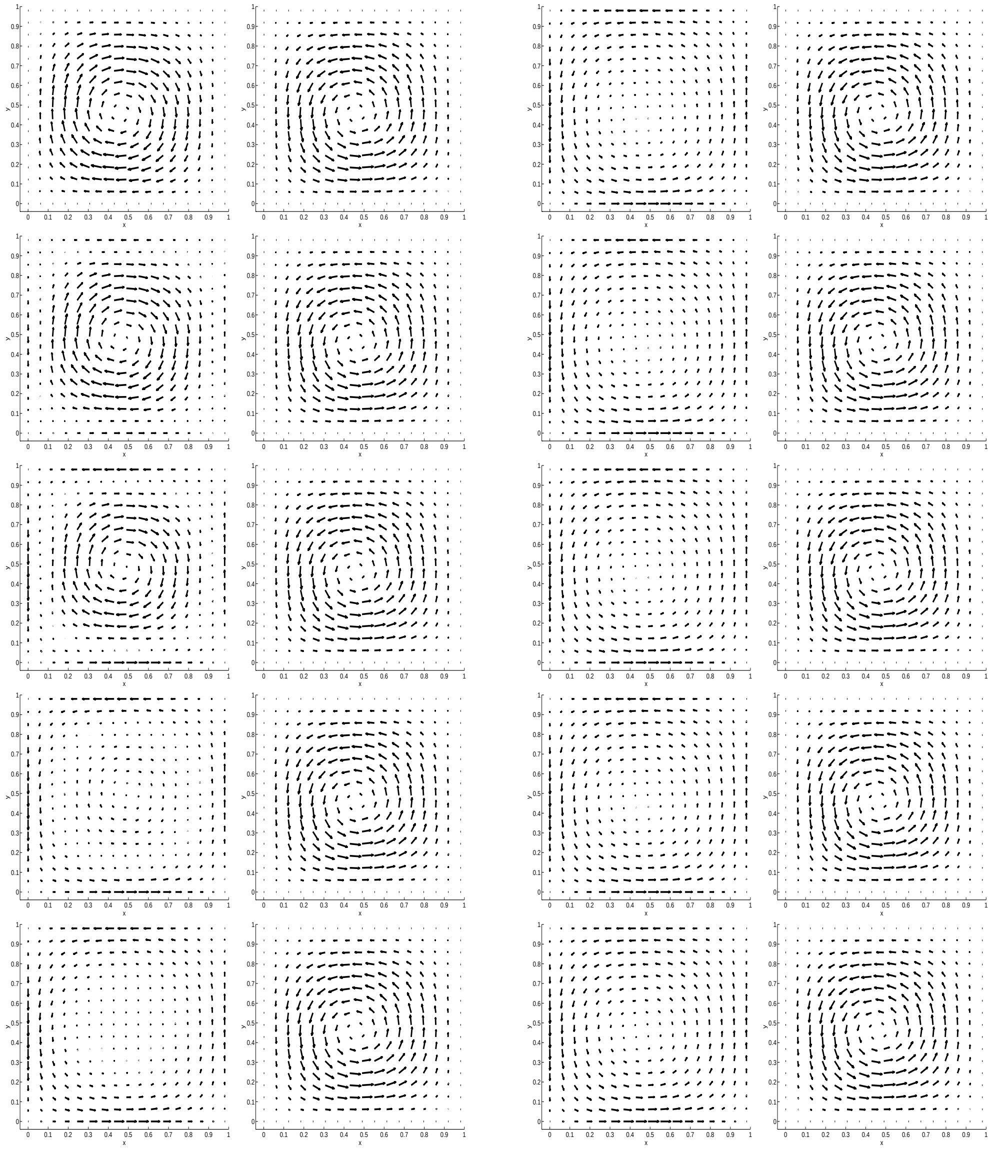


FIG. 16. *Test* 1. *Controlled (left) and target (right) flows at* $t = 0, 0.1, 0.2, 0.4, 0.6$ *(left pair of columns) and* $t = 0.8, 1, 2, 3, 4$ *(right pair of columns) for four-sided control.*

6.2. Test 2. We consider a unit square domain $(0,1) \times (0,1) \subset \mathbb{R}^2$. We assume that the time interval $[0,T]$ is divided in equal intervals of time $\Delta t = T/N$. The Taylor–Hood finite elements are used in this calculation on a rectangular mesh. We report only the final result with $h = 1/16$, but calculations with varying mesh sizes has been performed. The target velocity $\vec{U}$ for this test is time dependent and is given by

$$\phi(k,t,z) = (1 - \cos(2k\pi tz)) \times (1-z)^2 ,$$

$$a(k,t,x,y) = \frac{d}{dy}\left(\phi(k,t,x)\phi(k,t,y)\right), \qquad b(k,t,x,y) = -\frac{d}{dx}\left(\phi(k,t,x)\phi(k,t,y)\right),$$

$$U = a(.25,.4,x,y) + a(.5,t,x,y)/(4\pi t + 1),$$

$$V = b(.25,.4,x,y) + b(.5,t,x,y)/(4\pi t + 1).$$

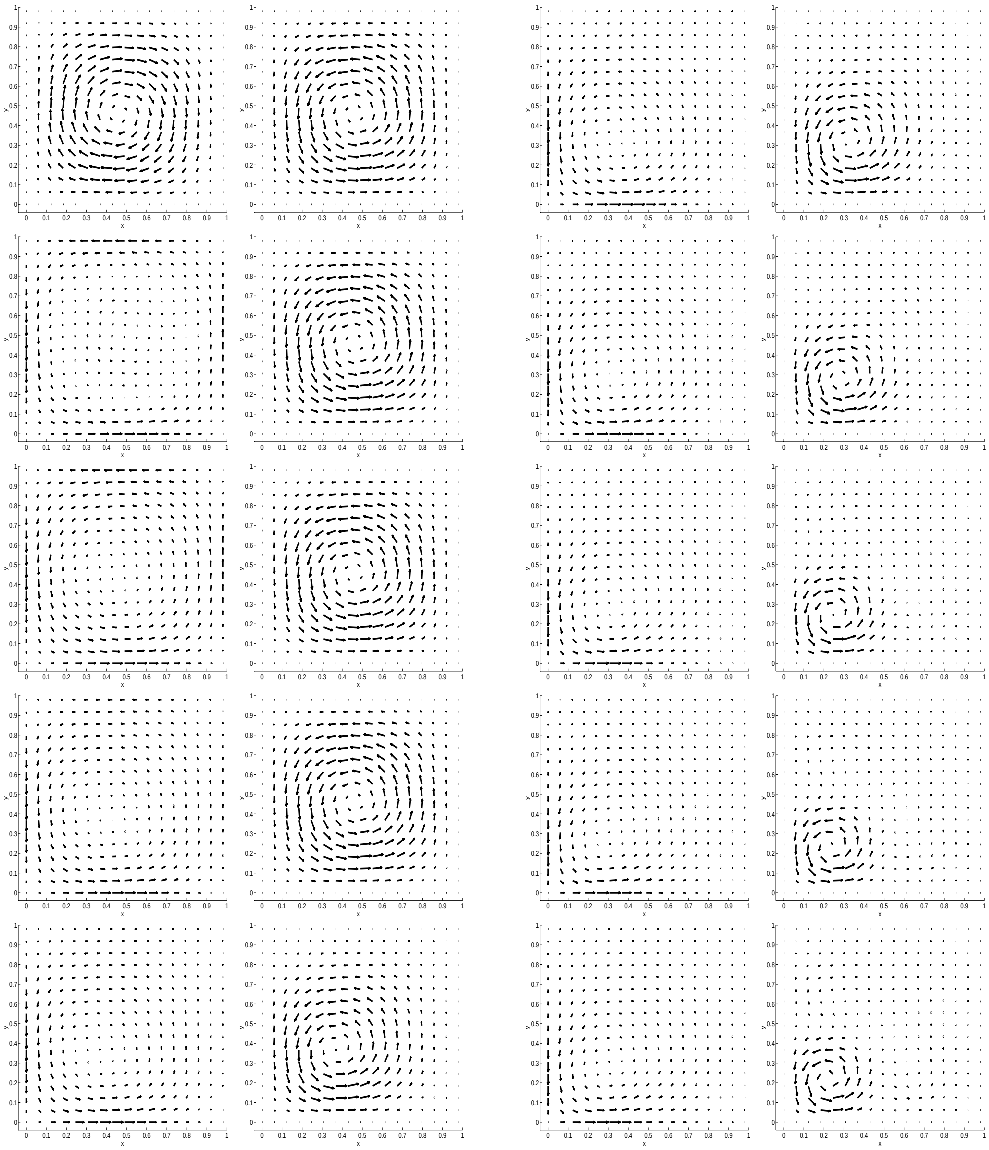


FIG. 17. *Test* 2. *Controlled (left) and target (right) flows at* $t = 0, 0.5, 1, 1.5, 2$ *(left pair of columns) and* $t = 2.5, 3, 3.5, 3.75, 4$ *(right pair of columns) for four-sided control.*

With this velocity field we have the superposition of two flows: one flow with a vortex at the center of the domain and another flow with four vortices. Each of these flows prevails at different times of the evolution. The initial velocity for the controlled flow is

$$u_0(x, y) = -5U(1, x, y), \qquad v_0(x, y) = -5V(1, x, y). \tag{93}$$

The evolution is given in Figure 17. In this computation $\alpha = 1$, $\beta = 0.001$, and $\beta_1 = \beta_2 = 0.1$, and $1/\nu = 300$. The control $\vec{g}$ is a four-sided control and that covers the whole boundary Γ. The controlled fluid is on the left, the desired flow is on the right, and all the pictures are normalized. As we can see at $t = 0.5$, the controlled flow reaches the optimal approximation and follows the motion of the target fluid.

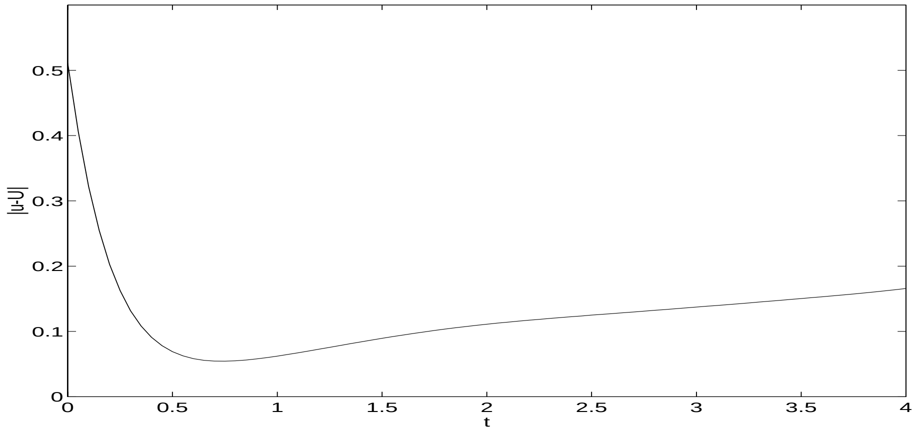


Fig. 18. *Test 2. Error* $\|\vec{u} - \vec{U}\|$.

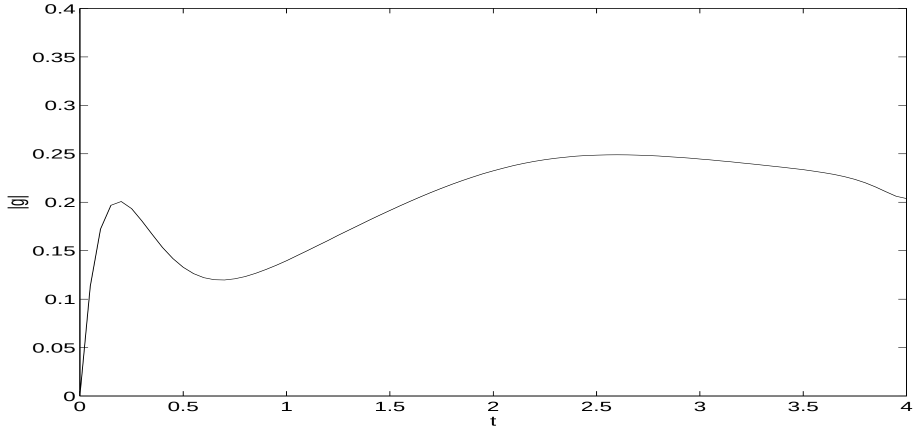


Fig. 19. *Test 2. Control norm* $\|g\|$.

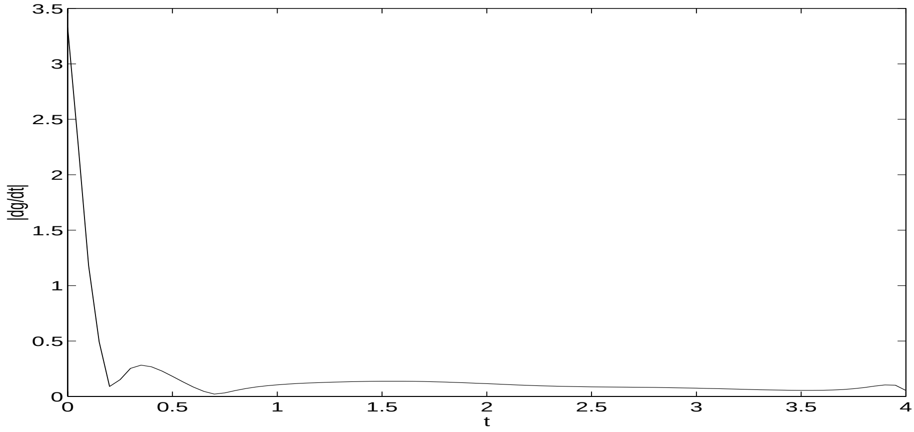


Fig. 20. *Test 2. Control norm* $\|\vec{g}_t\|$.

Figure 18 shows the error $\|\vec{u} - \vec{U}\|$ between the controlled flow $\vec{u}$ and the target flow $\vec{U}$. At the beginning the error rapidly decreases but after this initial interval of time this error increases due to changes in the desired flow. The boundary velocity cannot control the interior of the domain if the desired flow moves rapidly. However, this represents the optimum that can be achieved with the energy available. For the same flow, Figures 19 and 20, respectively, show the values of the norm of the control variable $\vec{g}$ and its time derivative.

REFERENCES

[1] F. Abergel and R. Temam, *On some control problems in fluid mechanics*, Theoretical and Computational Fluid Dynamics, 1 (1990), pp. 303–326.

[2] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[3] I. Babuska, *The finite element method with Lagrangian multipliers*, Numer. Math., 16 (1973), pp. 179–192.

[4] M. Berggren, *Numerical solution of a flow control problem: Vorticity reduction by dynamic boundary action*, SIAM J. Sci. Comput., 19 (1998), pp. 829–860.

[5] M. Berggren, R. Glowinski, and J.-L. Lions, *Controllability Issues for Flow-Related Models: A Computational Approach*, Technical report TR94-47, Rice University, Houston, TX, 1994.

[6] T. Bewley and P. Moin, *Optimal and Robust Approaches for Linear and Nonlinear Regulation Problems in Fluid Mechanics*, AIAA Paper 97-1872, AIAA, New York, 1997.

[7] T. Bewley, R. Temam, and M. Ziane, *A Robust Framework for Robust Control in Fluid Mechanics*, Report, Center for Turbulence Research, Stanford University, Stanford, CA, 1998.

[8] L. Cattabriga, *Su un problema al contorno relativo al sistema di equazioni di Stokes*, Rend. Sem. Mat. Univ. Padova, 31 (1961), pp. 308–340.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[10] P. G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimization*, Cambridge University Press, Cambridge, UK, 1989.

[11] P. Constantin and C. Foias, *Navier-Stokes Equations*, The University of Chicago Press, Chicago, IL, 1989.

[12] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2, Springer-Verlag, New York, 1993.

[13] M. Desai and K. Ito, *Optimal controls of Navier-Stokes equations*, SIAM J. Control Optim., 32 (1994), pp. 1428–1446.

[14] A. Fursikov, *On some control problems and results related to the unique solution of mixed problems by the three-dimensional Navier-Stokes and Euler equations*, Dokl. Akad. Nauk SSSR, 252 (1980), pp. 1066–1070.

[15] A. Fursikov, *Control problems and results on the unique solution of mixed problems by the three-dimensional Navier-Stokes and Euler equations*, Math. Sbornik, 115 (1981), pp. 281–306.

[16] A. Fursikov, M. D. Gunzburger, and L. S. Hou, *Boundary value problems and optimal boundary control for the Navier-Stokes system: the two-dimensional case*, SIAM J. Control Optim., 36 (1998), pp. 852–894.

[17] V. Girault and P. Raviart, *The Finite Element Method for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, New York, 1986.

[18] V. Girault and P. Raviart, *The Finite Element Approximation of the Navier-Stokes Equations*, Springer-Verlag, Berlin, 1979.

[19] G. Grubb, *Solution of parabolic pseudo-differential initial-boundary value problems*, J. Differential Equations, 87 (1990), pp. 256–304.

[20] G. Grubb and V. Solonnikov, *Boundary value problems for the nonstationary Navier-Stokes equations treated by pseudo-differential methods*, Math. Scand., 69 (1991), pp. 217–290.

[21] M. D. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.

[22] M. D. Gunzburger and S. Manservisi, *Analysis and approximation of the velocity tracking problem for Navier–Stokes flows with distributed control,* SIAM J. Numer. Anal., 37 (2000),

pp. 1481–1512.

[23] M. D. Gunzburger and S. Manservisi, *The velocity tracking problem for Navier-Stokes flows with bounded distributed controls*, SIAM J. Control Optim., 37 (1999), pp. 1913–1945.

[24] M. Hinze and K. Kunish, *Control Strategies for Fluid Flows—Optimal Versus Suboptimal Control*, World Science, River Edge, NJ, 1998.

[25] L. Hou and Y. Yan, *Dynamics for controlled Navier-Stokes systems with distributed controls*, SIAM J. Control Optim., 35 (1997), pp. 654–677.

[26] R. Joslin, M. Gunzburger, R. Nicolaides, G. Erlebacher, and M. Hussaini, *A methodology for the automated optimal control of flows including transitional flows*, AIAA J., 35 (1997), pp. 816–824.

[27] P. Koumoutsakos, *Active control of vortex-wall interactions*, Phys. Fluids, 9 (1997), pp. 3808–3816.

[28] J.-L. Lions and E. Magenes, *Problemes aux Limites Non Homogenes et Applications*, Vol. 1, Dunod, Paris, 1968.

[29] S. Manservisi, *Optimal Boundary and Distributed Control of the Time Dependent Navier-Stokes Equations*, Ph.D. thesis, Virginia Tech, Blacksburg, VA, 1997.

[30] B. Protas and A. Stuczek, *Theoretical and Computational Study of the Wake Control Problem*, Preprint, Department of Aerodynamics, Warsaw University of Techology, Warsaw, 1998.

[31] S. Sritharan, *Dynamic programming of Navier-Stokes equations*, Systems Control Lett., 16 (1991), pp. 229–307.

[32] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1979.

[33] R. Temam, T. Bewley, and P. Moin, *Control of turbulent flows*, in Proceedings of the 18th IFIP TC7 Conference on System Modelling and Optimization, Detroit, MI, 1997.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 2, pp. 635–659

OPTIMAL POLICIES FOR n-DIMENSIONAL SINGULAR STOCHASTIC CONTROL PROBLEMS. PART II: THE RADIALLY SYMMETRIC CASE. ERGODIC CONTROL*

LUKASZ KRUK†

Abstract. We consider a singular stochastic control problem with a radially symmetric running cost. We show that the value function is smooth, the nonaction region is a ball, and the problem has an explicit solution in terms of power series. Also, for a singular ergodic control problem with the class of admissible processes constrained to Brownian motions reflected normally at the boundary of some open, connected Caccioppoli set, we show existence, regularity, and basic properties of optimal domains using a geometric measure-theoretic approach.

Key words. singular stochastic control, radial symmetry, optimal policy, normally reflected Brownian motion, geometric measure theory

AMS subject classifications. 93E20, 34A25, 49F22

PII. S0363012998347547

1. Introduction.

The goal of the first part of this article is to minimize

$$u_M(x) = E^x \int_0^\infty e^{-t}[h(X_t)dt + d\xi_t], \tag{1}$$

where $x \in R^n$ is fixed and h is a nonnegative, convex, radially symmetric function (suitably smooth),

$$X_t = x + \sqrt{2}W_t + M_t, \tag{2}$$

with W_t an n-dimensional Brownian motion, M_t any adapted process of bounded variation over finite time-intervals, and ξ_t the variation of M up to time t. This problem, without the assumption of radial symmetry of h, has been investigated thoroughly in the existing literature. It is known that in one (see, e.g., [6, 10, 27, 28, 29, 34, 35, 40]; consult [41] or [14] for more complete references) or two [45] dimensions the corresponding value function u is twice continuously differentiable (smooth fit) and the optimal policy is the Brownian motion reflected at the boundary of the nonaction region $\mathcal{C} = \{x \in R^n : |\nabla u(x)| < 1\}$. In higher dimensions it is only known that the value function is $W^{2,\infty}_{loc}$ and that the optimal policy exists, is unique [42, 46], and solves a generalized Skorokhod problem [36].

Under our additional assumption of radial symmetry, thanks to symmetry of the Brownian motion, the value function u itself is radially symmetric. This allows us to use one-dimensional methods, yielding a simple n-dimensional example of smooth fit and providing an explicit solution in terms of power series under an additional assumption that h is real analytic. Our work generalizes some of the recently published results of [11], where a special case of the value function $h(x) = \lambda|x|^2$ was considered.

*Received by the editors November 19, 1998; accepted for publication (in revised form) May 11, 2000; published electronically September 27, 2000.
http://www.siam.org/journals/sicon/39-2/34754.html

†Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (kruk@cims.nyu.edu). Current address: Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (kruk@asdf4.math.cmu.edu).

In the second part of this article we consider a similar (not necessarily radially symmetric) problem for ergodic control in which we limit the admissible controlled processes to Brownian motions reflected normally at the boundary of some open, connected Caccioppoli set. We show existence, regularity, and some properties of optimal domains like their simple-connectedness, convexity for $n = 2$, and the mean curvature equation satisfied by their boundaries. Singular ergodic control of diffusion processes has been investigated, e.g., in [41] and the references given there. Our approach, however, differs from the results contained in those papers in both constraining the class of admissible controls and the methods used. In the context of optimizing the ergodic cost of controlling the *normally* reflected Brownian motion, regularity theory for domain-dependent functionals and the prescribed mean curvature equation seem to be convenient tools for establishing the desired properties of an optimal region.

2. The radially symmetric case.

2.1. The smooth fit. Let $(W_t, \mathcal{F}_t, t \geq 0)$ be a standard n-dimensional Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, P)$. Let $\{\mathcal{F}_t\}$ be the augmentation of the filtration generated by W (see [33]). For a given $x \in R^n$, let a process X_t be defined by (2), where M_t is a left-continuous process adapted to $\mathcal{F}_t$ such that, for all $T > 0$ P-a.s., the variation of $M_{.}(\omega)$ on the interval $[0, T]$ is finite. We write

$$M_t = \int_0^t N_s d\xi_s, \tag{3}$$

where $|N_t| = 1$ for every $t \geq 0$ almost surely (a.s.) and ξ is nondecreasing and left-continuous. In what follows, we shall always describe M_t as (N_t, ξ_t).

Let $h : R^n \to R$ be a strictly convex, radially symmetric function satisfying, for appropriate positive constants C_0, c_0, and $q > 1$, the following conditions:

$$h \in C^{2,1}(R^n), \tag{4}$$

$$0 \leq h(x) \leq C_0(1 + |x|^q), \tag{5}$$

$$|h(x) - h(x + x')| \leq C_0(1 + h(x) + h(x + x'))^{1-1/q}|x'|, \tag{6}$$

$$h(x + \lambda x') + h(x - \lambda x') - 2h(x) \leq C_0\lambda^2(1 + h(x))^r, r = \left(1 - \frac{2}{q}\right)^+, \tag{7}$$

$$c_0|y|^2 \leq D^2h(x)y \cdot y \tag{8}$$

for all $x, x' \in R^n$, $|x'| \leq 1$, and $0 \leq \lambda \leq 1$ (see [36]). We assume that $\inf_{x \in R^n} h(x) = h(0) = 0$.

For a given control process (N, ξ), we define the corresponding cost by

$$V_{(N,\xi)}(x) = E^x \int_0^\infty e^{-t}[h(X_t)dt + d\xi_t]. \tag{9}$$

The task is to minimize $V_{(N,\xi)}(x)$ in the class of all admissible controls, i.e., to find

$$u(x) = \inf_{N,\xi} V_{(N,\xi)}(x), \tag{10}$$

where (N_t, ξ_t) are as described above. If this minimum is attained for some $(\tilde{N}, \tilde{\xi})$, we say that

$$\nu_t = \int_0^t \tilde{N}_s d\tilde{\xi}_s \tag{11}$$

is an optimal policy (for x).

Recall [42, 45] that the value function u is a unique, nonnegative, convex $W^{2,\infty}_{loc}$ solution of the HJB equation

$$\max(u - \triangle u - h, |\nabla u|^2 - 1) = 0 \tag{12}$$

in R^n.

Moreover, because h is radially symmetric, the value function u also enjoys this property, thanks to the following simple lemma.

LEMMA 2.1. *Let O be any $n \times n$ orthogonal matrix (i.e., a matrix of an isometry in R^n), and let h be such that $h \circ O \equiv h$. Then $u \circ O \equiv u$.*

Proof. Let $v = u \circ O$. Then, by (12), spherical symmetry of the Laplacian, and $h \circ O \equiv h$, for any $x \in R^n$ we have

$$\begin{aligned} &\max(v(x) - \triangle v(x) - h(x), |\nabla v(x)|^2 - 1) \\ &\quad = \max(u(Ox) - \triangle u(Ox) - h(Ox), |\nabla u(Ox)|^2 - 1) = 0, \end{aligned}$$

so v also solves (12). u is a $W^{2,\infty}_{loc}$ nonnegative, convex function, so v also enjoys this property. But (12) admits a unique $W^{2,\infty}_{loc}$ nonnegative, convex solution, so $v \equiv u$ and the proof is complete.

In particular, if O is a matrix of any rotation in R^n (i.e., a rotation about some hyperplane in R^n of codimension 2 containing 0) and $h \circ O \equiv h$, then $u \circ O \equiv u$. This easily yields radial symmetry of u if h is radially symmetric. In particular, in this case $\mathcal{C}$ is radially symmetric. As was noticed in [14, pp. 332–333], the argument given in [45] shows that in any dimension $\mathcal{C}$ is connected. Also, by [45], $\mathcal{C}$ is bounded and contains the unique minimizer of u, which in our case is clearly 0 (otherwise we get a contradiction with convexity and symmetry of u). Thus, $\mathcal{C}$ is a ball $B_R(0)$ with center 0 and radius $R > 0$. Also, by spherical symmetry of u, the vector field ∇u on $\partial\mathcal{C}$ is perpendicular to $\partial\mathcal{C}$, so in $B_R(0)$ u solves

$$u - \triangle u = h \tag{13}$$

with the Neumann boundary condition

$$\frac{\partial u}{\partial \eta} = 1 \tag{14}$$

on $\partial B_R(0)$, where η denotes the outer normal to $\partial B_R(0)$, or with the Dirichlet boundary condition

$$u = c = \text{const} \tag{15}$$

on $\partial B_R(0)$, where c is the value u takes on $\partial B_R(0)$. Thus, by the well-known elliptic regularity theory (see, e.g., [22]), if $h \in C^{k,\alpha}(R^n)$ for any $\alpha \in (0,1)$, then $u \in C^{k+2,\alpha}(\overline{B}_R(0))$. Let

$$w(x) = |\nabla u(x)|^2. \tag{16}$$

Corollary 5.3 in [45] shows that $w \in C^{1,\alpha}(R^n)$ for every $\alpha \in (0,1)$. But in our case u is radially symmetric, so $w = (\frac{\partial u}{\partial r})^2$. Thus, at least away from 0, where ∇u vanishes,

$$\frac{\partial u}{\partial r} = \sqrt{w} \in C^{1,\alpha} \tag{17}$$

for $r = |x|$. In particular, u, considered as a function of r only, is $C^{2,\alpha}$ away from 0 for every $\alpha \in (0,1)$. We have arrived at the following theorem.

THEOREM 2.2. *If, in the control problem* (9)–(10)*, we assume that h is radially symmetric, then $u \in C^{2,\alpha}$ for every $\alpha \in (0,1)$ and $\mathcal{C} = B_R(0)$ for some $R > 0$. Moreover, if $h \in C^{k,\alpha}(R^n)$ for some $k \geq 2$ and $\alpha \in (0,1)$, then $u \in C^{k+2,\alpha}$ in $\overline{B}_R(0)$.*

Of course, in this case the optimal policy makes X_t a Brownian motion reflected at $\partial B_R(0)$ in the inner normal direction, with a possible initial jump to the closest point of $B_R(0)$ if the process starts outside $B_R(0)$.

2.2. The O.D.E. approach. As we have just shown, for a radially symmetric h the value function u is "essentially one-dimensional." Thus, it seems natural to solve our problem for u using the O.D.E. approach similar to that used for the 1-d control problem (see, e.g., [6], [12], [13]). Let $u(x) = v(r)$, where $r = |x|$, R as in Theorem 2.2. For $r > R$

$$\frac{dv}{dr}(r) \equiv 1 \tag{18}$$

and, by the well-known form of the Laplacian in polar coordinates (see, e.g., [31]) and the smooth fit, v satisfies

$$v(r) - \frac{d^2v}{dr^2}(r) - \frac{n-1}{r}\frac{dv}{dr}(r) - h(r) = 0 \tag{19}$$

in $(0, R)$ with boundary conditions

$$\frac{dv}{dr}(0) = 0 \tag{20}$$

(by the fact that u has a minimum at 0) and

$$\frac{dv}{dr}(R) = 1 \tag{21}$$

and the "smooth fit" condition

$$\frac{d^2v}{dr^2}(R) = 0 \tag{22}$$

resulting from (18) and the fact that u (hence v) $\in C^2$. (We are, in fact, slightly abusing the notation by using the same symbol h for the running cost, defined on R^n, and the function of one variable r equal to $h(x)$ for $|x| = r$, but it should not lead to any ambiguity.) If we know h and want to solve for v, R is also unknown, so our task is to solve (19)–(22) for both v and R. Of course, $v(r)$ defined as the value of u (the value function for our stochastic control problem) for an argument x such that $|x| = r$ solves (19)–(22), but it is not clear whether it is the *only* solution, because, as we shall soon see, (19)–(22) lead to a nonlinear equation for R. We want to show that (19)–(22) indeed admits a unique $C^2([0, R])$ solution.

From now on, we assume that v is any function satisfying (19)–(22) on some interval $[0, R]$, $R > 0$, and belonging to $C^2([0, R])$. It is easy to show, using the maximum principle, that both v_r and v_{rr} are nonnegative on $[0, R]$. Thus v is convex in $[0, R]$; in particular, v_r achieves its extreme values at the endpoints of this interval. By (20) and (21) we get

$$0 \leq \frac{dv}{dr}(r) \leq 1 \tag{23}$$

for every $r \in [0, R]$. Extending v to $[0, \infty)$ by

$$v(r) = v(R) + (r - R) \quad \text{for } r > R, \tag{24}$$

we get, by (21)–(22), a $C^2([0, \infty))$ convex function satisfying (23) on $[0, \infty)$. Now we want to show that

$$v(r) - \frac{d^2 v}{dr^2}(r) - \frac{n-1}{r}\frac{dv}{dr}(r) - h(r) \le 0 \tag{25}$$

for $r > R$. By (19),

$$v(r) - \frac{n-1}{r}\frac{dv}{dr}(r) \ge h(r) \quad \text{on } [0, R]. \tag{26}$$

Also, by (19) and (22), equality in (26) holds for $r = R$. Thus

$$\left(\frac{d}{dr}\left(v - \frac{n-1}{r}\frac{dv}{dr}\right)\right)(R) \le h_r(R), \tag{27}$$

which yields, by (21) and (22) once again,

$$1 + \frac{n-1}{R^2} \le h_r(R). \tag{28}$$

Also, by (19)–(21),

$$v(R) - \frac{n-1}{R} = h(R). \tag{29}$$

To prove (25) we need, by (24),

$$v(R) + (r - R) - \frac{n-1}{r} \le h(r), \tag{30}$$

i.e.,

$$h(R) + \frac{n-1}{R} + (r - R) - \frac{n-1}{r} \le h(r). \tag{31}$$

By convexity of h and (28) we have

$$\begin{aligned} h(r) &\ge h(R) + h_r(R)(r - R) \\ &\ge h(R) + \left(1 + \frac{n-1}{R^2}\right)(r - R). \end{aligned} \tag{32}$$

Thus, if we show

$$\frac{n-1}{R} - \frac{n-1}{r} \le \frac{n-1}{R^2}(r - R) \tag{33}$$

for $r > R$, then, by (32), (31), (so also (25)) is true. But (33) follows easily from the mean value theorem. Thus, by (19), (23), (24), and (25), if we define, for $x \in R^n$,

$$u(x) = v(|x|), \tag{34}$$

we see that u is a C^2 convex solution of (12) in R^n. Invoking uniqueness of a nonnegative, convex $W^{2,\infty}_{loc}$ solution of the HJB equation (12), once again we see that such u (thus v) is uniquely determined. We have proved the following theorem.

THEOREM 2.3. *A function $v \in C^2([0,R])$ for some $R > 0$ satisfies* (19)–(22) *iff $v(x) = u(|x|)$ for every $x \in R^n$, $|x| \leq R$, where u is the value function for the stochastic control problem* (9)–(10) *with a radially symmetric running cost function h. Moreover, extending v to $[0,\infty)$ by* (24)*, we get $v(x) \equiv u(|x|)$ in the whole R^n.*

Thus, for a solution v of the boundary value problem (19)–(21), the smooth fit condition (22) is not only necessary, but also sufficient for $v(x) \equiv u(|x|)$. An analogous result—sufficiency of (22) for describing the value function—for one-dimensional singular stochastic control problems may be found in [32] and, for more general diffusions, in [38].

It is natural to pose a similar question in dimension n. Suppose we have a bounded region G in R^n, say simply connected, and a vector field w defined on ∂G and pointing outside G. Let v be a real function defined in $\overline{G}$ such that

$$v - \triangle v = h \text{ in } G, \tag{35}$$

$$\frac{\partial v}{\partial w} = 1, \tag{36}$$

and

$$\frac{\partial^2 v}{\partial w^2} = 0 \tag{37}$$

on ∂G. Is v the restriction of the value function u for the stochastic control problem (9)–(10) to $\overline{G}$? If it does not have to be the case, is it true at least under an additional assumption that $w = \nabla u$ on ∂G ? We do not know the answer to this question and, because of the fact that the n-dimensional problem is much more complicated, it might be negative.

Now we turn to the special case of a radially symmetric real analytic running cost h satisfying the conditions given above. Then it is well known that u satisfying (13) with the boundary condition (15) is itself real analytic in $\overline{B}_R(0) = \overline{\mathcal{C}}$ and, as we have explained before, also radially symmetric. Then we can easily find u explicitly as the sum of a power series.

Let again $u(x) = v(r)$, $r = |x|$. v is real analytic on $[0,R]$, and let

$$h = \sum_{i=0}^{\infty} b_{2i}\, r^{2i} \tag{38}$$

(the odd coefficients drop out because of symmetry of h). We want to find the real analytic solution v of (19)–(22) on $[0,R]$. The corresponding homogeneous equation

$$v(r) - \frac{d^2 v}{dr^2}(r) - \frac{n-1}{r}\frac{dv}{dr}(r) = 0 \tag{39}$$

is similar to the well-known Bessel equation (see, e.g., [30]),

$$z'' + \frac{1}{x} z' + z = 0, \tag{40}$$

and can be solved in more or less the same way. Let

$$v(r) = \sum_{i=0}^{\infty} a_i r^i \tag{41}$$

be a solution of (39). Differentiating (41) term by term twice, plugging the results into (39), and equating the corresponding terms, we get

$$v_1(r) = a_0 \left(1 + \sum_{i=1}^{\infty} \frac{r^{2i}}{2 \cdot 4 \cdot \cdots \cdot (2i) \cdot n \cdot (n+2) \cdot \cdots \cdot (n+2i-2)}\right) \tag{42}$$

(we took $a_1 = 0$ to cancel the singular term $\frac{1}{r}$ in (39)). Of course, the radius of convergence of this series is ∞. As in the case of the Bessel equation, it may be shown that the second solution v_2, linearly independent on v_1, behaves at 0 similarly as the function $\log r$. Solving (19) using (38), we get

$$a_{i+2} = \frac{a_i - b_i}{(i+2)(i+n)} \tag{43}$$

for $i = 0, 2, 4, \dots$ and $a_i = 0$ for i odd. Thus, once a_0 is determined, all the remaining coefficients of the solution v can be found from (43) and, because of the singularity of v_2 at 0, all real analytic solutions of (19) in intervals containing 0 have this form. For any such solution, (20) is satisfied automatically because $a_1 = 0$. Thus, our task amounts to find constants a_0 and $R > 0$ such that the solution v of (19)–(21) on $[0, R]$ given by (41), (43) satisfies (22). By Theorem 2.3 and our previous remarks, there always exists a unique pair a_0, R ($R > 0$) satisfying this condition, and the value function $u(x)$ for the stochastic control problem (9)–(10) equals $v(|x|)$.

As an example, let us take perhaps the simplest possible running cost satisfying our assumptions, namely $h(x) = |x|^2 = r^2$. Then

$$v(r) = a_0 + \frac{a_0}{2n} r^2 + \sum_{i=2}^{\infty} \frac{a_0 - 2n}{2 \cdot 4 \cdot \cdots \cdot (2i) \cdot n \cdot (n+2) \cdot \cdots \cdot (n+2i-2)} \, r^{2i}. \tag{44}$$

Regarding R as a variable and computing a_0 in terms of R from (21), we get

$$a_0 = a_0(R) = \frac{1 + 2nS(R)}{\frac{R}{n} + S(R)}, \tag{45}$$

where

$$S(R) = \sum_{i=2}^{\infty} \frac{R^{2i-1}}{2 \cdot 4 \cdot \cdots \cdot (2i-2) \cdot n \cdot (n+2) \cdot \cdots \cdot (n+2i-2)}. \tag{46}$$

Thus, (22) gives us the following equation for R:

$$\frac{1}{n} + 2S(R) + (1 - 2R) \cdot S_1(R) = 0, \tag{47}$$

where

$$S_1(R) = \sum_{i=2}^{\infty} \frac{(2i-1)R^{2i-2}}{2 \cdot 4 \cdot \cdots \cdot (2i-2) \cdot n \cdot (n+2) \cdot \cdots \cdot (n+2i-2)}. \tag{48}$$

These equations can be solved numerically, yielding, e.g., for $n = 2$, $R = 1.4857650660\dots$, $a_0 = 1.9624905574\dots$ with accuracy to ten decimal points. See [11, Theorem 8] for a more elegant description of this solution in terms of the modified Bessel function (42).

Another simple experiment we can do is to take $h \equiv 0$ and to solve (19)–(21) in $[0, R]$, obtaining the total discounted cost of control

$$v(x) = E^x \int_0^\infty e^{-t} d\xi_t \tag{49}$$

necessary to keep the process in the ball $\overline{B}_R(0)$. The solution is of the form (42), where a_0 is chosen to match (21). We can easily check that the solution approaches ∞ as $R \to 0$. This is the reason why the optimal region $\mathcal{C}$ cannot be too small: if we try to keep the process very close to the minimum $h(0) = 0$ of the running cost, the control costs grow enormously. Thus, the optimal control can be thought of as a compromise between keeping the running cost $h(X_t)$ small and using a small amount of control.

3. Ergodic control.

3.1. Posing a problem. Now we turn to investigating the following problem. Let E be an open, connected set in R^n. To simplify the exposition, we initially assume that E is bounded and with sufficiently smooth (say C^2) boundary. (Eventually, we will remove the boundedness assumption and significantly relax the regularity requirement.) Let $x \in E$ and let $(W_t, \mathcal{F}_t, t \geq 0)$ be, as before, a standard n-dimensional Brownian motion. Finally, let X_t be defined by (2), (3), where (N_t, ξ_t) is the solution to the Skorokhod problem for the domain E and the normal reflection direction at ∂E starting at x. See, e.g., [37] for a definition and a construction of such a process. Let h be a nonnegative, Borel measurable function bounded on compact subsets of R^n with $\lim_{|x|\to\infty} h(x) = \infty$ (for some of the results to follow, the assumptions on h will be strengthened). Now let, for $T > 0$,

$$w_E(x, T) = \frac{1}{T} E^x \int_0^T [h(X_t)dt + d\xi_t], \tag{50}$$

$$u_E(x) = \lim_{T\to\infty} w_E(x, T). \tag{51}$$

By the ergodic theorem, the time-average of a quantity converges to its mean under the equilibrium measure, which is the Lebesgue measure on the domain E, properly normalized, i.e., divided by $|E|$, the Lebesgue measure of E. The limit of the expected average time spent by X_t inside E weighted by h is equal to $\frac{1}{|E|}\int_E h(x)dx$. The limit of the occupation time on the boundary for the Brownian motion $\sqrt{2}W_t$ is equal to the surface area $|\partial E|$ of E divided by its volume:

$$\lim_{T\to\infty} \frac{1}{T} E^x \int_0^T d\xi_t = \frac{|\partial E|}{|E|}. \tag{52}$$

Thus, actually, u_E does not depend on the starting point $x \in E$ (i.e., is a functional of the domain E only) and equals

$$J(E) = \frac{|\partial E|}{|E|} + \frac{1}{|E|}\int_E h(x)dx. \tag{53}$$

Our goal is to find a region E^* minimizing the ergodic value function u_E, i.e., a suitably smooth minimizer of (53).

This problem is, in a sense, a counterpart of the one with exponential discounting considered before. See [41] for an analysis of connections between singular ergodic control and singular control with exponentially discounted running cost for multidimensional Gaussian processes.

Instead of seeking an optimal control for (51) in the class of C^2 domains directly, we enlarge the domain of definition of (53) to the class of Caccioppoli sets E in R^n. Below, we recall some basic definitions; for a more complete discussion we refer to [23].

Definition 3.1 (see [23, p. 3]). *Let $G \subseteq R^n$ be an open set and let $f \in L^1(G)$. Define*

$$\begin{aligned} \int_G |Df| = \sup\{\int_G f \operatorname{div} g dx \ : \ g = (g_1, \dots, g_n) \in C_0^1(G, R^n), \\ \text{and } |g(x)| \leq 1 \text{ for all } x \in G\}, \end{aligned} \tag{54}$$

where, as usual, $\operatorname{div} g = \sum_{i=1}^n \frac{\partial g_i}{\partial x_i}$.

Definition 3.2 (see [23, p. 4]). *A function $f \in L^1(G)$ is said to have bounded variation in G if $\int_G |Df| < \infty$. We define $BV(G)$ as the space of all functions in $L^1(G)$ with bounded variation.*

It can be shown [23, Remark 1.12] that, under the norm

$$\|f\|_{BV} = \|f\|_{L^1} + \int_G |Df|, \tag{55}$$

$BV(G)$ is actually a Banach space.

In what follows, let ϕ_E denote the characteristic function (indicator) of a Borel set $E \subseteq R^n$.

Definition 3.3 (see [23, p. 5]). *Let E be a Borel set and G an open set in R^n. Define the perimeter of E in G as*

$$P(E, G) = \int_G |D\phi_E| = \sup\left\{\int_E \operatorname{div} g dx : g \in C_0^1(G, R^n),\ |g(x)| \leq 1\right\}. \tag{56}$$

If $G = R^n$, denote $P(E, R^n)$ by $|\partial E|$.

It is known that if the boundary of E is C^2, then $P(E, G) = H_{n-1}(\partial E \cap G)$, where H_{n-1} denotes the $n-1$-dimensional Hausdorff measure.

Definition 3.4 (see [23, p. 6]). *If a Borel set E has locally finite perimeter, that is, if $P(E, G) < \infty$ for every bounded open set G, then E is called a Caccioppoli set.*

Thus, we are looking for a minimizer of (53) in the class of all Caccioppoli sets, i.e., for a Caccioppoli set E^* such that

$$J(E^*) \leq J(E) \ \text{ for every Caccioppoli set } E. \tag{57}$$

For every $x \in R^n$ and $\rho > 0$, denote

$$B_\rho(x) = \{y : |x - y| < \rho\}.$$

Definition 3.5 (see [23, p. 43]). *A point x belongs to the reduced boundary, $\partial^* E$, of a set E if*

$$\int_{B_\rho(x)} |D\phi_E| > 0$$

for all $\rho > 0$, the limit $v(x) = \lim_{\rho\to 0} v_\rho(x)$ exists, where

$$v_\rho(x) = \frac{\int_{B_\rho(x)} D\phi_E}{\int_{B_\rho(x)} |D\phi_E|},$$

and $|v(x)| = 1$.

The vector $v(x)$ defined above can be regarded as the generalized unit inner normal vector to ∂E at x. Indeed, if ∂E is C^2, then $\partial^* E = \partial E$ and $v(x)$ is the unit inner normal vector to ∂E at x [23, p. 44]. Thus, $\partial^* E$ is the part of ∂E on which the generalized inner normal $v(x)$ to ∂E can be defined.

3.2. Existence of minimizers. Now we shall prove, using standard tools, that a minimizer indeed exists.

Let E_k be a minimizing sequence of Borel sets, i.e.,

$$J(E_k) \to c = \inf_E J(E). \tag{58}$$

Clearly, $c < \infty$.

Step 1. $|E_k| \le C$ for all k natural and some $C < \infty$.

Indeed, suppose $|E_k| \to \infty$, $k \to \infty$. Let $M > 0$ be an arbitrary constant. $h(x) \to \infty$ for $x \to \infty$, so the Lebesgue measure of a set $F = \{x \in R^n : h(x) \le M\}$ is finite. Let $|E_k| \ge 2|F|$. Then, by $h \ge 0$,

$$\begin{aligned}\int_{E_k} hdx &= \int_{E_k\cap F} hdx + \int_{E_k - F} hdx \\ &\ge \int_{E_k - F} hdx \\ &\ge M|E_k - F| \\ &\ge M(|E_k| - |F|) \\ &\ge \frac{M}{2}|E_k|,\end{aligned}$$

so

$$J(E_k) \ge \frac{1}{|E_k|}\int_{E_k} hdx \ge \frac{M}{2}. \tag{59}$$

Thus, $c \ge \frac{M}{2}$ for an arbitrary M, which is a clear contradiction.

Step 2. $\frac{|\partial E_k|}{|E_k|} \le C$ for all k natural and some $C < \infty$.

This follows at once from (53) and (58).

Step 3. $|\partial E_k| \le C$ for all k natural and some $C < \infty$.

This follows easily from (58) and the previous two steps.

Step 4. $|E_k| \ge c_0$ for all k and some $c_0 > 0$.

Suppose this is not true, e.g., there exists a minimizing sequence E_k such that

$$|E_k| \to 0 \text{ as } k \to \infty. \tag{60}$$

Let us recall the isoperimetric inequality [23, Corollary 1.29]: there exists a positive constant $c_1 = c_1(n)$ depending only on the dimension n such that, for every Caccioppoli set $E \subseteq R^n$, we have

$$|E|^{\frac{n-1}{n}} \le c_1(n)\, |\partial E|.$$

(Actually, as is well known, the best constant $c_1(n) = \frac{|B|^{\frac{n-1}{n}}}{|\partial B|}$, where B is an n-dimensional ball.) Using this inequality, we get

$$J(E_k) > \frac{|\partial E_k|}{|E_k|} \geq \frac{1}{c_1(n)|E|^{\frac{1}{n}}}. \tag{61}$$

But, by (60), the right-hand side of (61) diverges to ∞, and thus so does $J(E_k)$, which is a contradiction.

Step 5. We want to prove that a subsequence of E_k converges to a minimizer of (53). Say that $J(E_k) \leq c+1$, where c is given by (58). Let $M > 0$ and let $R > 0$ be large enough to assure $h \geq M$ outside $B_R(0)$. Then, for all k,

$$\begin{aligned} |E_k - B_R(0)| &\leq |E_k \cap [h \geq M]| \\ &\leq \frac{1}{M}\int_{E_k} hdx \\ &\leq \frac{|E_k|}{M}\, J(E_k) \\ &\leq \frac{C}{M}\,(c+1) \;\to\; 0 \end{aligned} \tag{62}$$

as $M \to \infty$ ($R \to \infty$) (the last inequality follows from step 1). Also, it is known that

$$|\partial(E \cap B_r)| \leq |\partial E| \tag{63}$$

for any Borel set $E \subseteq R^n$ such that $|E| < \infty$ and any ball B_r ([50]; see also [47]). Thus,

$$|\partial(E_k \cap B_R(0))| \leq |\partial E_k| < C \tag{64}$$

for every k, $R > 0$ by step 3. Thus, we see that $\phi_{E_k \cap B_R(0)}$ is a bounded subset of $BV(B_R(0))$. It is known (see Theorem 1.19 in [23]) that bounded sets in $BV(B_R)$ are relatively compact in $L^1(B_R)$, so $\phi_{E_k \cap B_R(0)}$ has a convergent subsequence in $L^1(B_R)$. But a sequence converging in L^p, $1 \leq p < \infty$, contains a subsequence converging almost everywhere (a.e.), so the limit of $\phi_{E_k \cap B_R(0)}$ in $L^1(B_R)$ must also be an indicator of some set E_R^*. By an usual diagonal argument, we can extract a subsequence, still called E_k, such that for $E^* = \bigcup_{m=1}^{\infty} E_m^*$, $\phi_{E_k \cap B_m(0)} \to \phi_{E^* \cap B_m(0)}$ in $L^1(B_m)$ for every natural m. In particular, $\phi_{E_k} \to \phi_{E^*}$ in L^1_{loc}. By semicontinuity (see Theorem 1.9 in [23])

$$|\partial E^*| \leq \lim\inf_{n\to\infty} |\partial E_k|. \tag{65}$$

By Fatou's lemma

$$\int_{E^*} hdx \leq \lim\inf_{k\to\infty} \int_{E_k} hdx. \tag{66}$$

We want to show that also $|E_k| \to |E^*|$, because this, together with step 1, (65), and (66) finishes the proof. Without any loss of generality we can, by steps 1 and 4, assume that the sequence $|E_k|$ is convergent. Define a family of Borel measures μ_k on R^n by

$$\mu_k(A) = |A \cap E_k|. \tag{67}$$

Then (62) assures that this family is tight, so there exists a measure μ on R^n such that a subsequence of μ_k converges weakly to μ (see, e.g., [7, Theorem 5.1]) and thus $\mu(R^n) = \lim_{k\to\infty}\mu_k(R^n)$. But, as we know from our previous reasoning, $\mu(A) = |A \cap E^*|$ (it is true for bounded sets by $\phi_{E_k} \to \phi_{E^*}$ in L^1_{loc}, so it is also true for all Borel sets A), so $|E_k| \to |E^*|$. Existence of minimizers for (53) is shown.

3.3. Some properties of minimizers.

3.3.1. Boundedness. For convenience, we assume that a minimizer E^* under consideration is open (Caccioppoli sets are defined only up to a set of the Lebesgue measure zero and their boundaries have measure zero). We want to prove the boundedness of E^*. Assume, to the contrary, that E^* is unbounded. Choose a large constant M and let

$$R = \sup\{x \in R^n : h(x) \le M\}. \tag{68}$$

Let $E^{**} = E^* \cap B_R(0)$ and let $\epsilon = \epsilon(M) = |E^* - E^{**}|$. By assumption, $\epsilon > 0$. By (63), $|\partial E^{**}| \le |\partial E^*|$, and thus

$$\begin{aligned} J(E^*) &= \frac{|\partial E^*|}{|E^*|} + \frac{1}{|E^*|}\int_{E^*} h(x)dx \\ &> \frac{|\partial E^{**}| + \int_{E^{**}} h(x)dx + M\epsilon}{|E^{**}| + \epsilon} \\ &> \frac{|\partial E^{**}|}{|E^{**}|} + \frac{1}{|E^{**}|}\int_{E^{**}} h(x)dx \\ &= J(E^{**}) \end{aligned}$$

if M is large enough and, consequently, if ϵ is small enough. This contradicts the fact that E^* is a minimizer of J and proves our claim.

3.3.2. Regularity. At this stage, the only thing we can guarantee is that a minimizer E^* is a bounded Caccioppoli set. In general, the boundary ∂E of a Caccioppoli set E is, up to a set of $|D\phi_E|$ (or, equivalently, the Haussdorff measure H_{n-1}) measure zero, a countable union of C^1 hypersurfaces [23, Theorem 4.4], and usually you cannot expect anything better than that. Fortunately, problems of the same nature have been studied thoroughly in variational calculus and related domains, starting from a classic theory of minimal surfaces (see, e.g., [23]), through minimizing

$$I(E) = |\partial E| + \int_E h dx \tag{69}$$

[39] until more recent developments posed in a much more abstract and general framework (see, e.g., [1, 2]). In particular, regularity of E^* can be easily deduced from known results in the following way. Define, for a Caccioppoli set E and an open, bounded set $A \subseteq R^n$,

$$\psi(E, A) = \int_A |D\phi_E| - \inf\left\{\int_A |D\phi_F| \ : \ F \triangle E \subset\subset A\right\}, \tag{70}$$

where $F \triangle E = (F - E) \cup (E - F)$ and $B \subset\subset A$ means that $\overline{B} \subseteq A$. It is known [39, 48] that if G is an open subset of R^n, $n \ge 2$, and E is a Caccioppoli set such that

$$\psi(E, B_\rho(x)) \le C\rho^n \tag{71}$$

for every $x \in G$ and every $\rho \in (0, R)$, with C and R local positive constants, then the reduced boundary $\partial^* E$ of E is a $C^{1,\frac{1}{2}}$-hypersurface in G and the Haussdorf measure

$$H_s[(\partial E - \partial^* E) \cap G] = 0 \tag{72}$$

for every $s > n - 8$. In particular, if $n \leq 7$, ∂E is $C^{1,\frac{1}{2}}$. Taking $E = E^*$, a minimizer of (53), and $G = R^n$, we show (71) using an argument similar to that given in [48] for a minimizer of (69). For completeness, we provide a sketch of the argument.

Suppose (71) is violated. Then for any $M > 0$ we can find a small ball $B_\rho(x)$ and a set $F \subset\subset B_\rho(x)$ such that

$$\int_{B_\rho(x)} |D\phi_F| \leq \int_{B_\rho(x)} |D\phi_E^*| - M\rho^n. \tag{73}$$

Then, if we take

$$E^{**} = (E^* - B_\rho(x)) \cup F,$$

we get, for M big enough,

$$J(E^{**}) < J(E^*), \tag{74}$$

because of (73) and the fact that $|E^*|$ and $\int_{E^*} h dx$ differ from $|E^{**}|$ and $\int_{E^{**}} h dx$, correspondingly, at most by $C\rho^n$ for some constant C depending on $\max_{B_\rho(x)} h$ only. However, (74) contradicts the definition of E^*.

Thus, we get $\partial^* E^* \in C^{1,\frac{1}{2}}$ and (72); in particular, for $n \leq 7$ we have $\partial E^* \in C^{1,\frac{1}{2}}$. In fact, the only assumption about h we need at this stage is $h \in L^\infty_{loc}$. Now we can improve the regularity of ∂E using the existence of multipliers in the nonparametric case (see [21, 24, 25]). Thus we have the following proposition.

PROPOSITION 3.6. *Let $h \in C^k$, $k \geq 1$, and let E^* be a minimizer of* (53). *Then $\partial^* E^*$ is a C^{k+1}- hypersurface in R^n and* (72) *holds for every $s > n - 8$. In particular, for $n \leq 7$, ∂E^* is a C^{k+1}-hypersurface in R^n.*

Let us mention that, in general, in problems of this type, smoothness of minimizers in dimensions bigger than 7 cannot be expected. A counterexample (the Simons cone) can be found in [23, Theorem 16.4].

Proposition 3.6 says that, for $n \leq 7$, the regularity of ∂E^* is precisely the same as that of a solution of the mean curvature equation (see, e.g., [22, Corollary 16.7]). We shall further discuss the reason for this in the next subsection.

An alternative way to establish an initial regularity result for ∂E^* is based on an observation that E^* also minimizes (69) under a constraint

$$|E| = \text{const} = |E^*|. \tag{75}$$

Thus, we can use known regularity results for the problem of minimizing (69) under a volume constraint in the parametric case. However, proofs of such results usually require serious work to show (71) or a similar inequality, either directly (see, e.g., [26, 49]) or by using Lagrange multipliers (see, e.g., [3]), and then use (71) in the way we have just done it, so it seems more natural to take advantage of a simple proof of (71) which can be given in the unconstrained case.

3.3.3. The mean curvature equation. Assume that $h \in C^1$. By Proposition 3.6, for $n \leq 7$, $\partial E^* \in C^2$. Thus we can obtain, from the first variation of (53), the mean curvature equation satisfied by ∂E^*. Let $x \in \partial E^*$. Let $\kappa_1, \dots, \kappa_{n-1}$ be the principal curvatures of $\partial E^* \in C^2$ at x, and let

$$H(x) = \frac{1}{n-1} \sum_{i=1}^{n-1} \kappa_i \tag{76}$$

be the mean curvature of $\partial E^* \in C^2$ at x (see, e.g., [22, section 14.6]). Then, computing the first variation of (53) at E^* under domain deformations, by the well-known fact that the first variation of the area equals $-(n-1)H$ [23, Theorem 10.4, equation (10.12), and Remark 10.6] (notice that our definition of the mean curvature differs from the one used in [22] by a normalizing factor $-(n-1)$), we get

$$H(x) = \frac{h(x) - c}{n-1}, \tag{77}$$

where c is defined by (58), because E^* is a minimizer of (53).

By Proposition 3.6, for dimensions $n \geq 8$ we have $\partial^* E^* \in C^2$. Thus, (77) holds for $x \in \partial^* E^*$, but $\partial E^* - \partial^* E^*$ may be nonempty. However, in this case we can still regard ∂E^* as a *generalized* solution to (77).

3.3.4. Connectedness and simple connectedness. From now on, in addition to $h \geq 0$ and $\lim_{|x| \to \infty} h(x) = \infty$, we assume that h is convex. We want to prove that, in any dimension n, a minimizer of (53) is simply connected in the following sense: its complement in $R^n \cup \{\infty\}$ (R^n compacted by adding a point at infinity) has only one component. Intuitively, it means that there are no holes inside a minimizer. We shall need the following lemma, which generalizes (63).

LEMMA 3.7. *Let E be a Caccioppoli set in R^n and let $C \subseteq R^n$ be convex. Then*

$$|\partial(E \cap C)| \leq |\partial E|. \tag{78}$$

This result is contained in [50, Lemma 8].

Now, let E^* be a minimizer of (53). We can, with no loss of generality, assume that E^* is open. Suppose, to the contrary, that E^* is not simply connected, i.e., its complement in $R^n \cup \{\infty\}$ has at least 2 components: A_∞ containing ∞ and a bounded component A. Choose a positive number b such that the set

$$C_b = \{x \in R^n : \ h(x) \leq b\}, \tag{79}$$

which is convex by convexity of h, contains A and $|A| > d := |E^* - C_b| > 0$. Let e be a fixed unit vector in R^n and let $\tilde{b}$ be such that

$$|A \cap H_{\tilde{b}}| = d, \tag{80}$$

where

$$H_{\tilde{b}} = \{x : \ e \cdot x \leq \tilde{b}\}. \tag{81}$$

Define

$$E^{**} = (E^* \cap C_b) \cup (A \cap H_{\tilde{b}}). \tag{82}$$

Clearly, E^{**} is a Caccioppoli set, $|E^{**}| = |E^*|$, and

$$\int_{E^{**}} hdx < \int_{E^*} hdx \tag{83}$$

by construction (h takes bigger values on $E^* - C_b$ than on $A \cap H_{\tilde{b}}$ and their measures are the same). Also, by Lemma 3.7 applied twice,

$$|\partial E^{**}| \leq |\partial(E^* \cap C_b)| \leq |\partial E^*|, \tag{84}$$

which yields

$$J(E^{**}) < J(E^*), \tag{85}$$

which is a clear contradiction.

Remark. We can always (also in the case of nonconvex h) take a *connected* minimizer E^*. Indeed, we know already that $\partial E^* \in C^2$. From now on, assume that E^* is actually open (Caccioppoli sets are defined up to a set of measure zero) and let $x \in E^*$. Let C be the component of E^* containing x. From our previous considerations it follows that the solution to the Skorokhod problem for E^* starting at x with normal reflection direction is optimal for ergodic discounting, i.e., minimizes (51). But it is clear that this solution never leaves C, so this policy also solves an analogous Skorokhod problem for C. Thus, by the ergodic theorem, C also minimizes (53).

3.3.5. Convexity in two dimensions. Let $n = 2$ and let the assumptions about h be the same as in the last subsection. We want to prove that E^*, a minimizer of (53), is convex. The proof is similar to the one we have just given. Suppose that our claim is false. Then, there exist points $x_1, x_2 \in \partial E^*$ such that

$$B \cap E^* = \emptyset, \tag{86}$$

where B is the open, nonempty set, the boundary of which consists of the interval joining x_1 to x_2 and the shorter arc joining x_1 to x_2 in ∂E^*, and moreover,

$$B \subseteq C_b \tag{87}$$

for C_b defined by (79), and b is such that

$$|E^* - C_b| = |B|. \tag{88}$$

Let

$$E^{**} = (E^* \cap C_b) \cup B. \tag{89}$$

Similarly as in the last proof, we check that $|E^{**}| = |E^*|$,

$$\int_{E^{**}} hdx < \int_{E^*} hdx, \tag{90}$$

and $|\partial(E^* \cap C_b)| \leq |\partial E^*|$. Finally, $|\partial E^{**}| < |\partial(E^* \cap C_b)|$ because the length of the interval $[x_1, x_2]$ is smaller than the length of the corresponding arc. Thus, again, (85) holds, leading to a contradiction.

Remark. This argument clearly does not go through for $n \geq 3$. In this case, a similar question, whether or not a minimizer of (69) with convex h is convex, remains open [47]. Modifications of the above argument, however, can give us partial answers to the convexity question in higher dimensions. For example, there is no point $x \in \partial^* E^*$ at which all the principal curvatures of $\partial^* E$ are strictly negative and if h (hence, by Proposition 3.8 of the next subsection, E^*) is axisymmetric, then E^* is convex.

3.3.6. Symmetry. Let $a \in R^n - \{0\}$, and let

$$N = N_a = \{x \in R^n : \ a \cdot x = 0\} \tag{91}$$

be a hyperplane in R^n. We say that a function h is symmetric with respect to N iff for every $x \in N$ and $t \in R$ we have $h(x - ta) = h(x + ta)$.

PROPOSITION 3.8. *Let the function h be nonnegative, strictly convex, and symmetric with respect to N and such that $\lim_{|x|\to\infty} h(x) = \infty$. Then, any minimizer of* (53) *is also symmetric with respect to N.*

For simplicity of the proof, assume that $a = e_n = (0, \dots, 0, 1)$, so $N = \{x : x_n = 0\}$. Let E^* be a nonsymmetric minimizer of (53). Let E^{**} be the set which results from E^* by its Steiner symmetrization with respect to the hyperplane N, i.e., for every $x = (x', 0) \in N$, $x' \in R^{n-1}$,

$$E^{**} \cap \{x + te_n : \ t \in R\} = \left\{x + te_n : \ -\frac{b}{2} \le t \le \frac{b}{2}\right\}, \tag{92}$$

where b is the one-dimensional Lebesgue measure of $E^* \cap \{x + te_n : \ t \in R\}$. By definition and Fubini's theorem, $|E^{**}| = |E^*|$. It is also well known (see, e.g., [20, 43]) that $|\partial E^{**}| \le |\partial E^*|$. Thus, if we show that

$$\int_{E^{**}} h(x)dx < \int_{E^*} h(x)dx, \tag{93}$$

we get $J(E^{**}) < J(E^*)$, which is a contradiction. (93) follows easily from a simple lemma.

LEMMA 3.9. *Let f be an even, strictly convex function on R and let $a > 0$ be a given finite number. Then, the minimum of $F(E) = \int_E f(x)dx$ over all Lebesgue measurable sets $E \subseteq R$ such that $|E| = a$ is*

$$\int_{-\frac{a}{2}}^{\frac{a}{2}} f(x)dx \tag{94}$$

and is attained iff E differs from $[-\frac{a}{2}, \frac{a}{2}]$ by a set of measure zero.

Proof. Let $E \subseteq R$ be a set such that $|E| = a$ and let $b_- = |E \cap \{x < 0\}|$, $b_+ = |E \cap \{x > 0\}|$. Obviously, by the fact that f is strictly increasing on $[0, \infty)$,

$$\int_{E\cap\{x>0\}} f(x)dx \ge \int_0^{b_+} f(x)dx, \tag{95}$$

and the equality holds iff $|E \cap \{x > 0\} \triangle [0, b_+]| = 0$. By the same reasoning applied to $E \cap \{x < 0\}$, we get that $F(E) \ge F([-b_-, b_+])$ with equality only if $E = [-b_-, b_+]$ a.e. To complete the proof, we need to show that the minimum is attained iff $b_+ = b_-$. Suppose, for example, that $b_+ > b_-$. (The argument in the opposite case is essentially the same.) Then

$$\begin{aligned}
\int_{-b_-}^{b_+} f(x)dx &= \int_{-b_-}^{b_-} f(x)dx + \int_{b_-}^{b_- + \frac{b_+ - b_-}{2}} f(x)dx + \int_{b_- + \frac{b_+ - b_-}{2}}^{b_+} f(x)dx \\
&> \int_{-b_- - \frac{b_+ - b_-}{2}}^{-b_-} f(x)dx + \int_{-b_-}^{b_-} f(x)dx + \int_{b_-}^{b_- + \frac{b_+ - b_-}{2}} f(x)dx \\
&= \int_{-b_- - \frac{b_+ - b_-}{2}}^{b_- + \frac{b_+ - b_-}{2}} f(x)dx
\end{aligned}$$

by symmetry of h and its monotonicity on $[0,\infty)$ $((-\infty,0])$ and $|[-b_-,b_+]| = |[-b_- - \frac{b_+-b_-}{2}, b_- + \frac{b_+-b_-}{2}]|$. The proof is complete.

In particular, if h is radially symmetric, then a minimizer of (53) is a ball $B_R = B_R(0)$. Putting $I(R) = J(B_R)$ and using calculus methods, we can find the radius of the optimal ball. For example, consider again the case of $n = 2$ and $h(x) = |x|^2$. We have

$$I(R) = \frac{2}{R} + \frac{\pi}{2}R^4. \tag{96}$$

This function attains its unique minimum at $R_0 = \frac{1}{\sqrt[5]{\pi}}$, so in this case $E^* = B_{\frac{1}{\sqrt[5]{\pi}}}(0)$.

3.4. A generalization. The initial assumption of smoothness and boundedness of the domains E under consideration was convenient for two reasons. First, path-by-path constructions of strong solutions to the corresponding Skorokhod problem for this case are known. Second, in this context the ergodic result

$$\lim_{T\to\infty} \frac{1}{T} E^x \int_0^T [h(X_t)dt + d\xi_t] = J(E) \tag{97}$$

for all $x \in \overline{E}$ is easy to get. In fact, we do not even need ergodic theorems here; it is enough to use the Ito's rule and the necessary and sufficient condition for solvability of the Neumann problem for the Laplace operator (see, e.g., [31, p. 95]). However, as we have already seen, it is natural to extend the class of allowed domains to Caccioppoli sets. See [9, 18, 51] for the construction of the Brownian motion reflected normally at the boundary of a bounded Caccioppoli set E; the case of an unbounded Caccioppoli domain is considered in [8, 19]. In the appendix we show that (97) actually holds for a.e. $x \in \overline{E}$ (with respect to the Lebesgue measure) if E is bounded Caccioppoli. Moreover, we have, for an arbitrary Caccioppoli set E,

$$\liminf_{T\to\infty} \frac{1}{T} E^x \int_0^T [h(X_t)dt + d\xi_t] \geq J(E) \tag{98}$$

quasi-everywhere (q.e.) in x, i.e., for all starting points $x \in \overline{E} - N$, where N is a set of the Newtonian capacity zero. $J(E)$, the right-hand side of (98), should be interpreted as $+\infty$ if $|E| = +\infty$. Both (97) and (98) follow from ergodic theorems given in [16, 17] after some technical details are taken care of. The only assumptions about h we need for (97)–(98) is that h is nonnegative, Borel measurable, and bounded on compact subsets of R^n, and $\lim_{|x|\to\infty} h(x) = \infty$. These facts, together with the arguments given in subsections 3.2-3.3, allow us to state the following theorem.

THEOREM 3.10. *Assume that $h \geq 0$ is Borel and locally bounded and $\lim_{|x|\to\infty} h(x) = \infty$. There exists an open, connected Caccioppoli set E^* such that the Brownian motion reflected normally at ∂E^* minimizes the ergodic cost*

$$\liminf_{T\to\infty} \frac{1}{T} E^x \int_0^T [h(X_t)dt + d\xi_t]$$

in the class of all Brownian motions X_t reflected normally at the boundary of some open, connected Caccioppoli set E for q.e. starting point $x \in \overline{E}$. The family of such minimizing sets E^* coincides with the family of open, connected minimizers of the functional $J(E)$ defined by (53) in the class of all Caccioppoli sets.

Recall that some properties of minimizing sets E^* were investigated in the last subsection.

Let us also remark that, in general, we cannot expect uniqueness of a minimizer E^* of J. See, e.g., [49] for related work.

3.5. Concluding remarks. Can we use analogous geometric-measure theoretic techniques in the exponentially discounted cost problem of minimizing (9)–(10)? In particular, can we establish or at least improve regularity of the free boundary $\partial\mathcal{C}$ by considerations based on regularity for the mean curvature equation? We have not found any results along this line. Admittedly, if we know from the outset that the optimal reflection direction $-\nabla u$ is *normal* to $\partial\mathcal{C}$ on the whole $\partial\mathcal{C}$, then u satisfies (13) in $\mathcal{C}$ with the Neumann boundary condition (14) and an additional free-boundary condition

$$\frac{\partial^2 u}{\partial \eta^2} = 0, \tag{99}$$

where $\eta = \nabla u$ is the outer normal at $\partial\mathcal{C}$. The last equation follows from the fact that $w = |\nabla u|^2$ is C^1,

$$\frac{\partial^2 u}{\partial \eta^2} = \frac{\partial \sqrt{w}}{\partial \eta} \tag{100}$$

on $\partial\mathcal{C}$ (compare (17)), and w attains its maximum value 1 on $\partial\mathcal{C}$. Moreover, because ∇u is perpendicular to $\partial\mathcal{C}$, we get another free-boundary condition (15). Thus, if $\partial\mathcal{C}$ is regular enough to assure that (13) holds up to the boundary (it suffices to have $\partial\mathcal{C} \in C^2$), writing (13) in the local normal-tangential system of coordinates at $x \in \partial\mathcal{C}$, we get

$$u - \triangle u = u - \frac{\partial^2 u}{\partial \eta^2} - (n-1)H\frac{\partial u}{\partial \eta} = h \tag{101}$$

on $\partial\mathcal{C}$, where H is again the mean curvature of $\partial\mathcal{C}$. Thus, by (14), (15), and (99) we get

$$H(x) = \frac{c - h(x)}{n-1} \tag{102}$$

resembling (77), with a different meaning of c of course. In fact, by Theorem 3.1 in [44], we can get both (15) and (99) under a weaker assumption that the Brownian motion reflected normally at $\partial\mathcal{C}$ minimizes (9)–(10) in the class of all solutions for the Skorokhod problem with normal reflection for C^2 domains containing x for every $x \in \mathcal{C}$. Thus, under this weaker condition (102) holds also, where, again, c is the (constant) value which the solution u of (13), (14) in $\mathcal{C}$ takes on $\partial\mathcal{C}$. Thus, one might attempt to find $\mathcal{C}$ as a *maximizer* of a (69)-type functional

$$|\partial E| + \int_E (c - h(x))dx \tag{103}$$

with a suitably chosen constant c.

As we have seen before, ∇u is normal to $\partial\mathcal{C}$ on $\partial\mathcal{C}$ if h is radially symmetric, but we are not sure whether a different example can be given. Additionally, in the radially symmetric case, $\partial\mathcal{C}$ is just a sphere, so any further considerations of its regularity are superfluous.

4. Appendix.

Let E be an open, connected Caccioppoli set in R^n and let m be the Lebesgue measure on E. Let the process X_t defined by (2), (3) be the Brownian motion reflected normally at ∂E. See [9, 18, 51] for the construction of such a process for bounded E, the case of an unbounded domain is considered in [8, 19]. The aim of this appendix is to show that (98) holds q.e. in x, i.e., for all starting points $x \in \overline{E} - N$, where N is a set of capacity zero and $J(E)$, the right-hand side of (98), is interpreted as $+\infty$ if $|E| = +\infty$. In this appendix we will use terms such as capacity, exceptional set, positive continuous additive functional, measure of finite energy integral, etc. from potential theory unified with the theory of Markov processes. Their definitions can be found, for example, in [15].

We also want to show that in the case of a bounded Caccioppoli domain E (98) can be refined to (97) for a.e. $x \in \overline{E}$.

Let $p_t = p_t(x, dy)$, $t > 0$, be the transition function of the process X_t. For a Borel function $f : R \to R$ let

$$p_t f(x) = \int_{\overline{E}} f(y) p_t(x, dy) = E^x f(X_t)$$

if the integral on the right-hand side exists.

It is clear that $\{p_t\}$ is *irreducible*; in other words, any p_t-invariant (i.e., such that $p_t f = f$ a.s. for all $t > 0$) bounded function on $\overline{E}$, integrable with respect to m, is constant a.e. in $\overline{E}$. This follows from the fact that E is open, connected, and from known averaging properties of the transition functions p_t.

First assume that E is bounded. Then h is bounded on E. Thus, by the corollary following [16, Theorem 1], we have

$$\lim_{t \to \infty} p_t h = \frac{1}{|E|} \int_E h(y) dy \tag{104}$$

for almost every $x \in \overline{E}$. In fact, this is true for q.e. $x \in \overline{E}$ by [17, Theorem 2]. This clearly implies

$$\lim_{T \to \infty} \frac{1}{T} E^x \int_0^T h(X_t) dt = \frac{1}{|E|} \int_E h(y) dy \tag{105}$$

for q.e. $x \in \overline{E}$.

Fix $s, t > 0$, and let, for $x \in \overline{E}$,

$$f(x) = E^x \xi_t. \tag{106}$$

By [18, Theorem 1.1], f is integrable with respect to m on $\overline{E}$.

We want to show that

$$E^x(\xi_{t+s} - \xi_s | X_s) = f(X_s) \tag{107}$$

P^x a.s. for q.e. $x \in \overline{E}$.

First note that, by the construction of the Brownian motion X_t reflected normally at ∂E [9, 18, 51], ξ_t is a positive continuous additive functional of X_t. Thus

$$\xi_{t+s}(\omega) - \xi_s(\omega) = \xi_t(\theta_s \omega) \tag{108}$$

for $\omega \in \Lambda$, where $P^x(\Lambda) = 1$ for all $x \in \overline{E} - N$, where N is an exceptional subset of $\overline{E}$ and $\theta_s : \Omega \to \Omega$ is the so-called *shift operator*

$$X_t \circ \theta_s(\omega) = X_{t+s}(\omega)$$

for all $\omega \in \Omega$, $t, s \geq 0$ (see, e.g., [15, p. 89], [33, p. 77]).

Fix $x \in \overline{E} - N$. It is not hard to show (107) under the assumption that ξ_t is $\mathcal{F}_t^X := \sigma\{X_u, 0 \leq u \leq t\}$-measurable. (Follow the proof below without modifying ξ_t.) However, in general we only know that ξ_t is $\{\tilde{\mathcal{F}}_t\}$-measurable, where $\{\tilde{\mathcal{F}}_t\}$ is the minimum completed admissible family for the Markov process X_t [15, pp. 89, 124], so some additional care must be taken. Using [15, Lemma 4.1.3], we can, as in the solution to [33, Problem 2.5.7], construct a $\mathcal{F}_{t+}^X$-measurable modification $\tilde{\xi}_t$ of the process ξ_t such that

$$P^y[\xi_t \neq \tilde{\xi}_t] = 0$$

for $y = x$ and for $P^x \circ X_s^{-1}$ a.e. $y \in \overline{E}$ (in [15, Lemma 4.1.3] use the measure $\mu = \frac{1}{2}(P^x \circ X_s^{-1} + \delta_x)$, where δ_x is the probability measure concentrated at the point x). To proceed further, we need to verify that

$$\xi_t \circ \theta_s = \tilde{\xi}_t \circ \theta_s \quad P^x \text{a.s.} \tag{109}$$

First, by construction,

$$P^{P^x \circ X_s^{-1}}[\xi_t \neq \tilde{\xi}_t] = 0. \tag{110}$$

Fix an $\epsilon > 0$. $[\xi_t \neq \tilde{\xi}_t] \in \mathcal{F}_{t+\epsilon}$, so, by the definition of the minimum completed admissible family for X_t, there exists a set $F \in \mathcal{B}(R^{[0,t+\epsilon]})$, the Borel σ-field on $R^{[0,t+\epsilon]}$ equipped with the product topology, such that

$$[\xi_t \neq \tilde{\xi}_t] \subseteq Z := [X. \in F],$$

where $X.$ in the above formula denotes the path of X on $[0, t+\epsilon]$, and

$$P^{P^x \circ X_s^{-1}}(Z) = 0,$$

i.e.,

$$P^y(Z) = 0 \tag{111}$$

for $P^x \circ X_s^{-1}$ a.e. y. Thus, by the Markov property of X_t,

$$\begin{aligned} P^x[\xi_t \circ \theta_s \neq \tilde{\xi}_t \circ \theta_s] &= P^x[\theta_s^{-1}[\xi_t \neq \tilde{\xi}_t]] \\ &\leq P^x[\theta_s^{-1} Z] \\ &= E^x P^x[\theta_s^{-1} Z | \mathcal{F}_s] \\ &= E^x P^{X_s}[Z] \ = \ 0, \end{aligned}$$

where the last equality follows from (111). This proves (109).

Thus, by (108), (109), and (110), we have

$$\begin{aligned} E^x(\xi_{t+s} - \xi_s | X_s) &= E^x(\xi_t \circ \theta_s | X_s) = E^x(\tilde{\xi}_t \circ \theta_s | X_s) \\ &= E^x(E^x[\tilde{\xi}_t \circ \theta_s | \mathcal{F}_s] | X_s) = E^x(E^{X_s} \tilde{\xi}_t | X_s) \\ &= E^{X_s} \tilde{\xi}_t = E^{X_s} \xi_t \\ &= f(X_s). \end{aligned}$$

The fourth inequality follows from the Markov property of X_t. (Here $E^{X_s}\tilde{\xi}_t$, $E^{X_s}\xi_t$ should be interpreted as the compositions of real functions $y \to E^y \tilde{\xi}_t$, $y \to E^y \xi_t$, and the random variable X_s.) We have shown (107).

Next, we need the fact that

$$\lim_{t\to\infty} p_t f = \frac{1}{|E|}\int_E f(y)dy \tag{112}$$

for almost every $x \in \overline{E}$. Similarly as in the proof of [16, Theorem 1], we argue that the limit $\lim_{t\to\infty} p_t f = g$ exists a.e. and in $L^1(\overline{E}, m)$ (we have $f \in L^1(\overline{E}, m)$ instead of $L^p(\overline{E}, m)$ for some $p > 1$, but the argument still goes through, because every backward martingale is convergent a.s. and in L^1). Moreover, it follows from the proof cited above that

$$E^m g = E^m f(X_0) = \int_E f(y)dy. \tag{113}$$

By [16, Corollary to Theorem 1], for any natural N we have

$$\lim_{t\to\infty} p_t(f\wedge N) = \frac{1}{|E|}\int_E (f\wedge N)(y)dy,$$

so

$$g = \lim_{t\to\infty} p_t f \geq g_0 := \frac{1}{|E|}\int_E f(y)dy.$$

But, by (113), g and g_0 have the same expectations under P^m, so they must be equal a.e. This proves (112).

Now take $t = 1$ in the definition (106) of f and denote by σ the Borel measure on R^n defined by

$$\sigma(G) = P(E, G) \tag{114}$$

for all open sets G. σ is a *smooth* measure, i.e., it charges no set of zero capacity [18]. It also follows from [18, Theorem 1.1 and Lemma 2.1] that σ is the Revuz measure corresponding to the positive continuous additive functional ξ_t in the sense of [15, Theorem 5.1.3].

We want to show that

$$\int_E f(y)dy = E^m\xi_1 = \sigma(\overline{E}) = |\partial E|. \tag{115}$$

If σ is of finite energy integral, this follows immediately from [15, Lemma 5.1.4(iii)]. In general, by [15, Theorem 3.2.3], there exists an increasing sequence $\{F_n\}$ of closed subsets of $\overline{E}$ such that

$$\sigma\left(\overline{E} - \bigcup_{n=1}^{\infty} F_n\right) = 0, \tag{116}$$

$\overline{E} - \bigcup_{n=1}^{\infty} F_n$ is an exceptional set, and the measures $I_{F_n}\cdot\sigma$, $n = 1, 2, \dots$, are of finite energy integral. By [15, Theorem 5.1.3], $I_{F_n}\cdot\sigma$ corresponds to the functional $I_{F_n}\xi$ defined by

$$(I_{F_n}\xi)_t = \int_0^t I_{F_n}(X_s)d\xi_s. \tag{117}$$

Using [15, Lemma 5.1.4(iii)] for $I_{F_n}\cdot\sigma$ and $I_{F_n}\xi$, we get

$$E^m(I_{F_n}\xi)_1 = (I_{F_n}\cdot\sigma)(\overline{E}) = \sigma(\overline{E}\cap F_n). \tag{118}$$

Letting $n\to\infty$, we get (115).

To finish the proof of (97), it suffices to show (still under the assumption that E is bounded), that

$$\lim_{T\to\infty}\frac{E^x\xi_T}{T} = \frac{|\partial E|}{|E|}$$

for a.e. $x\in\overline{E}$. Indeed, this together with (105) gives (97). To this end, let us remark that

$$\frac{E^x\xi_{[T]}}{[T]+1} \le \frac{E^x\xi_T}{T} \le \frac{E^x\xi_{[T]+1}}{[T]},$$

where $[T]$ denotes the integer part of T, so it actually suffices to show

$$\lim_{n\to\infty}\frac{E^x\xi_n}{n} = \frac{|\partial E|}{|E|} \tag{119}$$

for a.e. $x\in\overline{E}$, where the limit is taken over the integers n. We have, by (107),

$$\begin{aligned} E^x\xi_n &= E^x\left(\sum_{i=0}^{n-1}E^x(\xi_{i+1}-\xi_i|X_i)\right)\\ &= E^x\left(\sum_{i=0}^{n-1}f(X_i)\right)\\ &= \sum_{i=0}^{n-1}(p_if)(x) \end{aligned}$$

for a.e. $x\in\overline{E}$. Thus, by (112) and (115), (119) holds for a.e. $x\in\overline{E}$. This ends the proof of (97).

Now let us analyze the case of an *unbounded* Caccioppoli domain E. The construction of a normally reflected Brownian motion in E can be found in [8, 19].

Case 1. $|E|<\infty$.

Again, let σ be defined by (114). As in the previous case, σ is the Revuz measure corresponding to the positive continuous additive functional ξ_t and, by [15, Theorem 3.2.3], there exists an increasing sequence $\{F_n\}$ of closed subsets of $\overline{E}$ such that $\sigma(F_n)<\infty$ for each n, (116) holds, $\overline{E}-\bigcup_{n=1}^\infty F_n$ is an exceptional set, and the measures $I_{F_n}\cdot\sigma$, $n=1,2,\ldots$, are of finite energy integral.

Arguing as in the case of a bounded domain, but with $(I_{F_n}\xi)_t$ defined by (117) instead of ξ_t, and using (118), we get, for a.e. $x\in\overline{E}$,

$$\lim_{T\to\infty}\frac{E^x(I_{F_n}\xi)_T}{T} = \frac{E^m(I_{F_n}\xi)_1}{|E|} = \frac{\sigma(\overline{E}\cap F_n)}{|E|}.$$

Thus,

$$\liminf_{T\to\infty}\frac{E^x\xi_T}{T} \ge \frac{\sigma(\overline{E}\cap F_n)}{|E|}$$

for every $n \geq 1$, i.e.,

$$\liminf_{T\to\infty} \frac{E^x \xi_T}{T} \geq \frac{\sigma(\overline{E})}{|E|} = \frac{|\partial E|}{|E|} \tag{120}$$

for a.e. $x \in \overline{E}$.

Carefully examining this argument and using [17, Theorem 2], we can get (120) for q.e. $x \in \overline{E}$.

For any natural number N, $h \wedge N \in L^p$ for all $p \geq 1$. Thus, proceeding as in the bounded case, we get (104) and (105) with $h \wedge N$ instead of h, which yields

$$\liminf_{T\to\infty} \frac{1}{T} E^x \int_0^T h(X_t)dt \geq \frac{1}{|E|} \int_E h(y)dy \tag{121}$$

for q.e. $x \in \overline{E}$. This, together with (120), ends the proof of (98) in the case of $|E| < \infty$.

Case 2. $|E| = \infty$.

Define, for any natural number N,

$$f_N = N - (h \wedge N).$$

$f_N \in L^p$ for all $p \geq 1$, because $\lim_{|x|\to\infty} h(x) = \infty$. By [16, Corollary (ii) to Theorem 1] and [17, Theorem 2] once again, we get

$$\lim_{t\to\infty} p_t f_N = 0$$

q.e. in $\overline{E}$, so

$$\lim_{t\to\infty} p_t(h \wedge N) = N$$

q.e. in $\overline{E}$. From this we easily get

$$\liminf_{T\to\infty} \frac{1}{T} E^x \int_0^T h(X_t)dt = \infty$$

q.e. in $\overline{E}$ by the same reasoning as in the previous cases. This ends the proof of (98).

Acknowledgments. This work is a part of my Ph.D. thesis at the Courant Institute of Mathematical Sciences, New York University. I want to express my deep gratitude to my advisor, Professor Marco Avellaneda, and to Professor Robert V. Kohn for many helpful discussions. I also thank Professor Jose Luis Menaldi for his helpful comments on the first version of this paper, Professors Zhen-Qing Chen and Ruth William for references to recent papers, and the referees for their valuable suggestions.

REFERENCES

[1] F. Almgren, *Existence and regularity almost everywhere of solutions to elliptic variational problems with constraints*, Mem. Amer. Math. Soc., 165 (1976).

[2] F. Almgren, J. E. Taylor, and L. Wang, *Curvature-driven flows: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 387–438.

[3] E. Barozzi, M. Emmer, and E. H. A. Gonzalez, *Lagrange multipliers and variational methods for equilibrium problems of fluids*, Rend. Sem. Mat. Univ. Padova, 85 (1991), pp. 35–53.

[4] J. A. Bather and H. Chernoff, *Sequential decisions in the control of a spaceship*, in Procedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, III: Physical Sciencs, University of California Press, Berkeley, CA, 1967, pp. 181–207.

[5] J. A. Bather and H. Chernoff, *Sequential decisions in the control of a spaceship (finite fuel)*, J. Appl. Probab., 49 (1967), pp. 584–604.

[6] V. E. Beneš, L. A. Shepp, and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.

[7] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., Wiley, New York, 1999.

[8] Z. Q. Chen, *Reflecting Brownian motions and a deletion result for Sobolev spaces of order* $(1, 2)$, Potential Anal., 5 (1996), pp. 383–401.

[9] Z. Q. Chen, P. J. Fitzsimmons, and R. J. Williams, *Reflecting Brownian motions: Quasimartingales and strong Caccioppoli sets*, Potential Anal., 2 (1993), pp. 219–243.

[10] P. L. Chow, J. L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[11] M. H. A. Davis and M. Zervos, *A pair of explicitly solvable singular stochastic control problems*, Appl. Math. Optim. 38 (1998), pp. 327–352.

[12] A. Dixit, *A simplified treatment of the theory of optimal regulation of Brownian motion*, J. Econom. Dynam. Control, 15 (1991), pp. 657–673.

[13] B. Dumas, *Super contact and related optimality conditions*, J. Econom. Dynam. Control, 15 (1991), pp. 675–685.

[14] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[15] M. Fukushima, *Dirichlet Forms and Markov Processes*, North Holland, Amsterdam, Kodansha, New York, 1980.

[16] M. Fukushima, *A note on irreducibility and ergodicity of symmetric Markov processes,* in Stochastic Processes in Quantum Theory and Statistical Physics (Marseille, 1981), Lecture Notes in Phys. 173, Springer-Verlag, Berlin, New York, 1982, pp. 200–207.

[17] M. Fukushima, *Capacitary maximal inequalities and an ergodic theorem,* in Probability Theory and Mathematical Statistics (Tbilisi, 1982), Lecture Notes in Math. 1021, Springer-Verlag, Berlin, New York, 1983, pp. 130–136.

[18] M. Fukushima, *Dirichlet forms, Caccioppoli sets and the Skorokhod equation*, in Stochastic Differential and Difference Equations, Progr. Systems Control Theory 23, Birkhäuser Boston, Boston, MA, 1997, pp. 59–66.

[19] M. Fukushima, *On semimartingale characterization of functionals of symmetric Markov processes*, Electron. J. Probab., 4 (1999), pp. 1–31.

[20] P. R. Garabedian, *Partial Differential Equations*, Wiley, New York, 1964.

[21] C. Gerhardt, *On the capillary problem with constant volume*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 4 (1975), pp. 303–320.

[22] D. Gilbarg and N. Trudinger, *Elliptic Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1985.

[23] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Boston, Boston, MA, 1984.

[24] E. Giusti, *Generalized solutions for the mean curvature equation*, Pacific J. Math., 88 (1980), pp. 297–321.

[25] E. Gonzalez, U. Massari, and I. Tamanini, *Existence and regularity for the problem of a pendent liquid drop*, Pacific J. Math., 88 (1980), pp. 399–420.

[26] E. Gonzalez, U. Massari, and I. Tamanini, *Regularity of boundaries of sets minimizing perimeter with a volume constraint*, Indiana Univ. Math. J., 32 (1983), pp. 25–37.

[27] J. M. Harrison, *Brownian Motion and Stochastic Flow Systems*, Wiley, New York, 1985.

[28] J. M. Harrison and M. I. Taksar, *Instantaneous control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 454–466.

[29] J. M. Harrison and A. J. Taylor, *Optimal control of a Brownian storage system*, Stochastic Process Appl., 6 (1978), pp. 179–194.

[30] F. John, *Ordinary Differential Equations*, Lecture notes, Courant Institute of Mathematical Sciences, New York University, New York, NY, 1965.

[31] F. John, *Partial Differential Equations*, 4th ed., Springer-Verlag, New York, 1982.

[32] I. Karatzas, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.

[33] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[34] I. Karatzas and S. E. Shreve, *Connection between optimal stopping and singular stochastic control* I: *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.

[35] I. Karatzas and S. E. Shreve, *Connection between optimal stopping and singular stochastic control,* II: *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–541.

[36] L. Kruk, *Optimal policies for n-dimensional singular stochastic control problems. Part* I: *The*

Skorokhod problem, SIAM J. Control Optim., 38 (2000), pp. 1603–1622.

[37] P.-L. Lions and A. S. Sznitman, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37 (1984), pp. 511–537.

[38] J. Ma, *On the principle of smooth fit for a class of singular stochastic control problems for diffusions*, SIAM J. Control Optim., 30 (1992), pp. 975–999.

[39] U. Massari, *Esistenza e regolarità delle ipersuperfici di curvatura media assegnata in R^n*, Arch. Rational Mech. Anal., 55 (1974), pp. 357–382.

[40] J. L. Menaldi and M. Robin, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.

[41] J. L. Menaldi, M. Robin, and M. I. Taksar, *Singular ergodic control for multidimensional Gaussian processes*, Math. Control Signals Systems, 5 (1992), pp. 93–114.

[42] J. L. Menaldi and M. I. Taksar, *Optimal correction problem of a multidimensional stochastic system*, Automatica J. IFAC, 25 (1989), pp. 223–232.

[43] G. Polya and G. Szego, *Isoperimetric Inequalities in Mathematical Physics*, Princeton University Press, Princeton, NJ, 1951.

[44] J. Simon, *Variations with respect to domain for Neumann condition*, in Control of Distributed Parameter Systems, 1986: Proceedings of the 4th IFAC Symposium, Los Angeles, California, 1986, Pergamon Press, Elmsford, NY, Oxford, UK, 1987, pp. 395–399.

[45] H. M. Soner and S. E. Shreve, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.

[46] M. Taksar, *Convex solutions to variational inequalities and multidimensional singular control*, in The Dynkin Festschrift, Markov Processes and Their Applications, Progr. Probab. 34, Birkhauser Boston, Boston, MA, 1994, pp. 371–386.

[47] I. Tamanini, *Interfaces of prescribed mean curvature*, in Variational Methods for Free-Boundary Interfaces (Menlo Park, CA, 1985), Springer-Verlag, New York, Berlin, 1987, pp. 91–97.

[48] I. Tamanini, *Boundaries of Caccioppoli sets with Hölder-continuous normal vector*, J. Reine Angew. Math., 334 (1982), pp. 27–39.

[49] I. Tamanini, *Variational problems of least area type with constraints*, Ann. Univ. Ferrara Sez. VII (N.S.), 34 (1988), pp. 183–217.

[50] T. Vogel, *Unbounded parametric surfaces of prescribed mean curvature*, Indiana Univ. Math. J., 31 (1982), pp. 281–288.

[51] R. J. Williams, *Reflected Brownian motion: Hunt processes and semimartingale representation*, in Barcelona Seminar on Stochastic Analysis, Progr. Probab. 32, Birkhauser Boston, Boston, MA, 1999, pp. 216–221.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 661–672

A NONCONTROLLABILITY RESULT FOR SYSTEMS OF MIXED ORDER*

GIUSEPPE GEYMONAT† AND VANDA VALENTE‡

Abstract. In this paper we prove an abstract theorem of noncontrollability for the evolution equation associated to a Douglis–Nirenberg elliptic system of mixed order with nonempty essential spectrum. In particular, we show by Weyl's characterization of the essential spectrum that the condition of exact controllability does not hold true. We discuss examples concerning the elasticity operator in the framework of the linear membrane shell theory.

Key words. exact controllability, membrane shell theory, essential spectrum

AMS subject classifications. 93B05, 47A25, 74K25, 35P05

PII. S0363012998348322

1. Introduction.

The exact controllability of the vibrations of mixed order systems frequently arises in the control theory of distributed parameters systems and techniques of harmonic analysis, and results of spectral theory are often involved in the solution of the control problem. Excitable systems can otherwise produce vibrations which are not exactly controllable. It is the aim of our paper to describe such a situation and get an abstract noncontrollability result via spectral analysis.

In order to fix the notations we consider a system of differential equations

$$\begin{cases} \mathbf{v}_{tt} + \mathbf{A}\mathbf{v} = 0 & \text{in} \quad \Omega \times (0,T), \\ \\ \mathbf{B}\mathbf{v} = 0 & \text{on} \quad \partial\Omega \times (0,T), \end{cases} \tag{1.1}$$

where **A** is a Douglis–Nirenberg elliptic operator of mixed order and **B** is a system of normal boundary conditions. We restrict our attention to a linear selfadjoint operator associated with a sesquilinear form. Following the spectral analysis results given in [6], [7], [8], [9], we show that the exact controllability problem is strictly connected to the spectral properties of **A**; in particular, we point out that when the operator **A** has a nonempty essential spectrum, that is, the operator **A** contains a block of zero order, the system (1.1) is not exactly controllable. The proof of our main result is obtained by Weyl's characterization of the essential spectrum here recalled and reformulated. We adopt the Weyl sequences (or singular sequences) to show that the exact controllability problem cannot in general be solved.

In many physical situations the operator **A** depends on a small parameter ε and the limit as $\varepsilon \to 0$ may lead to a *singular perturbation* problem. When ε vanishes and the existence of eigenvalues which form a discrete spectrum is still ensured, the controllability problem may be solved, but if the limit problem has an essential spectrum, the system is no longer controllable; the phenomenon of the *loss of exact controllability* is therefore connected to the appearance of the *essential spectrum.*

*Received by the editors November 25, 1998; accepted for publication (in revised form) April 19, 2000; published electronically October 4, 2000.

http://www.siam.org/journals/sicon/39-3/34832.html

†Laboratoire de Mecanique et Technologie, E.N.S. de Cachan/C.N.R.S./ Université Paris VI, 61 Avenue du Presidente Wilson, 94235 Cachan, France (geymonat@lmt.ens-cachan.fr).

‡Istituto per le Applicazioni del Calcolo C.N.R., viale del Policlinico 137, 00161 Roma, Italy (valente@iac.rm.cnr.it).

For example, in the case of a thin shell, the perturbation parameter is represented by the shell thickness. In the spectral analysis for thin shells carried out in [5], the dependence on the thinness parameter in the formula that gives the distribution of the eigenvalues has been studied. The established result suggested, in particular, the existence of a finite accumulation point for a subsequence of eigenvalues of the limit problem. Moreover, an explicit computation of the eigenfunctions allowed us to deduce the nonexact controllability of the spherical membrane. Results of exact controllability of thin shells, including shallow shells, can be found in [4], [5], [12], [14].

This paper presents a result of independent interest which in the framework of the linear thin shell theory also generalizes the previous one and shows the phenomenology of the loss of the exact controllability in the transition shell-membrane. Although in this paper we repeatedly refer to the membrane shell operator, we insist on the interest in a criterion of noncontrollability for other applications in the control theory of distributed parameters systems. Moreover, the general formulation of our examples allows us to consider membrane shells of arbitrary shapes which can also take part in the modelling of more complex phenomena (see, for example, [11]).

2. Mixed order systems and spectral analysis.

Most of the arguments in this section have been introduced in [6], [7], [8], [9]. We consider mixed order operators which exhibit an essential spectrum.

2.1. Some assumptions.

Let $\bar{\Omega}$ be a compact d-dimensional C^∞ manifold with C^∞ boundary Γ, and let $\mathbf{A} = (A_{st})_{s,t=1,\dots,q}$ (with $q > 1$) be a selfadjoint matrix of differential operators A_{st} of order $m_s + m_t$, where

$$m_1 \geq m_2 \cdots \geq m_{q-p} > m_{q-p+1} = \cdots = m_q \geq 0 \qquad (1 \leq p < q).$$

We assume the following.

(A.I) $\mathbf{A}$ is elliptic in the sense of Douglis–Nirenberg (see [1]), i.e., the matrix of principal symbols $\sigma^0(\mathbf{A})(\mathbf{x}, \boldsymbol{\xi})$ is invertible for all $(\mathbf{x}, \boldsymbol{\xi})$ in the nonzero cotangent bundle $T^*(\bar{\Omega})\backslash 0$.

In this paper we consider operators $\mathbf{A}$ which contain a block of order 0, that is, $m_q = 0$. In this case it is useful to split $\mathbf{A}$ in blocks of positive order and zero order, respectively,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{NN} & \mathbf{A}_{NQ} \\ \mathbf{A}_{QN} & \mathbf{A}_{QQ} \end{pmatrix}, \tag{2.1}$$

where we have adopted the notations

$$\begin{array}{ll} N = 1, 2, \dots, q-p, & Q = q-p+1, \dots, q, \\ \mathbf{A}_{NN} = (A_{st})_{s,t \in N}, & \mathbf{A}_{NQ} = (A_{st})_{s \in N, t \in Q}, \\ \mathbf{A}_{QN} = (A_{st})_{s \in Q, t \in N}, & \mathbf{A}_{QQ} = (A_{st})_{s,t \in Q}. \end{array}$$

In (2.1) the block $\mathbf{A}_{QQ}$ is of order 0 while the other blocks are of positive order, and we assume for simplicity that the operator $\mathbf{A}_{NN}$ is elliptic.

Along with the operator $\mathbf{A}$ we consider a matrix of differential boundary operators $\mathbf{B} = (\mathbf{B}_{rs})_{r \in M, s=1,\dots q}$, where $M = \{1, 2, \dots, m\}$, $m = \sum_{s=1}^{q} m_s$. Each operator $\mathbf{B}_{rs}$ has order $\sigma_r + m_s$, where $-m_1 \leq \sigma_1 \leq \sigma_2 \cdots \leq \sigma_m \leq m_1 - 1$, and if $\sigma_r + m_s < 0$, then $\mathbf{B}_{rs} \equiv 0$. When $\sigma_r \leq -1$ for $r = 1, \dots, \ell$, then we consider the decomposition

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}^0 \\ \mathbf{B}^1 \end{pmatrix}, \qquad \mathbf{B}^0 = (\mathbf{B}_{rs})_{r=1,\dots\ell;\, s=1,\dots q}. \tag{2.2}$$

(A.II) $\mathbf{B}$ defines a system of normal boundary conditions, i.e., there exists a complementary system $\mathbf{C}$ of boundary conditions such that $\{\mathbf{B},\mathbf{C}\}$ are the reduced Cauchy data of $\mathbf{A}$ (see [9]), and the following Green formula holds for smooth functions:

$$((\mathbf{A}\mathbf{v},\tilde{\mathbf{v}}))_\Omega - ((\mathbf{v},\mathbf{A}\tilde{\mathbf{v}}))_\Omega = ((\mathbf{C}\mathbf{v},\,\mathbf{B}\tilde{\mathbf{v}}))_\Gamma + ((\mathbf{B}\mathbf{v},\,\mathbf{C}\tilde{\mathbf{v}}))_\Gamma.$$

In what follows, we denote by $\mathbf{H}=\mathbf{H}^0$ the vector space $\mathbf{L}^2(\Omega)$, and we denote by $\mathbf{V}$ the vector space of functions $\mathbf{u}\in H^{m_1}(\Omega)\times H^{m_2}(\Omega)\cdots\times H^{m_q}(\Omega)$ with the boundary condition $\mathbf{B}^0\mathbf{u}=0$, equipped with the norm $\|\mathbf{u}\|_V$ induced by $\prod_{i=1}^q H^{m_i}(\Omega)$.

The realization $\mathbf{A}_B$ of $\mathbf{A}$ associated to the boundary conditions $\mathbf{B}$ is defined via the following assumption.

(A.III) $\mathbf{A}_B$ is a selfadjoint lower bounded operator in $\mathbf{H}$ associated with the sesquilinear and continuous form $a(\mathbf{u},\mathbf{v})$ on $\mathbf{V}\times\mathbf{V}$ which satisfies the inequality

$$a(\mathbf{v},\mathbf{v})+\tau\|\mathbf{v}\|_H^2 \geq c_0\,\|\mathbf{v}\|_V^2\,, \qquad c_0>0 \text{ and } \tau\in\mathbb{R}.$$

Under this assumption $\mathbf{A}_B$ is the operator in $\mathbf{H}$ with domain

$$\mathrm{D}(\mathbf{A}_B)=\left\{\mathbf{u}\in\prod_{i=1,\dots,q}H^{m_i}(\Omega);\ \mathbf{A}\mathbf{u}\in\mathbf{H},\ \mathbf{B}\mathbf{u}=0\right\}\subset\mathbf{V},$$

defined by $(\mathbf{A}_B\mathbf{u},\mathbf{v})_H=a(\mathbf{u},\mathbf{v})$ for any $\mathbf{u}\in D(\mathbf{A}_B)$ and $\mathbf{v}\in\prod_{i=1}^q H^{m_i}(\Omega)$. Realizations $\mathbf{A}_B$ verifying assumption (A.III) are determined in [9].

We can also introduce the operator $\mathbf{A}_B^{min}$ defined by $\mathbf{A}_B^{min}\mathbf{u}=\mathbf{A}\mathbf{u}$ with domain

$$\mathrm{D}(\mathbf{A}_B^{min})=\left\{\mathbf{u}\in\prod_{i=1,\dots,q}H^{m_1+m_i}(\Omega),\ \mathbf{B}\mathbf{u}=0\right\}.$$

Examples. In the linear thin shell theory we consider the *membrane approximation* of the elasticity operator. Let Ω be a bounded open set of boundary Γ of the plane $\mathbb{R}^2$; the surface S of an elastic membrane is defined by two curvilinear coordinates x_1 and x_2; it is the image in $\mathbb{R}^3$ of Ω by the map

$$\boldsymbol{\varphi}:(x_1,x_2)\in\overline{\Omega}\to\mathbb{R}^3.$$

In each point of S two tangent vectors $\mathbf{a}_\alpha=\partial\boldsymbol{\varphi}/\partial x_\alpha$ $\alpha=1,2$ and a normal vector $\mathbf{a}_3=\frac{\mathbf{a}_1\times\mathbf{a}_2}{|\mathbf{a}_1\times\mathbf{a}_2|}$ are considered.

Adopting the summation convention (Greek indices take values in the set $\{1,\,2\}$ and the Latin indices take values in the set $\{1,\,2,\,3\}$) and denoting by $f_{,\alpha}$ the partial derivative of f with respect to x_α, the *first fundamental form* $(a_{\alpha\beta})$ and the *second fundamental form* $(b_{\alpha\beta})$ are given by

$$a_{\alpha\beta}=\mathbf{a}_\alpha\cdot\mathbf{a}_\beta, \qquad \alpha,\beta=1,2,$$

$$b_{\alpha\beta}=\mathbf{a}_3\cdot\mathbf{a}_{\alpha,\beta}, \qquad \alpha,\beta=1,2;$$

moreover, if $(a^{\alpha\beta})$ denotes the inverse matrix of $(a_{\alpha\beta})$, the reciprocal basis $\mathbf{a}^\alpha$ is defined by $\mathbf{a}^\alpha=a^{\alpha\beta}\mathbf{a}_\beta$. Let $\mathbf{v}(x_1,x_2)=v_1\mathbf{a}^1+v_2\mathbf{a}^2+v_3\mathbf{a}^3=v_i\mathbf{a}^i$ be the displacement vector of S; the deformed surface is given by $\boldsymbol{\varphi}+\mathbf{v}$. In the framework of a linearized theory small displacements $\mathbf{v}$ are considered.

The energy of membrane deformation is defined by the symmetric form

$$a^m(\mathbf{v},\tilde{\mathbf{v}}) = \int_S a^{\alpha\beta\lambda\mu}\gamma_{\alpha\beta}(\mathbf{v})\gamma_{\lambda\mu}(\tilde{\mathbf{v}})\,\mathrm{d}\,S, \tag{2.3}$$

where $\mathrm{d}\,S = |\mathbf{a}_1\times\mathbf{a}_2|\,\mathrm{d}\,x_1\,\mathrm{d}\,x_2$ and

$$a^{\alpha\beta\lambda\mu} = \frac{E}{2(1+\nu)}\left[a^{\alpha\lambda}a^{\beta\mu}+a^{\alpha\mu}a^{\beta\lambda}+\frac{2\nu}{(1-\nu)}\,a^{\alpha\beta}a^{\lambda\mu}\right] \tag{2.4}$$

is the tensor of "elastic moduli," with E and ν as the Young modulus and Poisson ratio, respectively.

The *deformation tensor* of the middle surface $(\gamma_{\alpha\beta}(\mathbf{v}))$ is given by

$$\gamma_{\alpha\beta}(\mathbf{v}) = \frac{1}{2}\,(v_{\beta|\alpha}+v_{\alpha|\beta}) - b_{\alpha\beta}v_3, \tag{2.5}$$

where the bar $|$ denotes the covariant derivative defined by means of the Christoffel symbols $\Gamma^{\alpha}_{\beta\lambda} = \mathbf{a}^{\alpha}\cdot\mathbf{a}_{\beta,\lambda}$ and $\mathrm{v}_{\alpha|\beta} = \mathrm{v}_{\alpha,\beta}-\Gamma^{\lambda}_{\alpha\beta}v_{\lambda}$.

Since the component v_3 appears in (2.3) by zero order derivative, while the component v_{α} appears by first order derivatives, we have that the differential selfadjoint operator $\mathbf{A}^m$, associated to the form a^m, is a linear system of differential operators of mixed order with indices $m_1 = m_2 = 1$, $m_3 = 0$, which we can write in block decomposition as

$$\mathbf{A}^m = \begin{pmatrix} \mathbf{A}^m_{NN} & \mathbf{A}^m_{N3} \\ \mathbf{A}^m_{3N} & A^m_{33} \end{pmatrix}, \tag{2.6}$$

where $A^m_{33} = a^{\alpha\beta\lambda\mu}b_{\alpha\beta}b_{\lambda\mu}$ is a zero order operator, and $\mathbf{A}^m_{NN}$ (resp., $\mathbf{A}^m_{N3}$, $\mathbf{A}^m_{3N}$)(N=1,2) is a matrix of differential operators of order $2m_1$ (resp., m_1+m_3, m_1+m_3). Let $\mathbf{u}$ be the vector (v_1,v_2) so that $\mathbf{v} = (v_1,v_2,v_3) = (\mathbf{u},v_3)$, and let $\mathbf{V} = (H^1(\Omega))^2\times L^2(\Omega)$. We may consider three different types of boundary conditions:

(a) $\mathrm{v}_{\alpha} = 0, \quad \alpha = 1,2$ (Dirichlet conditions);

(b) $a^{\alpha\beta\sigma\mu}\gamma_{\sigma\mu}(\mathbf{v})\nu_{\beta} = 0, \quad \alpha = 1,2$ (Neumann conditions),
where $\boldsymbol{\nu}$ is the unit outward normal vector in the surface at points of ∂S;

(c) $\mathrm{v}_{\alpha}\,\nu_{\alpha} = 0, \quad a^{\alpha\beta\sigma\mu}\gamma_{\sigma\mu}(\mathbf{v})\nu_{\beta}\,\tau_{\alpha} = 0$ (intermediate conditions),
where τ_{α} denotes the components of the tangent unit vector.

We define the operator $\mathbf{A}^m_D$, $\mathbf{A}^m_N$, and $\mathbf{A}^m_I$ corresponding to the boundary conditions (a), (b), and (c), respectively. According to the mentioned boundary conditions and the position (2.2), we introduce the following vector spaces:

$\mathbf{V}^{(\mathrm{a})} = \{\mathbf{v}\in\mathbf{V},\ |\ \mathbf{B}^0\mathbf{v} = \mathbf{0}$, with $\mathbf{B}^0 = (\mathbf{B}^0_{\alpha\beta}) = \mathbf{I}\}$,
$\mathbf{V}^{(\mathrm{b})} = \mathbf{V}$,
$\mathbf{V}^{(\mathrm{c})} = \{\mathbf{v}\in\mathbf{V},\ |\ \mathbf{B}^0\mathbf{v} = \mathbf{0}$, with $\mathbf{B}^0 = (\mathbf{B}^0_{1\beta}) = \nu_{\beta}\}$.

Moreover, let $\mathbf{H}$ be the space $(L^2(\Omega))^3$ equipped with the standard scalar product

$$(\mathbf{v},\tilde{\mathbf{v}})_H = ((\mathbf{v},\tilde{\mathbf{v}})) = \int_S a^{\alpha\beta}v_{\alpha}\tilde{v}_{\beta}+v_3\tilde{v}_3dS.$$

We refer to the papers of Ciarlet and Sanchez-Palencia for ellipticity results and uniqueness and existence theorems for linear membrane shell equations (see, for instance, [2] and the references therein).

2.2. The essential spectrum. Let $\sigma_{ess}(\mathbf{A}_B)$ be the *essential spectrum* of the selfadjoint operator $\mathbf{A}_B$, and one can prove (see [7], [8]) that

$$\sigma_{ess}(\mathbf{A}_B) = \omega \cup \omega_b, \tag{2.7}$$

where ω is the closed set of $\lambda \in \mathbb{R}$ such that $\mathbf{A} - \lambda\mathbf{I}$ is not Douglis–Nirenberg elliptic, and ω_b is the closed set of $\lambda \in \mathbb{R}\backslash\omega$ for which the boundary condition does not cover $\mathbf{A} - \lambda\mathbf{I}$, i.e., does not satisfy the Shapiro–Lopatinskii condition.

We recall that the set ω is formed by the points λ such that the determinant of the principal symbol of $\mathbf{A} - \lambda\mathbf{I}$, which we denote by $\sigma^0(\mathbf{A} - \lambda\mathbf{I})(\mathbf{x}, \boldsymbol{\xi})$, vanishes for some real $\boldsymbol{\xi}$ different from zero.

Since we consider the case $m_q = 0$, $\mathbf{A} - \lambda\mathbf{I}$ is not always Douglis–Nirenberg elliptic and the set ω is *nonempty.*

The characterization of the essential spectrum is also given in terms of *singular sequences.*

PROPOSITION 2.1 (Weyl characterization of the essential spectrum [7], [16]). *Let* $\mathbf{T}$ *be a selfadjoint linear operator on a Hilbert space* $\mathbf{H}$, $\lambda^* \in \sigma_{ess}(\mathbf{T})$ *if and only if there exists a singular sequence for* $\mathbf{T} - \lambda^*\mathbf{I}$, *i.e., a sequence* $\mathbf{w}_n$ *of elements of the Hilbert space* $\mathbf{H}$ *such that*

(i_0) $\mathbf{w}_n \in D(\mathbf{T})$,

(i_1) $\|\mathbf{w}_n\|_H = 1$,

(i_2) $(\mathbf{T} - \lambda^*\mathbf{I})\mathbf{w}_n \to 0$ *in* $\mathbf{H}$ *strongly,*

(i_3) $\mathbf{w}_n$ *has no strongly convergent subsequence in* $\mathbf{H}$ *(* $\mathbf{w}_n \to 0$ *in* $\mathbf{H}$ *weakly).*

In our situation we can verify further properties.

PROPOSITION 2.2. *Under the assumptions* (A.I)-(A.III) *if* $\lambda^* = \lambda(\mathbf{x}^*, \boldsymbol{\xi}^*) \in \omega$, *then there is a singular sequence for* $\mathbf{A}_B - \lambda^*\mathbf{I}$ *such that*

(i_4) $a(\mathbf{w}_n, \mathbf{w}_n) \to \lambda^*$,

(i_5) $\mathbf{C}\mathbf{w}_n \to 0$ *in* $\mathbf{L}^2(\Gamma)$ *strongly.*

Proof. In order to prove (i_5) we recall a technique for the construction of a singular sequence. We assume by translation $\mathbf{x}^* = 0$ and take $\mathbf{w} \in C_0^\infty(\mathbb{R}^d)$ with $\|\mathbf{w}\|_H = 1$.

Now we formally write the operator $\mathbf{A}$ in the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{NN} & \mathbf{0} \\ \mathbf{A}_{QN} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}_{NN}^{-1}\mathbf{A}_{NQ} \\ \mathbf{0} & \mathbf{S} \end{pmatrix}, \tag{2.8}$$

where $\mathbf{S} = \mathbf{A}_{QQ} - \mathbf{A}_{QN}\mathbf{A}_{NN}^{-1}\mathbf{A}_{NQ}$. Note that we use the same symbol $\mathbf{I}$ for unit matrices of different dimension. Since $\det\sigma^0(\mathbf{A}-\lambda^*\mathbf{I}) = \det\sigma^0(\mathbf{A}_{NN})\det\sigma^0(\mathbf{S}-\lambda^*\mathbf{I})$, we have that $\det\sigma^0(\mathbf{A}_{NN}) \neq 0$ implies that $\sigma^0(\mathbf{S}) - \lambda^*\mathbf{I}$ is not bijective at $(\mathbf{0}, \boldsymbol{\xi}^*)$.

We consider the eigenvector $\boldsymbol{\vartheta}$ corresponding to the eigenvalue λ^* for $\sigma^0(\mathbf{S})$; then the sequence

$$\tilde{\mathbf{w}}_n(\mathbf{x}) = n^{d/2} e^{i(\mathbf{x}, n^2\boldsymbol{\xi}^*)} \mathbf{w}(n\mathbf{x})\, \boldsymbol{\vartheta} \tag{2.9}$$

is a singular sequence for $\mathbf{S} - \lambda^*\mathbf{I}$. The singular sequence (2.9) is often quoted in literature (see, for example [6], [7], [15]) and can be used to construct a singular sequence for $\mathbf{A} - \lambda^*\mathbf{I}$. Let $\mathbf{P}$ be a parametrix of $\mathbf{A}_{NN}$, whose kernel as an integral operator has its support close to the diagonal; the sequence

$$\mathbf{w}_n = \frac{\mathbf{w}'_n}{\|\mathbf{w}'_n\|_H}, \qquad \mathbf{w}'_n = \begin{pmatrix} -\mathbf{P}\mathbf{A}_{NQ}\tilde{\mathbf{w}}_n \\ \tilde{\mathbf{w}}_n \end{pmatrix}, \tag{2.10}$$

is a singular sequence for $\mathbf{A}-\lambda^*\mathbf{I}$ and hence for $\mathbf{A}_B-\lambda^*\mathbf{I}$ since $\lambda^*\in\omega$. Indeed, conditions (i_0) and (i_1) are verified for construction, and since

$$(2.11)\qquad (\mathbf{A}-\lambda^*\mathbf{I})\,\mathbf{w}'_n=\begin{pmatrix} (-\mathbf{A}_{NN}\mathbf{P}+\mathbf{I})\,\mathbf{A}_{NQ}\tilde{\mathbf{w}}_n+\lambda^*\,\mathbf{P}\mathbf{A}_{NQ}\tilde{\mathbf{w}}_n, \\ (\mathbf{S}-\lambda^*\mathbf{I})\tilde{\mathbf{w}}_n, \end{pmatrix},$$

we have $(\mathbf{S}-\lambda^*\mathbf{I})\tilde{\mathbf{w}}_n\to 0$ strongly in $\mathbf{H}$, and $((-\mathbf{A}_{NN}\mathbf{P}+\mathbf{I})\mathbf{A}_{NQ}+\lambda^*\mathbf{P}\mathbf{A}_{NQ})\tilde{\mathbf{w}}_n\to 0$ strongly in $\mathbf{H}$ (using that $\tilde{\mathbf{w}}_n\to 0$ weakly in $\mathbf{H}$ and that $(-\mathbf{A}_{NN}\mathbf{P}+\mathbf{I})\mathbf{A}_{NQ}+\lambda^*\mathbf{P}\mathbf{A}_{NQ}$ is a regularizing operator); then $(\mathbf{A}-\lambda^*\mathbf{I})\mathbf{w}_n\to 0$ strongly in $\mathbf{H}$. The condition (i_2) is therefore satisfied, and hence, in the hypotheses defined by the assumptions (A.I)–(A.III), we get (i_4).

Moreover, we denote by T_n the application $T_nf=n^{d/2}f(nx)$ which defines an isometry of L^2 in itself; we see that $T_n\mathbf{w}$ concentrates the support of $\tilde{\mathbf{w}}_n$ (and hence of $\mathbf{w}'_n$, since the support of $\mathbf{P}\mathbf{A}_{NQ}\tilde{\mathbf{w}}_n$ is close to $\tilde{\mathbf{w}}_n$) near the origin (i.e., $\mathbf{x}^*=0$).

It follows that for any given $\varepsilon>0$ there exists an index n_ε such that for $n>n_\varepsilon$ the function $\mathbf{w}_n$ has compact support in Ω_ε, where $\Omega_\varepsilon\subset\Omega$ and ε is the positive distance from Ω_ε to the boundary of Ω. That implies on the boundary of Ω

$$(2.12)\qquad (\mathbf{C}\mathbf{w}_n,\mathbf{C}\mathbf{w}_n)_\Gamma\to 0 \qquad \text{as} \qquad n\to\infty,$$

and hence (i_5) is verified. □

2.3. The asymptotics of the unbounded discrete spectrum. When $\mathbf{A}_B$ is selfadjoint lower bounded, the essential spectrum is bounded and there is an unbounded discrete spectrum. Following the works of Grubb and Geymonat [7], [8], we find that there exists a sequence of real eigenvalues λ_j^+ of finite multiplicity and disjoint from the essential spectrum going to $+\infty$. Let λ^o be large enough in order that $\lambda^o\notin\sigma_{ess}(\mathbf{A}_B)$; then the asymptotic behavior of λ_j^+ is given in first approximation by the asymptotic behavior of the eigenvalues of $\mathbf{A}_{NN}$. That is, there exists a constant $c(\mathbf{A}_{NN})$ such that

$$(2.13)\qquad N(\lambda,\mathbf{A})=\sum_{\lambda^o<\lambda_j^+<\lambda}1=c(\mathbf{A}_{NN})\lambda^{d/2m_{q-p}}+o(\lambda^{d/2m_{q-p}}),\quad \lambda\to\infty,$$

where

$$c(\mathbf{A}_{NN})=\frac{1}{d(2\pi)^d}\int_\Omega\int_{|\xi|=1}tr(p(\mathbf{x},\xi)^{-d/2m_{q-p}})d\sigma d\mathbf{x}$$

with

$$p(\mathbf{x},\xi)=\sigma^0(\mathbf{A}_{q-p,q-p})-\sigma^0(\mathbf{A}_{q-p,\hat{N}})\sigma^0(\mathbf{A}_{\hat{N},\hat{N}})^{-1}\sigma^0(\mathbf{A}_{\hat{N},q-p})$$

and $\hat{N}=\{1,2,\dots,q-p-1\}$. We point out that the asymptotic behavior defined by (2.13) does not depend on the choice of the boundary conditions.

Examples (continued). The eigenvalue problem for the operator $\mathbf{A}^m$ reads

$$(2.14\text{a})\qquad -a^{\alpha\beta\sigma\mu}\gamma_{\sigma\mu}(\mathbf{v})_{|\beta}-\lambda a^{\alpha\beta}v_\beta=0,\qquad \alpha=1,2,$$

$$(2.14\text{b})\qquad -a^{\alpha\beta\sigma\mu}\gamma_{\sigma\mu}(\mathbf{v})b_{\alpha\beta}-\lambda v_3=0.$$

With reference to the operators $\mathbf{A}_D^m$, $\mathbf{A}_N^m$, and $\mathbf{A}_I^m$ defined above, we may state that if λ in the system (2.14) is large enough, the asymptotic behavior of that part of the discrete spectrum, formed by eigenvalues with finite multiplicity which accumulate to $+\infty$, can be easily computed according to (2.13). In each of the above situations (examples (a), (b), and (c)), taking into account that $m_{q-p} = m_1 = m_2 = 1$, we have

$$c(\mathbf{A}^m_{NN}) = \frac{1}{2(2\pi)^2} \int_\Omega \int_{|\xi|=1} tr(\sigma^0(\mathbf{A}^m_{NN})^{-1}) d\sigma d\mathbf{x}$$

with

$$\sigma^0(\mathbf{A}^m_{NN}) = (a_{\alpha\beta})(t^{\sigma\mu})$$

and

$$t^{\sigma\mu} = a^{\sigma\gamma\mu\alpha}\xi_\alpha\xi_\gamma, \qquad \sigma, \mu = 1, 2.$$

Although the asymptotic behavior of the unbounded discrete spectrum is independent on the choice of the boundary conditions, we remark that the essential spectrum and hence the problems related to the operators $\mathbf{A}^m_D$, $\mathbf{A}^m_N$, and $\mathbf{A}^m_I$ may be very different. We recall that for the axially symmetric deformations of hemispherical membranes, as considered in [4], [5], and [17], a simple computation of the essential spectrum gives $(1-\nu^2) \in \omega$. In this situation, since we treat a one-dimensional problem, we have to restrict our attention to the boundary conditions (a) and (b). In both cases the discrete spectrum is formed by two disjoint subsequences of eigenvalues: one going to ∞ and the other one with an accumulation point $(1-\nu^2)$.

3. A noncontrollability result.

In the hypotheses assumed for the operator $\mathbf{A}$ in the previous section, we consider the following exact controllability problem.

(EC) Given $T > 0$ and an initial state $\mathbf{\Phi}^0, \mathbf{\Phi}^1$ find the control function $\mathbf{g}$ such that the unique solution $\mathbf{\Phi}$ of

$$\ddot{\mathbf{\Phi}} + \mathbf{A\Phi} = 0 \qquad \text{in } Q = \Omega \times (0,T), \tag{3.1a}$$

$$\mathbf{B\Phi} = \mathbf{g} \qquad \text{on } \Sigma = \Gamma \times (0,T), \tag{3.1b}$$

$$\mathbf{\Phi}(0) = \mathbf{\Phi}^0 \ , \quad \dot{\mathbf{\Phi}}(0) = \mathbf{\Phi}^1 \quad \text{in } \Omega \tag{3.1c}$$

satisfies the following conditions:

$$\mathbf{\Phi}(T) = 0 \ , \quad \dot{\mathbf{\Phi}}(T) = 0 \quad \text{in } \Omega. \tag{3.1d}$$

To solve the exact controllability problem, we have to look at the existence of a solution $\mathbf{g}$ of the functional equation

$$\int_\Sigma \mathbf{C}\boldsymbol{\eta}\ \mathbf{g}(s,t)\,\mathrm{d}\,s\,\mathrm{d}\,t = ((\boldsymbol{\eta}^0, \mathbf{\Phi}^1)) - ((\boldsymbol{\eta}^1, \mathbf{\Phi}^0)) \tag{3.2}$$

for all initial data $\{\boldsymbol{\eta}^0, \boldsymbol{\eta}^1\}$ of the homogeneous problem associated to (3.1a)–(3.1c) which we refer to in its variational setting.

Let $T > 0$, $\boldsymbol{\eta}^0 \in \mathbf{V}$, and $\boldsymbol{\eta}^1 \in \mathbf{H}$ be given; find a function

$$\boldsymbol{\eta}(t) \in C([0,T];\ \mathbf{V}) \cap C^1([0,T]; \mathbf{H})$$

such that for any $\tilde{\boldsymbol{\eta}} \in \mathbf{V}$

$$((\ddot{\boldsymbol{\eta}}, \tilde{\boldsymbol{\eta}}))_\Omega + a(\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}}) = 0, \quad \boldsymbol{\eta}(0) = \boldsymbol{\eta}^0,\ \dot{\boldsymbol{\eta}}(0) = \boldsymbol{\eta}^1,\ \mathbf{C}\boldsymbol{\eta} \in L^2(\Sigma). \tag{3.3}$$

The existence, uniqueness, and regularity of the solution $\boldsymbol{\eta}$ of (3.3) follow from the assumptions (A.I)–(A.III) in section 2.

We have the existence of a solution $\mathbf{g}$ of (3.2) *if and only if* [3]

$$\left(\int_\Sigma (\mathbf{C}\boldsymbol{\eta})^2 \,\mathrm{d}\,s\,\mathrm{d}\,t\right)^{1/2} \geq \mathrm{const}\|\{\boldsymbol{\eta}^0, \boldsymbol{\eta}^1\}\|_{V\times H}. \tag{3.4}$$

The Hilbert uniqueness method of J. L. Lions [10], [13] gives a constructive result for exact controllability problems. If we put $\mathbf{g} = \mathbf{C}\mathbf{v}$ in (3.2), the problem is to find the initial data $\{\mathbf{v}^0, \mathbf{v}^1\}$ of the homogeneous problem which solve the functional equation

$$\int_\Sigma \mathbf{C}\boldsymbol{\eta} \quad \mathbf{C}\mathbf{v} \,\mathrm{d}\,s\,\mathrm{d}\,t = ((\boldsymbol{\eta}^0, \boldsymbol{\Phi}^1)) - ((\boldsymbol{\eta}^1, \boldsymbol{\Phi}^0)).$$

If we show that for T large enough

$$\left(\int_\Sigma (\mathbf{C}\mathbf{v})^2 \,\mathrm{d}\,s\,\mathrm{d}\,t\right)^{1/2} = \|\{\mathbf{v}^0, \mathbf{v}^1\}\|_{\mathbf{F}}$$

defines a *norm* on the set of initial data $\{\mathbf{v}^0, \mathbf{v}^1\} \in \mathbf{F} = \mathbf{V} \times \mathbf{H}$ of the homogeneous problem, then the controllability problem can be solved, and we have exact controllability for any $\{\boldsymbol{\Phi}^1, \boldsymbol{\Phi}^0\} \in \mathbf{F}'$ ($\mathbf{F}' = \mathbf{V}' \times \mathbf{H}$). Indeed, after introducing the linear operator $\Lambda : \mathbf{F} \to \mathbf{F}'$ we can solve the equation

$$\langle \Lambda\{\mathbf{v}^0, \mathbf{v}^1\}, \{\boldsymbol{\eta}^0, \boldsymbol{\eta}^1\}\rangle = ((\boldsymbol{\eta}^0, \boldsymbol{\Phi}^1)) - ((\boldsymbol{\eta}^1, \boldsymbol{\Phi}^0))$$

for any $\{\boldsymbol{\eta}^0, \boldsymbol{\eta}^1\}$ in $\mathbf{F}$ and

$$\langle \Lambda\{\mathbf{v}^0, \mathbf{v}^1\}, \{\mathbf{v}^0, \mathbf{v}^1\}\rangle = \int_\Sigma (\mathbf{C}\mathbf{v})^2 \,\mathrm{d}\,s\,\mathrm{d}\,t.$$

We are now in a position to prove the main result of this paper.

THEOREM 3.1. *We assume that the hypotheses* (A.I)–(A.III) *and the following assumption* (A.IV) *are satisfied.*

(A.IV) *There is a positive* λ^* *in the set* ω.

For each finite positive T, there exist some initial data $\{\boldsymbol{\Phi}^1, \boldsymbol{\Phi}^0\} \in \mathbf{F}' = \mathbf{V}' \times \mathbf{H}$ *such that the evolution system* (3.1a), (3.1b), (3.1c) *is not exactly controllable.*

Proof. We consider the singular sequence $\{\mathbf{w}_n\}$ for $\lambda^* \in \omega$ as we introduced in Proposition 2.2. If we choose the sequence $\{\mathbf{v}_n^0, \mathbf{v}_n^1\} \in \mathbf{V} \times \mathbf{H}$ of initial data for the homogeneous problem associated to (3.1a), (3.1b), (3.1c), with

$$\mathbf{v}_n^0 = \mathbf{w}_n\ , \qquad \mathbf{v}_n^1 = 0,$$

from (i_1) of Proposition 2.1, we have

$$\|\{\mathbf{v}_n^0, \mathbf{v}_n^1\}\|_{H\times H} = \|\mathbf{v}_n^0\|_H = 1.$$

Moreover, from assumption (A.IV),

$$\|\{\mathbf{v}_n^0, \mathbf{v}_n^1\}\|_{V\times H}^2 = \|\mathbf{v}_n^0\|_V^2 = a(\mathbf{v}_n^0, \mathbf{v}_n^0) \to \lambda^* > 0 \quad \text{as} \quad n \to \infty. \tag{3.5}$$

We denote by $\mathbf{v}_n$ the unique solution of the homogeneous problem

$$\ddot{\mathbf{v}}_n + \mathbf{A}\mathbf{v}_n = 0 \qquad \text{in } \; Q = \Omega \times (0,T), \tag{3.6}$$

$$\mathbf{B}\mathbf{v}_n = \mathbf{0} \qquad \text{on } \; \Sigma = \Gamma \times (0,T), \tag{3.7}$$

$$\mathbf{v}_n(0) = \mathbf{w}_n \;, \quad \dot{\mathbf{v}}_n(0) = \mathbf{0} \quad \text{in } \; \Omega, \tag{3.8}$$

and we introduce the function $\mathbf{f}_n = \cos(\sqrt{\lambda^*}\, t)\, \mathbf{w}_n$.

From the definition of singular sequence we have $\lambda^* \mathbf{w}_n - \mathbf{A}\mathbf{w}_n \to 0$ strongly in $\mathbf{H}$ as $n \to \infty$; moreover, we have that $\mathbf{S}_n = \ddot{\mathbf{f}}_n + \mathbf{A}\mathbf{f}_n = -\cos(\sqrt{\lambda^*}\, t)(\lambda^* \mathbf{w}_n - \mathbf{A}\mathbf{w}_n) \to 0$ strongly in $L^\infty(0,T;\mathbf{H})$ as $n \to \infty$. Now we consider the function $\mathbf{e}_n = \mathbf{v}_n - \mathbf{f}_n$ which is in $C([0,T];\mathbf{V}) \cap C^1([0,T];\mathbf{H})$ and satisfies

$$\ddot{\mathbf{e}}_n + \mathbf{A}\mathbf{e}_n = \mathbf{S}_n \qquad \text{in } \; Q = \Omega \times (0,T), \tag{3.9}$$

$$\mathbf{B}\mathbf{e}_n = \mathbf{0} \qquad \text{on } \; \Sigma = \Gamma \times (0,T), \tag{3.10}$$

$$\mathbf{e}_n(0) = \mathbf{0}, \quad \dot{\mathbf{e}}_n(0) = \mathbf{0} \quad \text{in } \; \Omega. \tag{3.11}$$

More regularity of the function $\mathbf{e}_n$ is a consequence of the compatibility condition and the regularity of the function $\mathbf{S}_n$. Indeed, we can introduce, by means of the standard technique involving the incremental quotient $(\mathbf{e}_n(t+dt) - \mathbf{e}_n(t))/dt$, the differential system in the unknown $\dot{\mathbf{e}}_n = \mathbf{k}_n$, that is,

$$\ddot{\mathbf{k}}_n + \mathbf{A}\mathbf{k}_n = \dot{\mathbf{S}}_n \qquad \text{in } \; Q = \Omega \times (0,T), \tag{3.12}$$

$$\mathbf{B}\mathbf{k}_n = \mathbf{0} \qquad \text{on } \; \Sigma = \Gamma \times (0,T), \tag{3.13}$$

$$\mathbf{k}_n(0) = \mathbf{0}, \quad \dot{\mathbf{k}}_n(0) = \mathbf{S}_n(0) - \mathbf{A}\mathbf{e}_n(0) = \mathbf{A}\mathbf{w}_n - \lambda^* \mathbf{w}_n \quad \text{in } \; \Omega. \tag{3.14}$$

Multiplying (3.12) by $\dot{k}_n$, it follows from Gronwall's inequality the energy estimate

$$E(\mathbf{k}_n(t)) = \frac{1}{2}\left\{ a(\mathbf{k}_n, \mathbf{k}_n) + \int_\Omega |\dot{\mathbf{k}}_n|^2 \right\} \le \frac{e^T}{2}\, [\|\mathbf{A}\mathbf{w}_n - \lambda^* \mathbf{w}_n\|_H^2 + T\|\dot{\mathbf{S}}_n\|^2_{L^\infty(0,T;H)}],$$

and since $\dot{\mathbf{S}}_n = \sqrt{\lambda^*}\; \sin(\sqrt{\lambda^*}\, t)(\lambda^* \mathbf{w}_n - \mathbf{A}\mathbf{w}_n)$, we obtain

$$E(\mathbf{k}_n(t)) \le C_1(T)\, \|\mathbf{A}\mathbf{w}_n - \lambda^* \mathbf{w}_n\|_H^2. \tag{3.15}$$

The previous inequality allows us to multiply (3.9) by $\mathbf{A}\dot{\mathbf{e}}_n$ and get an a priori estimate for the function $\mathbf{e}_n$. We have

$$\tilde{E}(\mathbf{e}_n(t)) = \frac{1}{2}\left\{ a(\dot{\mathbf{e}}_n, \dot{\mathbf{e}}_n) + \int_\Omega |\mathbf{A}\mathbf{e}_n|^2 \right\} = \int_Q \mathbf{S}_n\, \mathbf{A}\dot{\mathbf{e}}_n = \int_\Omega \mathbf{S}_n\, \mathbf{A}\dot{\mathbf{e}}_n - \int_0^t \int_\Omega \dot{\mathbf{S}}_n\, \mathbf{A}\mathbf{e}_n.$$

With simple algebraic manipulations, again from the Gronwall's inequality and from the definition of $\mathbf{S}_n$ and $\dot{\mathbf{S}}_n$, we get

$$\tilde{E}(\mathbf{e}_n(t)) \le C(T)\, \|\mathbf{A}\mathbf{w}_n - \lambda^* \mathbf{w}_n\|_H^2, \tag{3.16}$$

where $C(T)$ is a constant depending on T such that $\lim_{T\to\infty} C(T) = \infty$, and hence for any fixed positive T

$$\mathbf{e}_n \in L^\infty(0,T;\, D(\mathbf{A}_B^{min})), \qquad \dot{\mathbf{e}}_n \in L^\infty(0,T;\, \mathbf{V}).$$

Moreover, taking into account our assumptions, $\mathbf{A}_B^{min}$ is a maximal and monotone operator and from the Hille–Yosida theorem we have $\mathbf{e}_n \in C(0,T;D(\mathbf{A}_B^{min}))$ and $\mathbf{Ce}_n \in L^2(\Sigma)$. In these regularity hypotheses the boundary operator $\mathbf{C}$ is continuous and surjective [8], [9] and from the estimate (3.16)

$$\int_\Sigma (\mathbf{Ce}_n)^2 dsdt \to 0 \qquad \text{as } n \to \infty;$$

hence from Proposition 2.2

$$\int_\Sigma (\mathbf{Cv}_n)^2 dsdt \le 2\left\{\int_\Sigma (\mathbf{Ce}_n)^2 dsdt + \int_\Sigma (\mathbf{Cf}_n)^2 dsdt\right\} \to 0 \qquad \text{as } n \to \infty.$$

The last condition and the behavior of (3.5) are in contradiction with the necessary (and sufficient) condition for the exact controllability

$$\int_\Sigma (\mathbf{Cv}_n)^2\, dsdt \ \ge c\|\{\mathbf{v}_n^0, \mathbf{v}_n^1\}\|^2_{V\times H},$$

and that completes the proof. □

Examples (continued). The procedure proposed in the proof of the Proposition 2.2 can be applied to the operator $\mathbf{A}^m$ (see (2.6)). The vibrations of the membrane shell are described by the system

$$a^{\alpha\beta\sigma\mu}\gamma_{\sigma\mu}(\mathbf{z})_{|\beta} = a^{\alpha\beta}(z_\beta)_{tt}, \qquad \alpha = 1,2,$$

$$a^{\alpha\beta\sigma\mu}\gamma_{\sigma\mu}(\mathbf{z})b_{\alpha\beta} = (z_3)_{tt}.$$

On the boundary we assume one of these conditions:
$\mathrm{z}_\alpha = g_\alpha, \qquad \alpha = 1,2$ (the example (a));
$a^{\alpha\beta\sigma\mu}\gamma_{\sigma\mu}(\mathbf{z})\nu_\beta = g_\alpha, \qquad \alpha = 1,2$ (the example (b));
$\mathrm{z}_\alpha\, \nu_\alpha = g_1, \quad a^{\alpha\beta\sigma\mu}\gamma_{\sigma\mu}(\mathbf{z})\nu_\beta\, \tau_\alpha = g_2$ (the example (c)).
And at the initial time we prescribe

$$\mathrm{z}_\alpha(0) = \mathrm{z}_\alpha^0, \qquad (\mathrm{z}_\alpha)_t(0) = \mathrm{z}_\alpha^1.$$

The controllability problem requires us to find the control function g_α such that at a given time T

$$\mathrm{z}_\alpha(T) = 0, \qquad (\mathrm{z}_\alpha)_t(T) = 0.$$

From Theorem 3.1 it follows that the control function g_α $(\alpha = 1,2)$ does not exist for any choice of the initial data. The eigenfunctions corresponding to the eigenvalues with finite accumulation point are used to prove the lack of the exact controllability for spherical membranes (see [4] and [5]). We briefly mention that case. The axially symmetric vibrations are described by two unknowns u and w: the meridional and radial component, respectively, of the displacement vector $\mathbf{v} = (u, w)$. The equations for the hemispherical membrane vibrations take the form

$$u_{tt} = u'' + u'\mathrm{cot}\theta - u(\nu + \mathrm{cot}^2\theta) - (1+\nu)w',$$
$$w_{tt} = \frac{(1+\nu)}{\mathrm{sin}\theta}(u\mathrm{sin}\theta)' - 2(1+\nu)w,$$

where the prime stands by the derivative with respect to $\theta \in]0, \theta_0 = \frac{\pi}{2}[$. We consider the following types of homogeneous boundary conditions:

(a) $u(\theta_0, t) = 0$ (Dirichlet condition);

(b) $u' - (1+\nu)w|_{(\theta_0,t)} = 0$ (Neumann condition).

The corresponding boundary conditions in 0 are a consequence of the symmetry of the problem. The spectral analysis of these problems can be easily carried out (see [4], [5] for (a)); in particular, we recall that the related eigenfunctions are derived by the spherical functions of even order (in case (a)) and odd order (in case (b)).

The same spectral asymptotics allow us to treat the problems in an analogous way. The positive number $(1-\nu^2)$ which is in the set ω is the accumulation point of two subsequences of eigenvalues (for the eigenvalues problems defined by (a) and (b), respectively). In this particular situation a more simple proof of Theorem 3.1 can be established if we take the corresponding subsequences of eigenfunctions as initial data of the control problems. Indeed, we denote these last subsequences by $\mathbf{v}_n^{(a)}$ and $\mathbf{v}_n^{(b)}$, and we easily check as $n \to \infty$ that $(u_n^{(a)}(\theta_0))' - (1+\nu)w_n^{(a)}(\theta_0) \to 0$ and $u_n^{(b)}(\theta_0) \to 0$.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[2] P. G. Ciarlet and E. Sanchez-Palencia, *An existence and uniqueness theorem for the two dimensional linear membrane shell equations*, J. Math. Pures Appl. (9), 75 (1996), pp. 51–67.

[3] G. Fichera, *Linear Elliptic Differential Systems and Eigenvalue Problems*, Springer-Verlag, Berlin, New York, 1965.

[4] G. Geymonat, P. Loreti, and V. Valente, *Exact controllability of thin elastic hemispherical shell via harmonic analysis*, in Boundary Value Problems for Partial Differential Equations and Applications, Masson, Paris, 1993, pp. 379–385.

[5] G. Geymonat, P. Loreti, and V. Valente, *Spectral problem for thin shells and exact controllability*, in Spectral Analysis of Complex Structures, Travaux en Cours 49, Hermann, Paris, 1995, pp. 35–57.

[6] G. Grubb, *Essential spectra of elliptic systems on compact manifolds*, in Spectral Analysis, Centro Internazionale Matematico Estivo (C.I.M.E), Edizione Cremonese Roma, 1974, pp. 143–170.

[7] G. Grubb and G. Geymonat, *The essential spectrum of elliptic systems of mixed order*, Math. Ann., 227 (1977), pp. 247–276.

[8] G. Grubb and G. Geymonat, *Eigenvalue asymptotics for selfadjoint elliptic mixed order systems with nonempty essential spectrum*, in Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8) 16, Zanichelli Bologna, 1979, pp. 1032–1048.

[9] G. Grubb, *On the coerciveness of Douglis-Nirenberg elliptic boundary value problems*, in Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8) 16, Zanichelli Bologna, 1979, pp. 1049–1080.

[10] J. E. Lagnese and J. L. Lions, *Modelling Analysis and Control of Thin Plates*, Masson, Paris, 1988.

[11] I. Lasiecka, *Mathematical control theory in structural acoustic problems*, Math. Models Methods Appl. Sci., 8 (1998), pp. 1119–1153.

[12] I. Lasiecka, R. Triggiani, and V. Valente, *Uniform stabilization of a spherical shells with boundary dissipation*, Adv. Differential Equations, 4 (1996), pp. 635–674.

[13] J. L. Lions, *Contrôlabilité exacte, perturbations et stabilization des systèmes distribués*, 1 and 2, Masson, Paris, 1988.

[14] B. Miara and V. Valente, *Exact controllability of a Koiter shell by a boundary action*, J. Elasticity, 52 (1999), pp. 267–287.

[15] R. S. Palais, *Seminar on the Atiyah-Singer Index Theorem*, Ann. of Math. Stud. 57, Princeton University Press, Princeton, NJ, 1965.

[16] J. Sanchez-Hubert and E. Sanchez-Palencia, *Vibration and Coupling of Continuous Systems*, Springer-Verlag, Berlin, New York, 1989.

[17] V. Valente, *Relaxed exact spectral controllability of membrane shells*, J. Math. Pures Appl. (9), 76 (1997), pp. 551–562.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 673–681

AVERAGE COST DYNAMIC PROGRAMMING EQUATIONS FOR CONTROLLED MARKOV CHAINS WITH PARTIAL OBSERVATIONS[*]

V. S. BORKAR[†]

Abstract. The value function for the average cost control of a class of partially observed Markov chains is derived as the "vanishing discount limit," in a suitable sense, of the value functions for the corresponding discounted cost problems. The limiting procedure is justified by bounds derived using a simple coupling argument.

Key words. average cost control, controlled Markov chains, partial observations, dynamic programming, vanishing discount limit

AMS subject classification. 93E20

PII. S0363012998345172

1. Introduction.

Deriving the dynamic programming equations for average cost control of partially observed Markov chains has been an elusive task. Only scattered results are available which achieve this under very restrictive conditions [5], [6], [9], [10], [11]. (See [1], section 7, for a slightly dated survey.) Our aim here is to achieve this under a fairly broad condition. This is done by first relaxing the control problem to include the class of the so-called "wide sense admissible" controls, a notion borrowed from continuous time stochastic control literature [7], and then using a simple coupling argument to get the kind of bounds we need on the discounted cost value function in order to justify the "vanishing discount" limit.

The paper is organized as follows. The next section introduces the notation and the formal statement of the average cost control problem. It also introduces the associated nonlinear filter and the class of "wide sense admissible" controls alluded to above. Section 3 contains the coupling argument leading to the key estimates on the discounted value functions. Section 4 presents the main results, viz., the derivation of the average cost dynamic programming equation as the vanishing discount limit of the dynamic programming equations for the discounted cost problem.

2. The control problem.

We consider a controlled Markov chain $\{X_n\}$ on a finite state space $S = \{1, 2, \ldots, s\}$, controlled by a control process $\{Z_n\}$ taking values in a compact metric "action space" A, and with an associated observation process $\{Y_n\}$ taking values in a finite "observation space" W with cardinality $d \geq 1$. These are realized on an underlying probability space $(\Omega, \mathcal{F}, P)$. The evolution law of (X_n, Y_n) is given by

$$P(X_{n+1} = i, Y_{n+1} = j / X_m, Y_m, Z_m, m \leq n) = p(X_n, Z_n, i, j) \tag{2.1}$$

for $i \in S, j \in W$, where $p : X \times A \times S \times W \to [0, 1]$ is a continuous function satisfying

$$\sum_{j,k} p(i, u, j, k) = 1 \ \forall i, u.$$

[*] Received by the editors September 28, 1998; accepted for publication (in revised form) April 21, 2000; published electronically October 4, 2000.
http://www.siam.org/journals/sicon/39-3/34517.html

[†] School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@tifr.res.in). This research was supported in part by grant DST/MS/III-045/96 of the Department of Science and Technology, Government of India.

Call $\{Z_n\}$ strict sense admissible if it is adapted to $\sigma(Y_m, m \leq n),\ n \geq 0$. The average cost control problem under partial observations, in its original formulation, is to minimize over all such $\{Z_n\}$ the average cost

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} E[h(X_m, Z_m)] \tag{2.2}$$

for a prescribed "running cost" function $h \in C(S \times A)$.

We shall denote by $\mathcal{P}(S)$ the space of probability measures on S, with topology of weak convergence. Since S is finite, this is the simplex of probability vectors in R^s. Keeping in mind the larger class of $\{Z_n\}$ we introduce later, define $\mathcal{F}_n = \sigma(Y_m, Z_m, m \leq n),\ n \geq 0$, and let $\{\pi_n\}$ be the $\mathcal{P}(S)$-valued process of regular conditional laws of X_n given $\mathcal{F}_n$ for $n \geq 0$. Define $\bar{h} \in C(\mathcal{P}(S) \times A)$ by

$$\bar{h}(\nu, a) = \sum_{i\in S} \nu(i) h(i, a),$$

where we write $\nu(i)$ for $\nu(\{i\})$ by abuse of notation. The average cost control problem above can then be shown to be equivalent to the "separated" average cost control problem of controlling the $\mathcal{P}(S)$-valued controlled Markov process $\{\pi_n\}$, so as to minimize over strict sense admissible $\{Z_n\}$ the quantity

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} E[\bar{h}(\pi_m, Z_m)]. \tag{2.3}$$

This in fact equals (2.2), as seen by a simple conditioning argument. The dynamics of $\{\pi_n\}$ are given by the nonlinear filter (derivable by an elementary application of the Bayes rule—see, e.g., [3, Chapter VIII])

$$\pi_{n+1} = \pi_n P(Z_n, Y_{n+1})/(\pi_n P(Z_n, Y_{n+1}) 1_c),\ n \geq 0, \tag{2.4}$$

where $\{\pi_n\}$ is written as a row vector, $P(u, k)$ is the matrix $[[p(\cdot, u, \cdot, k)]]$ for $u \in A, k \in W$, and 1_c is the column vector of all l's. The initial condition π_0 is the known law of X_0.

Before we proceed, we need to exhibit $\{X_n\}$ explicitly as a noise-driven dynamical system, i.e., as

$$X_{n+1} = F(X_n, Z_n, \xi_n),\ n \geq 0, \tag{2.5}$$

where $F : S \times A \times [0, 1] \to S$ is measurable and $\{\xi_n\}$ are independently and identically distributed (i.i.d.) uniformly on [0,1]. This is possible by the stochastic realization theoretic results of [2] and may require an augmentation of the underlying probability space.

We shall now reproduce (in law) the above processes on a more convenient probability space. Let $\bar{\Omega} = [0, 1]^\infty \times A^\infty \times W^\infty \times S$, and let $\mathcal{F}$ be the corresponding product σ-field. Let (u, v, w, x) denote a typical element of $\bar{\Omega}$ with $u = [u_0, u_1, \ldots] \in [0, 1]^\infty, v = [v_0, v_1, \ldots] \in A^\infty, w = [w_0, w_1, \ldots] \in W^\infty$. Let $u^n = [u_0, u_1, \ldots, u_n]$, and define v^n, w^n accordingly for $n \geq 0$. By the definition of strict sense admissible controls, we have

$$Z_n = \psi_n(Y_0, \ldots, Y_n),\ n \geq 0, \tag{2.6}$$

for some $\psi_n : W^{n+1} \to A$. Define

$$\bar{\psi}_n(w^n) = [\psi_0(w^0), \psi_1(w^1), \dots, \psi_n(w^n)],\ n \geq 0.$$

Let l_1, l_2 denote the uniform probability measures on [0,1], W, resp. Define a probability measure P_0 on $(\bar{\Omega}, \bar{\mathcal{F}})$ by the following. If $B_1 \subset [0,1]^{n+1}$, $B_2 \subset A^{n+1}$, $B_3 \subset W^{n+1}$, and $B_4 \subset S$ are Borel, then

$$P_0\left(\Pi_{i=1}^4 B_i\right) = l_1^{n+1}(B_1) l_2^{n+1}(B_3 \cap \{w^n : \bar{\psi}_n(w^n) \in B_2\})\pi_0(B_4),$$

where l_i^n denote the appropriate product measures. Define processes $\{\xi_n\}$, $\{Z_n\}$, $\{Y_n\}$, and a random variable X_0 canonically on $(\bar{\Omega}, \bar{\mathcal{F}}, P_0)$ by

$$\begin{aligned} \xi_n(u,v,w,x) &= u_n, \\ Z_n(u,v,w,x) &= v_n, \\ Y_n(u,v,w,x) &= w_n, \\ X_0(u,v,w,x) &= x \end{aligned}$$

for $n \geq 0$. Then under P_0,

(i) $\{\xi_i\}$ are i.i.d. uniformly distributed on [0,1],
(ii) $\{Y_n\}$ are i.i.d. uniformly distributed on W,
(iii) law of X_0 is π_0,
(iv) $(\{\xi_n\}, \{Y_n\}, X_0)$ is an independent family, and,
(v) $\{Z_n\}$ is specified by (2.6).

Define $\{X_n\}$ recursively using (2.5). Then by construction, $\{X_n\}$ is a controlled Markov chain satisfying

$$P(X_{n+1} = j / X_m,\ Z_m,\ m \leq n) = \bar{p}(X_n, Z_n, j),$$

$j \in S,\ n \geq 0$, where

$$\bar{p}(i,u,j) \triangleq \sum_k p(i,u,j,k),\ i,\ j \in S,\ u \in A.$$

For $n \geq 0$, let $\mathcal{G}_n = \sigma(X_m, Y_m, \xi_m, Z_m, m \leq n)$, and let P_{0n} be the restriction of P_0 to $(\bar{\Omega}, \mathcal{G}_n)$. Define a new probability measure $\bar{P}$ on $(\bar{\Omega}, \bar{\mathcal{F}})$ as follows. If $\bar{P}_n$ denotes its restriction to $(\bar{\Omega}, \mathcal{G}_n)$, then $\bar{P}_n << P_{0n}$ $\forall n$ with

$$\Lambda_n \triangleq \frac{d\bar{P}_n}{dP_{0n}} = \prod_{m=0}^{n-1} \frac{p(X_m, Z_m, X_{m+1}, Y_{m+1})}{\bar{p}(X_m, Z_m, X_{m+1})(1/d)},\ n \geq 0.$$

It is easily verified that $(\Lambda_n, \mathcal{G}_n)$ is a nonnegative martingale with mean 1, and therefore this defines in a consistent and unique manner a probability measure $\bar{P}$ on $(\Omega, \bigvee_n \mathcal{G}_n)$. Since $\mathcal{F} = \bigvee_n \mathcal{G}_n$ by construction, we are through. Furthermore, under $\bar{P}, (X_n, Y_n, Z_n, \xi_n)$ have the same joint law as the corresponding processes on $(\Omega, \mathcal{F}, P)$ we started with, again by construction.

This construction permits us to define wide sense admissible controls along the lines of [7]: $\{Z_n\}$ is said to be wide sense admissible if for each n, Z_n is independent of $(\{\xi_m\}, X_0, \{Y_i, i > n\})$ under P_0. Note that this includes strict sense admissible controls. (Intuitively, this relaxation allows for extraneous randomization of controls that does not use any information that it shouldn't.) Our "relaxed" partially observed separated control problem then is to minimize (2.3) over all wide sense admissible $\{Z_n\}$.

This is to be interpreted in the following sense. Under P_0, the laws of $\{Y_n\}, \{\xi_n\}, X_0$ are fixed and $\{X_n\}$ is specified once $\{Z_n\}$ is. Thus the above framework is specified "in law" by specifying the conditional law of $\{Z_n\}$ given $\{Y_n\}$ or, equivalently, the joint law of $\{Y_n\}, \{Z_n\}$ (where the marginal corresponding to the law of $\{Y_n\}$ is fixed). Thus it makes sense to refer to either of these as *the* wide sense admissible control. We denote by Φ the set of wide sense admissible controls, and, by a slight abuse of notation, we denote by $\{Z_n\}$ a typical element thereof.

3. The discounted value function. In anticipation of the "vanishing discount" argument to be used later, we consider here the family of value functions associated with the discounted cost problem, indexed by the discount factor $\alpha > 0$. Recall that the discounted cost under a wide sense admissible control $\{Z_n\} \in \Phi$ and initial law π is

$$J_\alpha(\{Z_n\}, \pi) \triangleq E\left[\sum_{m=0}^{\infty} \alpha^m h(X_m, Z_m)/\pi_0 = \pi\right]$$

$$\triangleq E\left[\sum_{m=0}^{\infty} \alpha^m \bar{h}(\pi_m, Z_m)/\pi_0 = \pi\right].$$

The associated value function

$$V_\alpha(\pi) = \inf_\Phi J_\alpha(\{Z_n\}, \pi)$$

then satisfies the dynamic programming equation

$$V_\alpha(\pi) = \min_u(\bar{h}(\pi, u) + \alpha \int \eta(\pi, u, d\pi')V_\alpha(\pi')), \ \pi \in \mathcal{P}(S), \tag{3.1}$$

where $(\pi, u) \in \mathcal{P}(S) \times A \to \eta(\pi, u, d\pi') \in \mathcal{P}(\mathcal{P}(S))$ is the controlled transition kernel of the $\mathcal{P}(S)$-valued controlled Markov process $\{\pi_n\}$. From (2.4) and conditions on $p(\cdot,\cdot,\cdot,\cdot)$, one easily verifies that $\eta(\cdot,\cdot,d\pi')$ is a continuous map. (See [3, Chapter VIII] for a detailed treatment of the separated discounted cost problem.)

We shall need to compare $V_\alpha(\cdot)$ for two different values of its argument. With this in view, we construct on a common probability space two controlled Markov chains satisfying (2.1) with a "common" $\{Z_n\} \in \Phi$, but different initial laws, say $\tilde{\pi}$ and $\hat{\pi}$. This is done by a small modification of the construction of the preceding section. Note that specification of $\{Z_n\} \in \Phi$ for initial law $\tilde{\pi}$ entails specification of its joint law with $\{Y_n\}$ under P_0, assumed to satisfy the independence/conditional independence constraints in the definition of wide sense admissible controls. Denote this joint law by $\phi(dy^\infty, dz^\infty) \in \mathcal{P}(W^\infty \times A^\infty)$. Define

$$\hat{\Omega} = ([0,1]^\infty \times S) \times ([0,1]^\infty \times S) \times A^\infty \times W^\infty \times W^\infty$$

with $\hat{\mathcal{F}} =$ the corresponding product σ-field and $\hat{P}_0$ the probability measure on $(\hat{\Omega}, \hat{\mathcal{F}})$ defined by

$$\begin{aligned}&\hat{P}_0((d\tilde{u}^\infty \times d\tilde{x}) \times (d\hat{u}^\infty \times d\hat{x}) \times dz^\infty \times d\tilde{y}^\infty \times d\hat{y}^\infty)\\ &\qquad = \ell_1^\infty(d\tilde{u}^\infty)\tilde{\pi}(d\tilde{x})\ell_1^\infty(d\hat{u}^\infty)\hat{\pi}(d\hat{x})\phi(d\tilde{y}^\infty, dz^\infty)\ell_2^\infty(d\hat{y}^\infty).\end{aligned}$$

On $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P}_0)$, define processes $\{\tilde{\xi}_n\}$, $\{\hat{\xi}_n\}$, $\{Z_n\}$, $\{\tilde{Y}_n\}$, $\{\hat{Y}_n\}$, and random variables $\tilde{X}_0, \hat{X}_0$ canonically as follows: For $\omega = (\tilde{u}^\infty, \tilde{x}, \hat{u}^\infty, \hat{x}, z^\infty, \tilde{y}^\infty, \hat{y}^\infty)$, let $\tilde{\xi}_n(\omega) =$

$\tilde{u}_n,\ \hat{\xi}_n(\omega) = \hat{u}_n,\ \tilde{X}_0 = \tilde{x},\ \hat{X}_0 = \hat{x},\ Z_n = z_n,\ \tilde{Y}_n = \tilde{y}_n,\ \hat{Y}_n = \hat{y}_n,\ n \geq 0$. Define $\{\tilde{X}_n\}$, $\{\hat{X}_n\}$ recursively by

$$\begin{aligned} \tilde{X}_{n+1} &= F(\tilde{X}_n, Z_n, \tilde{\xi}_n), \\ \hat{X}_{n+1} &= F(\hat{X}_n, Z_n, \hat{\xi}_n), \end{aligned}$$

$n \geq 0$. For $n \geq 0$, let $\Gamma_n = \sigma(\hat{X}_m, \tilde{X}_m, \hat{\xi}_m, \tilde{\xi}_m, \hat{Y}_m, \tilde{Y}_m, Z_m,\ m \leq n)$. Then $\hat{\mathcal{F}} = \bigvee_n \Gamma_n$. Define a new probability measure $\hat{P}$ on $(\hat{\Omega}, \hat{\mathcal{F}})$ by the following. If $\hat{P}_n, \hat{P}_{0n}$ are restrictions of $\hat{P}, \hat{P}_0$, resp. to $(\hat{\Omega}, \Gamma_n)$, then

$$\hat{\Lambda}_n \triangleq \frac{d\hat{P}_n}{d\hat{P}_{0n}} = \prod_{m=0}^{n-1} \left(\frac{p(\tilde{X}_m, Z_m, \tilde{X}_{m+1}, \tilde{Y}_{m+1})}{\bar{p}(\tilde{X}_m, Z_m, \tilde{X}_{m+1})(1/d)} \right) \left(\frac{p(\hat{X}_m, Z_m, \hat{X}_{m+1}, \hat{Y}_{m+1})}{\bar{p}(\hat{X}_m, Z_m, \hat{X}_{m+1})(1/d)} \right),$$

$n \geq 0$. Then the controlled Markov chains $\{\tilde{X}_n\}$, $\{\hat{X}_n\}$ defined on $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ form the desired pair in so far as their initial laws are $\tilde{\pi}, \hat{\pi}$, resp., and they are governed by a "common" $\{Z_n\} \in \Phi$, as argued below.

LEMMA 3.1. $\{\tilde{X}_n\}$ *(resp.,* $\{\hat{X}_n\}$*) is a controlled Markov chain on* $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ *with associated observation process* $\{\tilde{Y}_n\}$ *(resp.,* $\{\hat{Y}_n\}$*) and wide sense admissible control* $\{Z_n\}$, *with its evolution governed by* (2.1).

Proof. It suffices to observe that if we consider the probability measure

$$P_1(d\tilde{u}^\infty \times d\tilde{x} \times dz^\infty \times d\tilde{y}^\infty) = \hat{P}(d\tilde{u}^\infty \times d\tilde{x} \times [0,1]^\infty \times S \times dz^\infty \times d\tilde{y}^\infty \times W^\infty)$$

on $[0,1]^\infty \times S \times A^\infty \times W^\infty$, it is precisely a special case of the construction in the preceding section of $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P})$. Thus $(\{\tilde{X}_n\}, \{\tilde{Y}_n\}, \{Z_n\})$ obey (2.1) with $\{Z_n\}$ wide sense admissible and the law of $\tilde{X}_0 = \tilde{\pi}_0$. Likewise, consider the probability measure

$$P_2(d\hat{u}^\infty \times d\hat{x} \times dz^\infty \times d\hat{y}^\infty) = \hat{P}([0,1]^\infty \times S \times d\hat{u}^\infty \times d\hat{x} \times dz^\infty \times W^\infty \times d\hat{y}^\infty)$$

on $[0,1]^\infty \times S \times A^\infty \times W^\infty$ to draw a similar conclusion about $(\{\hat{X}_n\}, \{\hat{Y}_n\}, \{Z_n\})$. □

What this construction achieves is to identify each wide sense admissible control $\{Z_n\}$ for $\tilde{\pi}$ with one wide sense admissible control for $\hat{\pi}$. (This identification can be many-one.) By a symmetric argument that interchanges the role of $\tilde{\pi}, \hat{\pi}$, one may identify every wide sense admissible control for $\hat{\pi}$ with one for $\tilde{\pi}$. Suppose $V_\alpha(\tilde{\pi}) \leq V_\alpha(\hat{\pi})$. Then for a wide sense $\{\tilde{Z}_n\}$ that is optimal for $\tilde{\pi}$,

$$\begin{aligned} |V_\alpha(\tilde{\pi}) - V_\alpha(\hat{\pi})| &= V_\alpha(\hat{\pi}) - V_\alpha(\tilde{\pi}) \\ &\leq J_\alpha(\{\tilde{Z}_n\}, \hat{\pi}) - J_\alpha(\{\tilde{Z}_n\}, \tilde{\pi}) \\ &\leq \sup_\Phi |J_\alpha(\{Z_n\}, \hat{\pi}) - J_\alpha(\{Z_n\}, \tilde{\pi}))|, \end{aligned}$$

where the above identification is used to derive the first and the second inequalities. If $V_\alpha(\tilde{\pi}) > V_\alpha(\hat{\pi})$, a symmetric argument works, interchanging the roles of $\tilde{\pi}, \hat{\pi}$. Thus

$$|V_\alpha(\tilde{\pi}) - V_\alpha(\hat{\pi})| \leq \sup_\Phi |J_\alpha(\{Z_n\}, \tilde{\pi}) - J_\alpha(\{Z_n\}, \hat{\pi})|,$$

a fact we use in Lemma 3.2 below.

The stage is now set for stating our key assumption in this paper. Let $\tau = \min\{n \geq 0 : \tilde{X}_n = \hat{X}_n\}(= \infty$ if the set in question is empty).

Assumption A. There exist $K_0 > 0, \delta \in (0, 1)$ such that

$$\sup_{i,j} \sup_{\Phi} P(\tau > n/\hat{X}_0 = i, \tilde{X}_0 = j) \leq K_0\delta^n, n \geq 0. \tag{3.2}$$

Note that for an uncontrolled chain, (3.2) would follow from irreducibility and aperiodicity. The latter, in fact, are necessary. In our case, however, we need a statement over all wide sense admissible controls, which is what (3.2) does. Simpler sufficient conditions imposing restrictions on the graph of the Markov chain can be given (see Appendix), but these would be much more restrictive. Note also that the above conditional probability summed over n gives the corresponding conditional expectation, which will then be bounded by $\frac{K_0}{1-\delta}$. As our proofs will show, the weaker hypothesis that

$$\sup_{i,j} \sup_{\Phi} E[\tau/\hat{X}_0 = i, \tilde{X}_0 = j] \leq K_0$$

will in fact suffice. However, in most interesting cases where this is available, so is (3.2).

Note also that $\tau = 0$ almost surely (a.s.) on $\{\hat{X}_0 = \tilde{X}_0\}$. Using this, we shall cast (3.2) in a more convenient form. View $S = \{1, 2, \ldots, s\} \subset R$, allowing S to inherit the Euclidean metric from R. Then for $i \neq j$ in $S, |i - j| \geq 1$. Thus without any loss of generality, we may rewrite (3.2) as

$$\sup_{\Phi} P(\tau > n/\hat{X}_0 = i, \tilde{X}_0 = j) \leq K_0\delta^n|i - j|, \; n \geq 0, i, \; j \in S. \tag{3.3}$$

The last bit of technicality we need is the metric $\rho(\cdot, \cdot)$ on $\mathcal{P}(S)$ defined by

$$\rho(\mu_1, \mu_2) = \inf E[|X_1 - X_2|],$$

where the infimum is over all pairs (X_1, X_2) of S-valued random variables such that the law of X_i is μ_i for $i = 1, 2$. This metric is equivalent to the Prohorov metric [4, p. 29] and thus metrizes the Prohorov topology which, because S is finite, is the same as the total variation norm topology, which in turn coincides with the Euclidean topology of $\mathcal{R}^s$ relativized to $\mathcal{P}(S)$. Note also that as $\mathcal{P}(S)$ is compact, every compatible metric on it, in particular ρ, is complete. The main result of this section is the following lemma.

LEMMA 3.2. *There exists a* $K_1 > 0$ *such that*

$$|V_\alpha(\mu_1) - V_\alpha(\mu_2)| \leq K_1\rho(\mu_1, \mu_2) \; \forall \mu_1, \mu_2 \in \mathcal{P}(S), \; \alpha > 0. \tag{3.4}$$

Proof. Consider the above construction with $\hat{\pi} = \mu_1, \tilde{\pi} = \mu_2$. Define an S-valued process $\bar{X}_m, m \geq 0$, and a W-valued process $\bar{Y}_m, m \geq 0$, by

$$\bar{X}_m = \hat{X}_m, \bar{Y}_m = \hat{Y}_m \quad \text{for} \quad m < \tau,$$

$$\bar{X}_m = \tilde{X}_m, \bar{Y}_m = \tilde{Y}_m \quad \text{for} \quad m \geq \tau.$$

That is $(\bar{X}_m, \bar{Y}_m)$ is obtained from $(\hat{X}_m, \hat{Y}_m)$ by "gluing" it to $(\tilde{X}_m, \tilde{Y}_m)$ from τ on, a standard construction in coupling arguments. We claim that $\{\bar{X}_m\}$ is a controlled Markov chain as in (2.1) with $\{\bar{Y}_m\}$ the associated observation process, with initial law $\hat{\pi} = \mu_1$ and the same control process $\{Z_n\} \in \Phi$ as before. To see this, observe first that the claim (except for the wide sense admissibility of $\{Z_n\}$) is equivalent to the statement that for $i \in S$, $j \in W$, $\sum_{m=1}^{n}(I\{\bar{X}_m = i, \bar{Y}_m = j\} - p(\bar{X}_{m-1}, Z_{m-1}, i, j))$, $n \geq 1$, is a $\{\Gamma_n\}$-martingale. The corresponding statements for $(\{\tilde{X}_n\}, \{\tilde{Y}_n\})$ and $(\{\hat{X}_n\}, \{\hat{Y}_n\})$ are immediate. Thus

$$E[(I\{\tilde{X}_n = i, \tilde{Y}_n = j\} - p(\tilde{X}_{n-1}, Z_{n-1}, i, j))/\Gamma_{n-1}] = 0$$

and

$$E[(I\{\hat{X}_n = i, \hat{Y}_n = j\} - p(\hat{X}_{n-1}, Z_{n-1}, i, j))/\Gamma_{n-1}] = 0,$$

whereas we need to prove

$$E[(I\{\bar{X}_n = i, \bar{Y}_n = j\} - p(\bar{X}_{n-1}, Z_{n-1}, i, j))/\Gamma_{n-1}] = 0.$$

But the left-hand side equals

$$\begin{aligned}
&E[(I\{\tilde{X}_n = i, \tilde{Y}_n = j\} - p(\tilde{X}_{n-1}, Z_{n-1}, i, j))I\{\tau \geq n\} \\
&\quad + (I\{\hat{X}_n = i, \hat{Y}_n = j\} - p(\hat{X}_{n-1}, Z_{n-1}, i, j))I\{\tau < n\}/\Gamma_{n-1}] \\
= \quad &E[(I\{\tilde{X}_n = i, \tilde{Y}_n = j\} - p(\tilde{X}_{n-1}, Z_{n-1}, i, j)/\Gamma_{n-1}]I\{\tau \geq n\} \\
&\quad + E[(I\{\hat{X}_n = i, \hat{Y}_n = j\} - p(\hat{X}_{n-1}, Z_{n-1}, i, j))/\Gamma_{n-1}]I\{\tau < n\} \\
= 0,&
\end{aligned}$$

proving the claim. The wide sense admissibility of $\{Z_n\}$ for $(\{\bar{X}_n\}, \{\bar{Y}_n\})$ can be verified easily by reference to the probability measure $\hat{P}_0$ above. Under $\hat{P}_0$, $\tilde{Y}_n$ and $\hat{Y}_n$ are both independent of $\Gamma_{n-1} \bigvee \sigma(\tilde{\xi}_m, \hat{\xi}_m, m \geq 0)$ and are identically distributed, therefore so will be $\theta\tilde{Y}_n + (1-\theta)\hat{Y}_n$ for any Γ_{n-1}-measurable $\{0, 1\}$-valued random variable θ. $\bar{Y}_n$ is a special case of this. Now let $\{\bar{\pi}_n\}$ denote the corresponding process of conditional laws. Now,

$$\begin{aligned}
|V_\alpha(\mu_1) - V_\alpha(\mu_2)| &\leq \sup_{\{Z_n\}\in\Phi} |J_\alpha(\{Z_n\}, \mu_1) - J_\alpha(\{Z_n\}, \mu_2)| \\
&\leq \sup_{\Phi} \sum_{m=0}^{\infty} \alpha^m E[|h(\bar{X}_m, Z_m) - h(\tilde{X}_m, Z_m)|] \\
&= \sup_{\Phi} \sum_{m=0}^{\infty} \alpha^m E[|h(\bar{X}_m, Z_m) - h(\tilde{X}_m, Z_m)|I\{\tau > m\}] \\
&\leq 2K_2 \sup_{\Phi} \sum_{m=0}^{\infty} \alpha^m P(\tau > m) \\
&\leq 2K_2 K_0 \sum_{m=0}^{\infty} \alpha^m \delta^m E[|\bar{X}_0 - \tilde{X}_0|],
\end{aligned}$$

where K_2 is a bound on $|h(\cdot,\cdot)|$. Let $\epsilon > 0$. By a judicious choice of the joint law of $(\bar{X}_0, \tilde{X}_0)$, this can be made

$$\begin{aligned}
&\leq \frac{2K_2K_0}{1-\alpha\delta}(\rho(\mu_1, \mu_2) + \epsilon) \\
&\leq \frac{2K_2K_0}{1-\delta}(\rho(\mu_1, \mu_2) + \epsilon).
\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, the claim follows. □

4. The dynamic programming equations. Fix a $\mu^* \in \mathcal{P}(S)$ and set $\bar{V}_\alpha(\mu) = V_\alpha(\mu) - V_\alpha(\mu^*)$ for $\mu \in \mathcal{P}(S), \alpha \in (0,1)$. By Lemma 3.2, $\{\bar{V}_\alpha(\cdot), \alpha \in (0,1)\}$ is bounded equicontinuous. Letting $\alpha \to 1$, we conclude from the Arzela–Ascoli theorem that $\bar{V}_\alpha(\cdot)$ converges in $C(\mathcal{P}(S))$ to some $V(\cdot)$ along a subsequence $\{\alpha(n)\}, \alpha(n) \to 1$. By dropping to a further subsequence if necessary, we may also suppose that $\{(1-\alpha(n))V_{\alpha(n)}(\mu^*)\}$, which is clearly bounded, converges to some $\Delta \in R$ as $n \to \infty$.

Before stating our main theorem, recall that for a separated control problem, a control policy is said to be stationary if $Z_n = v(\pi_n)\ \forall n$ and some measurable $v : \mathcal{P}(S) \to A$, and stationary randomized if for all n, Z_n is conditionally independent of $X_0,\ \xi_m,\ \pi_m,\ Z_{m-1},\ Y_m,\ m \le n$, given π_n and the regular conditional law thereof given the latter is $\varphi(\pi_n)$ for some measurable $\varphi : \mathcal{P}(S) \to \mathcal{P}(A)$ ([3, Chapter VIII]). It is also easy to verify that these are wide sense admissible. By abuse of terminology, we refer to the map $v(\cdot)$ (resp., $\varphi(\cdot)$) itself as the stationary (resp., stationary randomized) policy.

THEOREM 4.1. (i) $(V(\cdot), \Delta)$ *solve the dynamic programming equation for the average cost control problem*

$$V(\pi) = \min_u \left(\bar{h}(\pi, \mu) + \int \eta(\pi, u, d\pi')V(\pi') - \Delta \right). \tag{4.1}$$

(ii) Δ *is the optimal cost, independent of the initial condition. Furthermore, a stationary policy* $v(\cdot)$ *(resp., a stationary randomized policy* $\varphi(\cdot)$*) is optimal for any initial condition if*

$$v(\pi) \in \text{(resp., support}(\varphi(\pi)) \subset) \text{ Argmin} \left(\bar{h}(\pi, \cdot) + \int \eta(\pi, \cdot, d\pi')V(\pi') \right).$$

In particular, an optimal stationary policy exists.

Proof. (i) Rewrite (3.1) as

$$\bar{V}_\alpha(\pi) = \min_u \left(\bar{h}(\pi, u) + \alpha \int \eta(\pi, u, d\pi')\bar{V}_\alpha(\pi') - (1-\alpha)V_\alpha(\mu^*) \right).$$

Letting $\alpha \to 1$ along an appropriate subsequence, we get (4.1).

(ii) The first two statements follow by a standard argument which may be found, e.g., in [8, Theorem 5.2.4, pp. 80–81]. The last claim follows from a standard measurable selection theorem—see, e.g., [12]. □

THEOREM 4.2. *If* φ *is an optimal stationary policy and* γ *is a corresponding ergodic probability measure, then, for* $\varphi(\pi, du) \stackrel{\Delta}{=} \varphi(\pi)(du)$,

$$V(\pi) = \int \bar{h}(\pi, u)\varphi(\pi, du) + \int\int \varphi(\pi, du)\eta(\pi, u, d\pi')V(\pi') - \Delta \tag{4.2}$$

for γ*—a.s.* π*. (In particular, if* v *is an optimal stationary policy, then* (4.2) *holds for* γ*—a.s.* π *for every ergodic probability measure* γ *under* v*).*

Proof. Clearly, (4.3) always holds with "$\le$" replacing "=." If the claim were false, we can integrate both sides of this inequality with respect to $\gamma(d\pi)$ to obtain

$$\Delta < \int\int \bar{h}(\pi, \cdot)d\varphi(\pi)\gamma(d\pi),$$

a contradiction to the optimality of φ. The claim follows. □

COROLLARY 4.3. *If V, V' are two bounded measurable solutions to* (4.1), $V(\pi) - V'(\pi)$ *is a constant a.s. with respect to any ergodic probability measure under any optimal stationary randomized policy.*

Proof. Let φ, γ be as in Theorem 4.2, and let $\{\pi_n\}$ be the corresponding ergodic process of conditional laws, with law of $\pi_n = \gamma \ \forall n$. By Theorem 4.2,

$$V(\pi_n) - V'(\pi_n),\ n \geq 0,$$

is a bounded martingale with respect to the natural filtration of $\{\pi_n\}$ and therefore converges a.s. But since it is also a stationary process, this is possible only if the claim holds. □

Appendix. Here is a simple sufficient condition for (3.2) to hold. Suppose that there exist $i_0 \in S,\ \eta > 0$, and $N \geq 1$ such that

$$\inf_{i,j} \inf_{\Phi} P(\hat{X}_N = \tilde{X}_N = i_0 / \hat{X}_0 = i, \tilde{X}_0 = j) \geq \eta.$$

That is, there is a path of length N from any $i \in S$ to i_0 with a minimum probability of η. (For uncontrolled chains, aperiodicity would ensure this condition for all N sufficiently large. Thus we need aperiodicity that is in some sense uniform in Φ.) Now,

$$P(\tau > N / \hat{X}_0 = i, \tilde{X}_0 = j) \leq 1 - \eta.$$

Also, by a simple conditioning argument, for $m \geq 1$,

$$\begin{aligned} P(\tau > mN/\hat{X}_0 = i, \tilde{X}_0 = j) &= E[P(\tau > mN/\Gamma_{(m-1)N}) I\{\tau > (m-1)N\}/\hat{X}_0 = i, \tilde{X}_0 = j] \\ &\leq (1-\eta) P(\tau > (m-1)N/\hat{X}_0 = i, \tilde{X}_0 = j). \end{aligned}$$

Iterating, $P(\tau > mN/\hat{X}_0 = i, \tilde{X}_0 = j) \leq (1-\eta)^m,\ m \geq 1$, from which (3.2) follows.

REFERENCES

[1] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.

[2] V.S. Borkar, *White-noise representations in stochastic realization theory*, SIAM J. Control Optim., 31 (1993), pp. 1093–1102.

[3] V.S. Borkar, *Topics in Controlled Markov Chains*, Pitman Res. Notes Math. 240, Longman Scientific and Technical, Harlow, UK, 1991.

[4] V.S. Borkar, *Probability Theory: An Advanced Course*, Springer-Verlag, New York, 1995.

[5] E. Fernández-Gaucherand, A. Arapostathis, and S.I. Marcus, *On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes*, Ann. Oper. Res., 29 (1991), pp. 439–470.

[6] E. Fernández-Gaucherand, A. Arapostathis, and S.I. Marcus, *Remarks on the existence of solutions to the average cost optimality equation in Markov decision processes*, Systems Control Lett., 15 (1990), pp. 425–432.

[7] W.H. Fleming and E. Pardoux, *Optimal control of partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.

[8] O. Hernández-Lerma and J.B. Lasserre, *Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1996.

[9] L.K. Platzman, *Optimal infinite-horizon undiscounted control of finite probabilistic systems*, SIAM J. Control Optim., 18 (1980), pp. 362–380.

[10] W.J. Runggaldier and L. Stettner, *Approximations of Discrete Time Partially Observed Control Problems*, Applied Maths. Monographs 6, Giardini Editori e Stampatori, Pisa, 1994.

[11] L. Stettner, *Ergodic control of partially observed Markov processes with equivalent transition probabilities*, Appl. Math. (Warsaw) 22 (1993), pp. 25–38.

[12] D.H. Wagner, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859–903.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 682–706

GLOBAL STABILIZATION OF NONLINEAR SYSTEMS WITH INPUTS SUBJECT TO MAGNITUDE AND RATE BOUNDS: A PARAMETRIC OPTIMIZATION APPROACH*

JULIO SOLÍS-DAUN†, RODOLFO SUÁREZ†, AND JOSÉ ÁLVAREZ-RAMÍREZ†

Abstract. A bounded feedback control design approach is proposed for the global asymptotic stabilization of a class of nonlinear systems with stable free dynamics. The control inputs and their derivatives are constrained to take values on sets defined by a Cartesian product of η-dimensional closed balls $\mathcal{B}_{\mathbf{r}}^{\eta}(p)$, which are defined by means of a p-norm and a radius vector parameter $\mathbf{r}$. In order to derive the bounded control stabilizer, the resulting procedure implies that gains (as state-functions) are obtained from the solution to a set of c-parameterized nonlinear programming problems. In general, the resulting closed-loop system could be *implicitly defined*, i.e., consisting of a system of differential equations plus a set of nonlinear algebraic equations (required to compute the control). Special interest is focused on an important class of homogeneous systems that includes a class of globally asymptotically stabilizable systems by linear feedback and bilinear systems. For those systems, the problem of inputs subject to *globally bounded rates* is also addressed.

Key words. bounded control, bounded control rate, nonlinear systems, global stabilization, parametric programming

AMS subject classifications. 93D15, 93C10, 90C31

PII. S0363012997330233

1. Introduction.

Consider the multiple input continuous-time affine system

$$\dot{x} = f(x) + \sum_{j=1}^{m} u_j\, g_j(x), \tag{1.1}$$

where $x \in \mathbb{R}^n$ and $f, g_j : \mathbb{R}^n \to \mathbb{R}^n$, for $j = 1, \ldots, m$, are smooth functions. Without loss of generality, we shall assume that the origin is an equilibrium point of the associated free dynamics of (1.1), i.e., $f(0) = 0$.

Define the η-dimensional $p, \mathbf{r}$-normed closed ball by

$$\mathcal{B}_{\mathbf{r}}^{\eta}(p) := \left\{ v \in \mathbb{R}^{\eta} : \|v\|_{p,\mathbf{r}} \le 1 \right\}, \text{ where } \|v\|_{p,\mathbf{r}} := \left[\left(\tfrac{v_1}{r_1}\right)^p + \cdots + \left(\tfrac{v_\eta}{r_\eta}\right)^p \right]^{1/p}, \tag{1.2}$$

for $\mathbf{r}$ a radius vector parameter $\mathbf{r} := (r_1, \ldots, r_\eta)^\top, r_i > 0$, for $i = 1, \ldots, \eta$, and $p := s/d > 1$, with s even and d odd positive numbers. We will say that $\|v\|_{p,\mathbf{r}}$ is the $\mathbf{r}$-weighed p-norm. Observe that the *usual* p-norm $\|\cdot\|_p$ is simply the $p, \mathbf{r}$-norm when $\mathbf{r} = (1, \ldots, 1)$. We assume that, by renaming the control input entries (if necessary), the control input is given by $u = (\mathbf{u}_1, \ldots, \mathbf{u}_\mu)^\top$ with *control blocks* $\mathbf{u}_i$ taking values in different $p, \mathbf{r}$-normed balls. Thus, the *control-value set* consists on a Cartesian product of $p, \mathbf{r}$-normed closed balls,

$$U = \prod_{i=1}^{\mu} \mathcal{B}_{\mathbf{r}_i}^{m_i}(p_i) := \mathcal{B}_{\mathbf{r}_1}^{m_1}(p_1) \times \cdots \times \mathcal{B}_{\mathbf{r}_\mu}^{m_\mu}(p_\mu) \tag{1.3}$$

*Received by the editors November 17, 1998; accepted for publication (in revised form) May 8, 2000; published electronically October 4, 2000. This work was partially supported by CONACyT grant 400200-5-C036E.

http://www.siam.org/journals/sicon/39-3/33023.html.

†División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana-Iztapalapa, Apdo. Postal 55-534, 09340, México, D.F., México (jesd@xanum.uam.mx, rsua@xanum.uam.mx, jjar@xanum.uam.mx).

for $1 \leq \mu \leq m$ and $m_1 + \cdots + m_\mu = m$. We have, for instance, the following examples: (a) if $p = 2$, then $\mathcal{B}^m_{\mathbf{r}}(2)$ is an ellipsoid in $\mathbb{R}^m$ with r_j the length of its jth-semiaxe; (b) if $r_1 = \cdots = r_m = \overline{r} > 0$, then $\mathcal{B}^m_{\overline{\mathbf{r}}}(p) = \{v \in \mathbb{R}^m : \|v\|_p \leq \overline{r}\}$ is a p-normed (Minkowski) ball of radius $\overline{r}$ in $\mathbb{R}^m$; (c) if $m = m_1 + m_2 = 3$, then $\mathcal{B}^1_{r_1}(\infty) \times \mathcal{B}^2_{\overline{\mathbf{r}}_2}(p)$ is a "cylinder" with its base being a p-normed ball of radius $\overline{r} > 0$; and (d) if $\mu = m$, then (1.3) is an m-dimensional $\mathbf{r}$-hyperbox $\mathcal{B}^m_{\mathbf{r}}(\infty) = [-r_1, r_1] \times \cdots \times [-r_m, r_m]$.

By an *admissible* input u, it is meant a C^α_L *function* (i.e., a C^α ($\alpha \geq 1$) function with the possible exception of $x = 0$, and everywhere Lipschitz continuous) that takes values in a (*prescribed*) Cartesian product of $p, \mathbf{r}$-normed closed balls, that is, the set of admissible controls is given by

$$(1.4) \qquad \mathcal{U} = C^\alpha_L(\mathbb{R}^n, U) := \left\{ u : \mathbb{R}^n \to \prod_{i=1}^{\mu} \mathcal{B}^{m_i}_{\mathbf{r}_i}(p_i) : u(\cdot) \text{ is a } C^\alpha_L \text{ function} \right\}.$$

In this paper, we first address the problem of *global asymptotic stabilization* (GAS) of the system (1.1) by means of a *bounded feedback control* (BFC) function $u = u(x)$, admissible in the aforementioned sense. As a second aim, we will study the GAS problem when further subject to input derivatives restricted to lie in *prescribed* constraint sets (1.3). Throughout this paper, stabilization will always be understood as stabilization at the origin.

In the last years, there has been a growing interest concerning the GAS problem of systems by means of BFC functions. The characterization of a class of linear systems for which it can be solved led to the concept of *asymptotic null controllability with bounded controls* (ANCBC). It is well known that a linear system is ANCBC iff it is *stabilizable by arbitrarily small controllers* (SASC) (cf. Sontag [24] and references therein). Different methodologies have been proposed for GAS of linear systems via BFC functions. See, for instance, Gavrilyako, Korobov, and Sklyar [10], Sussmann, Sontag, and Yang [27] and Suárez, Álvarez-Ramírez, and Solís-Daun [26]. When further assuming the problem of *rate-limited actuators*, see Lin [16] and Solís-Daun, Álvarez-Ramírez, and Suárez [23] for *semiglobal* results, and Shewchun and Feron [22] for a *global* solution. In contrast to the case of linear control systems, the characterization of the class of nonlinear systems which have the SASC property remains as an open problem.

Lyapunov analysis is a tool usually employed for the stabilization of nonlinear systems. Indeed, many systems admit a variational approach for deriving their dynamical equations and, in that framework, the energy function provides typically a positive-definite and proper function, such that it is nonincreasing along the solution of the uncontrolled system. This idea was the subject of many works on feedback stabilization, see, for instance, [11, 13, 4, 14] and references therein. For that aim, Jurdjevic and Quinn (J–Q) [13] developed a successful approach, based on a controllability-like rank condition defined in terms of Lie derivatives: the *ad-condition.* In [4], Byrnes, Isidori and Willems introduced the concept of *zero state detectability* (ZSD) in order to address the GAS problem for the important class of *passive* (and *feedback passive*) nonlinear systems by means of smooth (generally unbounded) feedback control functions. Further, they presented a criterion (an *ad*-like condition) for testing ZSD. Recently, Coron [5] and Lin [14] obtained GAS of passive systems by using arbitrarily small stabilizers. Those results can be thought of as a first attempt towards the characterization of the class of nonlinear systems sharing the SASC property. Lin also provided a criterion for testing ZSD that generalized previously reported results.

Different methods of the J–Q approach for the stabilization of nonlinear systems by means of bounded controls have been proposed. Mazenc and Praly [20] have given sufficient conditions under which GAS of a certain subsystem by saturated control implies GAS of the overall system; further, their method served to solve the stabilization problem of feedforward systems. Teel [28] presented a small gain theorem that was used for analyzing control system design: e.g., a stabilizing algorithm for feedforward systems was obtained, which in turn was applied to the control of stabilizable linear systems with input magnitude and rate saturation. Freeman and Praly [6] presented a backstepping procedure for the design of global feedback stabilizers which are bounded both in magnitude and rate (though the achieved bounds do not satisfy prescribed constraints). Finally, Sontag and coworkers have proposed *explicit formulae* for almost smooth BFC laws that stabilize a general system (1.1), under the assumption that an appropriate *control-Lyapunov function* (CLF) (see [1]) is known. Such formulae were designed to guarantee CLF stabilization under specific control-value sets U: the Euclidean open unit ball in [15] and a family of Minkowski open unit balls (with $1 < p \leq 2$) in [19].

On the other hand, it is important to note that depending on some applications, e.g., robot manipulators, chemical control processes, etc., unlimited feedback control rate can be conceived as a disadvantage of most of the existing methods for designing bounded stabilizers. In the case of polynomial systems, the designed control often resembles a *bang-bang control*, especially when states are far from the origin. That behavior precludes their use in the mentioned applications, due to the natural responses (e.g., controller inertia) of those systems to external stimuli. Hitherto, a general methodology for the GAS problem of nonlinear systems subject to prescribed input magnitude and rate bounds is still lacking.

Taking into account the above problems, in this paper we propose a design approach that allows the GAS of a class of nonlinear systems with stable free dynamics, using BFC functions with bounded derivatives. Following the basic ideas introduced in [26], the (high-gain) control law proposed in this work increases the feedback "gain" as the controlled trajectory converges towards the origin, guaranteeing that input bounds will not be exceeded. In the general nonlinear systems case with control-value set given by a $p, \mathbf{r}$-normed ball, the proposed procedure implies that: (a) in order to derive the bounded stabilizer, the "gains" (as state-functions) are obtained from the solution of a *c-parameterized nonlinear programming* problem; and (b) the resulting closed-loop system could be *implicitly defined*, in the sense that it consists of a system of differential equations plus a nonlinear algebraic equation (required to compute the control). With respect to existing results available in the literature [5, 6, 14, 20, 28], in this work we make the following contributions.

(a) We extend Lyapunov-based designs to account for more general control-value sets.
(b) We address the problem of both magnitude and rate constraints in the control input.
(c) We introduce a state-dependent control gain to achieve a suboptimal stability margin.

The paper is organized as follows. In section 2 we present some definitions, review some previous work, and state the problem. In section 3, a stabilizing BFC function is designed in the case when the control-value set is given by an m-dimensional $p, \mathbf{r}$-normed ball. In section 4, we show that the proposed control is suboptimal. In section 5, the result is extended to the case when the control-value set is a Cartesian product of $p, \mathbf{r}$-normed balls. In section 6, we study an important class of homogeneous systems (that includes those GAS systems via linear feedback and bilinear systems), for which the constructed control law is *explicitly defined*, so that the problem of inputs subject to prescribed *globally bounded rates* can be addressed. Finally, in section 7, we present some concluding remarks.

2. Preliminaries and statement of the problem. Recall that a function $V : \mathbb{R}^n \to \mathbb{R}$ is said to be *proper* iff, for any $c \in \mathbb{R}$, the set $V^{-1}(c) = \{x \in \mathbb{R}^n : V(x) = c\}$ is compact, and it is *positive-definite* iff $V(0) = 0$ and $V(x) > 0$ for all $x \neq 0$. In general, $L_{g_j}V(x)$ denotes a Lie derivative of function $V(x)$ in the direction of function $g_j(x)$. If g denotes the matrix $g(x) = (g_1(x), \dots, g_m(x))$, $g_j : \mathbb{R}^n \to \mathbb{R}^n$, $j = 1, \dots, m$, then we define $L_gV(x) := (L_{g_1}V(x), \dots, L_{g_m}V(x))$.

In this work, we assume the following.

Hypothesis H1. *Suppose there exists a C^α ($\alpha \geq 2$) function $V : \mathbb{R}^n \to \mathbb{R}$ that is positive-definite and proper on $\mathbb{R}^n$, such that the uncontrolled system $\dot{x} = f(x)$ ((1.1) with $u = 0$) satisfies*

$$L_fV(x) \leq 0. \tag{2.1}$$

In particular, it is Lyapunov stable.

Choose a dummy output $y := (L_gV(x))^\top$ for the system (1.1). Then, the input-output system (1.1) with dummy output

$$h(x) := (L_gV(x))^\top \tag{2.2}$$

is *passive* [4, 14] with *storage function* $V(x)$.

In addition to Hypothesis H1, system (1.1) is assumed to satisfy the following.

Hypothesis H2. *The input-output system* (1.1)–(2.2) *is zero state detectable (ZSD). That is, for all $x \in \mathbb{R}^n$,*

$$\textit{if } \forall t \geq 0,\ h\left(x(t, x_0; 0)\right)|_{u=0} = 0, \quad \textit{then } \lim_{t\to\infty} x(t, x_0; 0) = 0, \tag{2.3}$$

where $x(t, x_0; u)$ is a trajectory of (1.1) *with initial condition $x(0) = x_0$.*

We will need the following simple extension of a result presented in [4].

Proposition 2.1. *Assume that system* (1.1) *satisfies Hypotheses* H1–H2. *Then, any C_L^α feedback control function*

$$u(x) = \rho\, v(x), \tag{2.4}$$

with $\rho > 0$, satisfying that

$$\begin{array}{l}\text{(i) } v(0) = 0, \quad \text{(ii) } L_gV(x)\, v(x) \leq 0, \ \textit{and} \\ \text{(iii) } L_gV(x)\, v(x) = 0 \ \textit{only if} \ L_gV(x) = 0\end{array} \tag{2.5}$$

achieves GAS of the system (1.1). *In particular, any control $v(x)$ such that $v(0) = 0$ and* $\operatorname{sign} v_j(x) = -\operatorname{sign} h_j(x)$ *for $j = 1, \dots, m$ satisfies conditions* (2.5).

Hitherto, the control given in (2.4) is not necessarily bounded. Nevertheless, in the case of passive systems, Coron and Lin, independently, provided the following SASC result.

Theorem 2.2 (see [5, 14]). *Assume that Hypotheses* H1–H2 *hold. Then GAS of system* (1.1) *is achieved by means of arbitrarily small smooth feedbacks.*

Although one can use a *smooth boundedness* technique to obtain a globally defined BFC function that globally asymptotically stabilizes the system, low-gain control designing should usually be avoided, because it can result in sluggish responses and poor performance for all initial conditions. One way to obtain a high-gain controller is to define the parameter $\rho = \rho(x)$ in (2.4) as a positive function that increases as the distance to the origin is reduced, in such a way that the controller uses the maximum input amplitude without violating the input constraint. Indeed, it is not difficult to design such a function $\rho(x)$. For instance, consider the scalar feedback control given by

$$u_\sigma(x) := \begin{cases} 0 & \text{if } h(x) = 0, \\ -\rho(x)\, h(x) = -\frac{r\,\sigma(|h(x)|)}{|h(x)|}\, h(x) & \text{otherwise,} \end{cases} \tag{2.6}$$

where $\sigma(s) \leq 1$ for all $s \geq 0$, $\sigma(s) > 0$ for $s > 0$, and σ is as smooth as desired. Clearly, this feedback control is bounded: $|u_\sigma(x)| \leq r$ for all $x \in \mathbb{R}^n$, and it can be considered as a smooth approximation to the singular bang-bang control $\vartheta(x) = -r\,\mathrm{sign}(h(x))$. However, the problem is if the obtained controller has a globally bounded input rate. Indeed, control (2.6) does not have a bounded derivative, even in the linear case. In fact, suppose that the linear system with scalar control input $\dot{x}= Ax + b\,u$, (A is an $n \times n$ matrix and $b \in \mathbb{R}^n$) satisfies Hypotheses H1–H2. In this case, $h(x) = b^\top Px$, where P is a positive-definite symmetric matrix. Consider only the case when $h(x) > 0$, so that $u_\sigma(x) = -r\,\sigma(h(x)) < 0$. The analysis of the case $h(x) < 0$ is analogous. Thus,

$$\dot{u}_\sigma= -r\ \dot{\sigma}= -r\,\nabla\sigma\ \dot{h}= -r\,\nabla\sigma(h(x))\ \left(b^\top PAx + (b^\top Pb)\,u_\sigma(x)\right). \tag{2.7}$$

Let $U(\overline{h}) := \{x \in \mathbb{R}^n : h(x) = b^\top Px = \overline{h}$, with fixed $\overline{h} > 0\}$. Then,

$$\dot{u}_\sigma|_{U(\overline{h})}= -r\,\nabla\sigma(\overline{h})\ \left(b^\top PAx + (b^\top Pb)\,\overline{u}_\sigma\right)|_{U(\overline{h})}, \tag{2.8}$$

where $u_\sigma\ |_{h=\overline{h}}= \overline{u}_\sigma$, is an *unbounded function* (unless $A = \lambda\, I_{n\times n}$, with $\lambda \leq 0$ and $I_{n\times n}$ is the $n\times n$ identity matrix). Observe that for all $x \in U(\overline{h})$, we have that $\rho(x) = \overline{\rho} = r\,\sigma(\overline{h})/\overline{h} > 0$. Then, in virtue that the set $U(\overline{h})$ ($\subseteq L(\overline{\rho}) = \{x \in \mathbb{R}^n : \rho(x) = \overline{\rho}$, $\overline{\rho} > 0\}$) is unbounded, we have that $|du_\sigma/dt|$ cannot be globally bounded.

The difference between control functions like (2.6) and the control function to be proposed in the next section is that $\rho(x)$ will be constant along the boundary of the level sets of certain proper function $E(x)$, i.e., $\rho(x)$ will be constant on compact sets. Based on this idea, we obtain control functions with derivatives restricted to lie in *prescribed* constraint sets.

3. Control design when $U = \mathcal{B}^m_{\mathrm{r}}(p)$. In general, $V(x)$ (the Lyapunov function considered in Hypothesis H1) does not necessarily satisfy the condition that its level sets are simply connected—a sufficient condition in the control function design presented in this section. One case where that condition is satisfied, is the important class of homogeneous systems studied in section 6. Thus, we assume the following definition.

DEFINITION 3.1. *Suppose that $E : \mathbb{R}^n \to \mathbb{R}$ is a smooth, proper, and positive-definite function satisfying that, for any $c > 0$, its c-level sets*

$$\mathcal{E}(c) := \{x \in \mathbb{R}^n : E(x) \leq c\} \tag{3.1}$$

are simply connected, with $0 \in \operatorname{int} \mathcal{E}(c)$*. Then, we will say that* E *is an* $\mathcal{E}$*-function.*

Suppose that the control-value set U is an m-dimensional $p, \mathbf{r}$-normed closed ball, $\mathcal{B}^m_{\mathbf{r}}(p)$, with $\mathbf{r}$ a radius vector parameter $\mathbf{r} := (r_1, \ldots, r_m)^\top$, and $p \in \mathbb{Q}^*_{>1}$, with

$$\mathbb{Q}^*_{>1} := \left\{ p \in \mathbb{Q} : p = \frac{s}{d} > 1, \text{ with } s, d \in \mathbb{N}\backslash\{0\}, \text{ even and odd numbers} \right\}, \tag{3.2}$$

which is a *dense subset* of the open interval $(1, \infty) \subset \mathbb{R}$.

In this work, the proposed control function shall depend on the *dual value* $q > 1$, related to the given $p, \mathbf{r}$-norm, defined by

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{3.3}$$

From this formula, we have that $p \in \mathbb{Q}^*_{>1}$ iff $q \in \mathbb{Q}^*_{>1}$. Consider the associated *dual* $q, 1/\mathbf{r}$*-norm* given by

$$\|v\|_{q,1/\mathbf{r}} := \left[\sum_{j=1}^{m} (r_j \, v_j)^q \right]^{1/q}. \tag{3.4}$$

The $p, \mathbf{r}$ and $q, 1/\mathbf{r}$ norms have the following properties. (1) Let $R = \operatorname{diag}(r_1, \ldots, r_m)$; then $\|Rv\|_q = \|v\|_{q,1/\mathbf{r}}$ and $\|R^{-1}v\|_p = \|v\|_{p,\mathbf{r}}$. (2) *Hölder inequality*: for all $u, w \in \mathbb{R}^m$, $|u^\top w| \leq \|u\|_{q,1/\mathbf{r}} \, \|w\|_{p,\mathbf{r}}$.

In what follows, we shall propose a control design methodology for GAS of system (1.1) with smooth BFC functions. In addition, this technique will allow us to address the problem of globally bounded input rates. To this aim, we define a smooth function $\tau(x)$ that superestimates $\|v(x)\|_{q,1/\mathbf{r}}$, and it is constant along the boundary of the c-level sets of the proper function $E(x)$.

Choose an arbitrary $0 \neq x_0 \in \mathbb{R}^n$ and set $c = E(x_0)$ (> 0). Therefore, we pose the optimization problem

$$\begin{aligned} \tau(x_0) = \ & \max\nolimits_x \|v(x)\|_{q,1/\mathbf{r}} \text{ subject to (s.t.)} \\ & x \in \partial\mathcal{E}(c), \quad \text{with } c = E(x_0), \end{aligned} \tag{3.5}$$

where $v(x)$ was defined in Proposition 2.1 and ∂ denotes the boundary of a given set. Optimization problem (3.5) is a one-parameter $(c \geq 0)$ family of programs with a varying equality constraint [7, 12] such that its objective function $\|v(x)\|_{q,1/\mathbf{r}}$ does not depend on c. In particular, since $E(x)$ is proper, the family of programs given above is proper [7]. A proper program (as (3.5)) has at least one global solution, since the set $\partial\mathcal{E}(c)$ is compact. In order to exclude the case that $\partial\mathcal{E}(c)$ be a discrete set for all $c > 0$, we should restrict the problem by taking $n \geq 2$ (i.e., one-dimensional systems will not be considered).

Remark 1. Note that $E(x)$ must be an $\mathcal{E}$-function because, otherwise, if for some $c > 0$ the set $\mathcal{E}(c)$ (3.1) would not be *simply* (not even) *connected*, then it would lead to hard troubles for implementation of program (3.5).

A solution $\tau = \tau(x)$ to the optimization problem (3.5) will be called *admissible* iff it is positive-definite. Hereafter, unless otherwise specified, $\tau(x)$ will be assumed admissible. Then, for $q \geq 2$ (more specifically, $q \in \mathbb{Q}^*_{>1} \cap [2, \infty) \subset \mathbb{R}$), we define the τ-dependent control function as

$$u_\tau(x) := \frac{1}{\tau^{q-1}(x)} \left(r_1 \, (r_1 v_1(x))^{q-1}, \ldots, r_m \, (r_m v_m(x))^{q-1} \right)^\top. \tag{3.6}$$

On the other hand, from (3.3) and the definition of p, it follows that $q := s/(s-d)$, so that $q-1 = d/(s-d)$, with both d and $s-d > 0$ being odd numbers. Thus, the exponent $q-1$ in (3.6) preserves the sign of each control component v_j for $j = 1, \dots, m$; and hence, based on Proposition 2.1, global stabilization using that control is guaranteed provided that $v(x)$ is a global stabilizer. Nevertheless, observe that for $p \geq 2$ we have $q-1 < 1$, and vice versa. Thus, for $q \in \mathbb{Q}^*_{>1} \cap (1,2)$, the proposed control given by (3.6) could be nonsmooth (even non-Lipschitz) whenever $v_j(x) = 0$ for some j, because all terms in (3.6) involve potential functions with exponent $0 < \underline{q-1} < 1$. To overcome this difficulty, one can apply the following smooth approximation ("regularization") to odd root-like functions.

First of all, denote $z_j = \xi_j^{q-1}$, where $\xi_j = r_j v_j/\tau$, for $j = 1, \dots, m$, so that control (3.6) becomes

$$u_\tau(x) = \left(r_1\, z_1(x), \dots, r_m\, z_m(x)\right)^\top . \tag{3.7}$$

Then, replace each function z_j with a new function $\widehat{z}_j$ being implicitly defined from the relation: $\xi_j = (2-q)\, \widehat{z}_j^{1/(q-1)} + (q-1)\, \widehat{z}_j = ((p-2)\, \widehat{z}_j^{p-1} + \widehat{z}_j)/(p-1)$, for $j = 1, \dots, m$. Function $\widehat{z}_j(\xi_j)$ subestimates $z_j(\xi_j)$ (i.e., $|\widehat{z}_j| \leq |z_j|$ for all $|\xi_j| \leq 1$) and $\widehat{z}_j = z_j$ iff $\xi_j = -1, 0$ or 1. Further, $\widehat{z}_j(\xi_j)$ is an everywhere differentiable and bijective function in the variable ξ_j

$$\widehat{z}'_j(\xi_j) = \frac{p-1}{\Big((p-1)(p-2)\, \widehat{z}_j^{p-2}(\xi_j) + 1\Big)} > 0 \quad \forall\, |\xi_j| \leq 1. \tag{3.8}$$

Observe that $\widehat{z}'_j(0) = p-1$, so that the slope at $\xi_j = 0$ increases as $p \to \infty$. Therefore, the proposed control function when $q \in \mathbb{Q}^*_{>1} \cap (1,2)$, is the following:

$$u_\tau(x) := \left(r_1\, \widehat{z}_1(x), \dots, r_m\, \widehat{z}_m(x)\right)^\top . \tag{3.9}$$

It should be pointed out that both controllers (3.6) and (3.9) are not defined at the origin. Hence, the origin is rather a *singularity* instead of an equilibrium point of the corresponding closed-loop system. That means that the resulting system can be non-Lipschitz at the origin (uniqueness of the solutions with respect to initial conditions is not guaranteed), so that all trajectories could converge to $x = 0$ in *finite-time.* In view that for a wide range of applications, this is an undesirable feature (e.g., causing the so-called *chattering* of the controller), those controllers are redesigned by introducing a tuning parameter $\varepsilon > 0$, which can be taken as small as desired. Denote

$$v_\varepsilon(x) := \frac{1}{\varepsilon + \tau(x)}\, \left(r_1\, v_1(x), \dots, r_m\, v_m(x)\right)^\top = \frac{1}{\varepsilon + \tau(x)}\, R\, v(x), \tag{3.10}$$

where $R = \operatorname{diag}(r_1, \dots, r_m)$. Consequently, the proposed BFC function is defined as

$$u_\theta(x) := \left(r_1\, \zeta_1\left(v_{\varepsilon_1}\right), \dots, r_m\, \zeta_m\left(v_{\varepsilon_m}\right)\right)^\top , \tag{3.11}$$

where the functions ζ_j , for $j = 1, \dots, m$, are defined

$$\begin{cases} \text{explicitly}: \quad \zeta_j = (\theta v_{\varepsilon_j})^{1/(p-1)} & \text{if } p \in \mathbb{Q}^*_{>1} \cap (1,2], \\ \text{implicitly}: \quad (p-2)\zeta_j^{p-1} + \zeta_j = (p-1)\, \theta v_{\varepsilon_j} & \text{if } p \in \mathbb{Q}^*_{>1} \cap [2,\infty), \end{cases} \tag{3.12}$$

with θ a parameter close to (and less than) 1, $0 < \theta \approx 1$ (needed for proving Proposition 3.3). In the implicit case, we have that $\zeta_j'(0) = p - 1$, so that the value of the slope of functions $\zeta_j(v_{\varepsilon_j})$ at $v_{\varepsilon_j} = 0$ increases as p does. Observe that if $p = 2$, both explicit and implicit functions ζ_j coincide, so that $\zeta_j = v_{\varepsilon_j}$.

Control (3.11) fulfills the following important items. (1) It is bounded in the $p, \mathbf{r}$-normed ball $\mathcal{B}_{\mathbf{r}}^m(p)$, and (2) it achieves GAS of the system (1.1) under the hypotheses of Proposition 2.1. These facts are proved in the following proposition.

PROPOSITION 3.2. *On the basis of the hypotheses of Proposition* 2.1, *if* $\tau(x)$ *is an admissible solution to* (3.5), *then for any* $\varepsilon > 0$, *the control* $u_\theta(x)$ *given by* (3.11) *satisfies* $\|u_\theta(x)\|_{p,\mathbf{r}} < 1$ *and the closed-loop system* (1.1)–(3.11) *is GAS.*

Proof. First of all, we show that control $u_\theta(x)$ is bounded in the $p, \mathbf{r}$-norm. In fact, departing from the fact that $\tau(x) \geq \|v(x)\|_{q,1/\mathbf{r}}$ for all $x \in \mathbb{R}^n$, it follows that

$$(3.13) \qquad r_j \left(\frac{r_j\,|v_j|}{\tau} \right)^{q-1} \leq r_j \left(\frac{r_j\,|v_j|}{\|v\|_{q,1/\mathbf{r}}} \right)^{q-1} \qquad \text{for } j = 1, \dots, m, \text{ and } x \neq 0.$$

Hence, the proposed control (3.11) subestimates the *singular control*

$$(3.14) \qquad \omega_p(x) := -\Big(r_1\, \xi_1^{q-1}(x), \dots, r_m\, \xi_m^{q-1}(x) \Big)^{\top}, \quad \text{where } \xi_j = \frac{r_j\, v_j}{\|v\|_{q,1/\mathbf{r}}},$$

for $j = 1, \dots, m$, that lies precisely on the boundary of the associated closed ball $\mathcal{B}_{\mathbf{r}}^m(p)$. Then, control $u_\theta(x)$ (3.11) takes values in the *open ball* $\operatorname{int} \mathcal{B}_{\mathbf{r}}^m(p)$ ($\|u_\theta\|_{p,\mathbf{r}} < 1$), for given p and vector parameter $\mathbf{r}$. That means, control (3.11) *never saturates.*

Finally, assume that control $v(x)$ fulfills the hypotheses of Proposition 2.1 and $\tau(x)$ is admissible. Then, GAS of the closed-loop system (1.1)–(3.11) is achieved observing that the control $u_\theta(x)$ satisfies the required conditions on $v(x)$, and further, $\operatorname{sign} v(x) = \operatorname{sign} u_\theta(x)$ for all $x \in \mathbb{R}^n$. □

Clearly, if $\tau(x)$ is replaced with $\|v(x)\|_{q,1/\mathbf{r}}$ in (3.11), the resulting function has the same features, but it would be a nonsmooth control: the problem of input rate bounds is nonsense. As mentioned above, for $p = 2$ we have that $\zeta_j = v_{\varepsilon_j}$, and if, further, $r_1 = \cdots = r_m = r$, then $u_\theta(x) = \rho(x)\, v(x)$, with $\rho(x) = r^2/(\varepsilon + \tau(x))$ being constant along $\partial\mathcal{E}(c)$—a condition for input rate boundedness that smooth estimates to $\|v(x)\|_{q,1/\mathbf{r}}$ might not satisfy.

We now state that the τ-based control design is globally rate-limited. Assuming a stabilization problem with control-value sets given by $p, \mathbf{r}$-normed balls might be conceived of only mathematical interest per se. Nevertheless, we will obtain an interpretation of control rate bounds in terms of the associated ball. In fact, formula (3.11) for u_θ entails a relation between expressions $u_\theta(x)$ and $v_\varepsilon(x)$ (3.10), in such a manner that du_θ/dt can satisfy any prescribed constraint (defined as a p, κ-normed ball), whenever dv_ε/dt is proven bounded for some system. Thus, the following result, based on the assumption that dv_ε/dt is bounded, is more subtle than one could determine at first sight. The problem of *bounded input rates* is solved in the following sense. Fix the control bound parameter $\mathbf{r}$, and consider that inputs are constrained to take values in the control-value set $\mathcal{B}_{\mathbf{r}}^m(\infty) := [-r_1, r_1] \times \cdots \times [-r_m, r_m] \subset \mathbb{R}^m$, with $r_j > 0$, for $j = 1, \dots, m$. Then, given an a priori control rate bound $\kappa = (\kappa_1, \dots, \kappa_m)$, with $\kappa_j > 0$, for $j = 1, \dots, m$, there is a $p, \mathbf{r}$-normed ball $\mathcal{B}_{\mathbf{r}}^m(p)$ ($\subset \mathcal{B}_{\mathbf{r}}^m(\infty)$) and a p, κ-normed ball $\mathcal{B}_{\kappa}^m(p)$, such that control function $u_\theta(x)$ and its derivative $du_\theta(x)/dt$ are bounded in the $p, \mathbf{r}$-norm and p, κ-norm, respectively.

PROPOSITION 3.3. *For a fixed control bound* $\mathbf{r}$, *assume that* $\tau(x)$ *is a* C^α *function and* $dv_\varepsilon(x)/dt$ *has a global bound, where* $v_\varepsilon(x)$ *is given in* (3.10). *Then, for any* $\kappa =$

$(\kappa_1, \dots, \kappa_m)^\top \in \mathbb{R}^m_{>0}$, *there exists a rational number* $q \in \mathbb{Q}^*_{>1}$ *(depending on* κ*), such that control* $u_\theta(x)$ (3.11) *satisfies* $\|u_\theta(x)\|_{p,\mathbf{r}} < 1$ *and its derivative* $\|du_\theta(x)/dt\|_{p,\kappa} < 1$ *for all* $x \in \mathbb{R}^n$*, where* $p \in \mathbb{Q}^*_{>1}$ *is obtained from* (3.3).

Proof. First of all, we have to consider two cases, depending on the value of q.

1. If $q \geq 2$, from (3.11) we obtain

$$\frac{du_{\theta_j}}{dt} = r_j \theta\, z_j'(\upsilon_{\varepsilon_j}) \left(\frac{d\upsilon_{\varepsilon_j}}{dt}\right) = r_j \theta\, (q-1)\, \left(\theta \upsilon_{\varepsilon_j}\right)^{q-2} \left(\frac{d\upsilon_{\varepsilon_j}}{dt}\right) \tag{3.15}$$

for $j = 1, \dots, m$. Observe that, since for all $x \in \mathbb{R}^n$, $|\theta \upsilon_{\varepsilon_j}(x)| < 1$, then $z_j'(\upsilon_{\varepsilon_j}(x))$ is globally bounded for fixed q, and $\lim_{q\to\infty} z_j'(\upsilon_{\varepsilon_j}) \leq \lim_{q\to\infty}(q-1)\theta^{q-2} = 0$.

2. If $1 < q < 2$, from (3.11), we obtain

$$\frac{du_{\theta_j}}{dt} = r_j \theta\, \widehat{z}_j'(\upsilon_{\varepsilon_j}) \left(\frac{d\upsilon_{\varepsilon_j}}{dt}\right) = \frac{r_j \theta\, (q-1)^3}{\left((2-q)\, \widehat{z}_j^{(2-q)/(q-1)}(\upsilon_{\varepsilon_j}) + (q-1)^4\right)} \left(\frac{d\upsilon_{\varepsilon_j}}{dt}\right) \tag{3.16}$$

for $j = 1, \dots, m$. From this, it follows that $\lim_{q\to 1} \max_{\upsilon_{\varepsilon_j}} \widehat{z}_j'(\upsilon_{\varepsilon_j}) = \lim_{q \to 1} \widehat{z}_j'(0) = \infty$. Moreover, $\widehat{z}_j'(\upsilon_{\varepsilon_j}(x))$ is globally bounded for fixed q.

Then, from the previous two items, we have that each function $du_{\theta_j}(x)/dt$ will have an a priori global bound value whenever $d\upsilon_{\varepsilon_j}(x)/dt$ is globally bounded for $j = 1, \dots, m$. Therefore,

$$\left\|\frac{du_\theta}{dt}\right\|_{p,\kappa} = \left(\sum_{j=1}^m \frac{1}{\kappa_j^p} \left(\frac{du_{\theta_j}}{dt}\right)^p\right)^{1/p} = \theta \left(\sum_{j=1}^m \left(\frac{r_j}{\kappa_j} \frac{d\zeta_j}{d\xi_j}(\upsilon_{\varepsilon_j})\right)^p \left(\frac{d\upsilon_{\varepsilon_j}}{dt}\right)^p\right)^{1/p}$$

$$\leq \theta \left\|\frac{d\upsilon_\varepsilon}{dt}\right\|_p \left(\sum_{j=1}^m \left(\frac{r_j}{\kappa_j} \frac{d\zeta_j}{d\xi_j}(\upsilon_{\varepsilon_j})\right)^p\right)^{1/p} \leq \theta \left\|\frac{d\upsilon_\varepsilon}{dt}\right\|_1 \left\|RK^{-1}\, D_\xi \zeta(\upsilon_{\varepsilon_j})\right\|_1 = (\triangle),$$

where $R = \operatorname{diag}(r_1, \dots, r_m)$, $K = \operatorname{diag}(\kappa_1, \dots, \kappa_m)$, and $\|\cdot\|_1$ denotes the usual 1-norm. Since by assumption $\|d\upsilon_\varepsilon/dt\|_1 \leq \widehat{\kappa} - \lambda$ globally bounded in any norm (equivalence of norms in $\mathbb{R}^n$), from (3.15) and (3.16) we obtain

$$(\triangle) \leq \varphi(q) := \begin{cases} \theta \widehat{\kappa}\, \left\|RK^{-1}\right\|_1 (q-1)^{-1} & \text{if } q \in \mathbb{Q}^*_{>1} \cap (1, 2], \\ \theta \widehat{\kappa}\, \left\|RK^{-1}\right\|_1 (q-1)\, \theta^{q-2} & \text{if } q \in \mathbb{Q}^*_{>1} \cap [2, \infty). \end{cases} \tag{3.17}$$

Denote by $\overline{\varphi}(q)$ the extension of $\varphi(q)$ to the open interval $(1, \infty) \subset \mathbb{R}$. Observe that $\overline{\varphi} : (1, \infty) \to (0, \infty)$ is well defined. Further, it is a continuous and (via a continuity argument) *surjective* function *onto* $(0, \infty)$. Given any $\xi \in (0, \infty)$, there exists $\overline{q} \in (1, \infty)$ such that $\overline{\varphi}(\overline{q}) = \xi$. Hence, recalling that set $\mathbb{Q}^*_{>1}$ is dense in $(1, \infty)$, there exists a $q^* \in \mathbb{Q}^*_{>1}$ which can be taken as closely as desired to $\overline{q}$ such that $\varphi(q^*) < \xi$. Therefore, we have that there exists $q^* \in \mathbb{Q}^*_{>1}$ such that $\varphi(q^*) < 1$.

Consequently, given any $\kappa = (\kappa_1, \dots, \kappa_m)$ with $\kappa_j > 0$ for $j = 1, \dots, m$, there is a rational number $q \in \mathbb{Q}^*_{>1}$ such that $\|du_\theta(x)/dt\|_{p,\kappa} < 1$ for all $x \in \mathbb{R}^n$, where $p \in \mathbb{Q}^*_{>1}$ is obtained from (3.3). □

Remark 2. On the basis of the above theorem, for fixed magnitude control bound $\mathbf{r}$, the $q, 1/\mathbf{r}$-norm measures the maximal rate bounds that control $u_\theta(x)$ is allowed to have, assigning the corresponding control-value sets $\mathcal{B}^m_{\mathbf{r}}(p) \subset \mathcal{B}^m_{\mathbf{r}}(\infty)$. Hence, it shows how close control $u_\theta(x)$ is to the singular control $\omega_\infty(x) = \operatorname{sign} v(x) \in \partial\mathcal{B}^m_{\mathbf{r}}(\infty)$.

In fact, for q close to 1 (i.e., p is large enough), $\mathcal{B}_{\mathbf{r}}^m(p) \nearrow \mathcal{B}_{\mathbf{r}}^m(\infty)$, the maximum value of rate bound κ is achieved, and $u_\theta(x) \to \omega_\infty(x)$; whereas for a larger q (i.e., p tends to 1), $\mathcal{B}_{\mathbf{r}}^m(p) \searrow \mathcal{B}_{\mathbf{r}}^m(1)$, the value of κ is minimized, and $u_\theta(x)$ is farther from $\omega_\infty(x)$.

In particular, if the Lyapunov function $V(x)$ (considered in Hypothesis H1) is an $\mathcal{E}$-function, the BFC function can be redefined to take values on the *closed* ball $\mathcal{B}_{\mathbf{r}}^m(p)$. First of all, we have that the equality $S_{\tau^*} = \{x \in \mathbb{R}^n : \tau(x) \leq \tau^*\} = \mathcal{E}(c(\tau^*))$ is fulfilled for any $\tau^* > 0$, so that S_{τ^*} is an *invariant set* under all the trajectories of the closed-loop system (1.1) with control $u_\tau(x)$ given by either (3.6) or (3.9). As mentioned above, both controls $u_\tau(x)$ are smooth except at the origin, where they are *singular*. Nevertheless, a redefinition yields the BFC function

$$u^\times(x) = \begin{cases} u_\tau(x) & \text{if } x \in \mathbb{R}^n \backslash \mathcal{E}(\tau^\times), \\ \\ u_{\tau^\times}(x) & \text{if } x \in \mathcal{E}(\tau^\times), \end{cases} \tag{3.18}$$

where $u_{\tau^\times}(x) := u_\tau(x)\mid_{\tau=\tau^\times}$ and $\tau^\times > 0$ is small enough. Observe that this controller is C_L^α in $\mathbb{R}^n \backslash \partial\mathcal{E}(\tau^\times)$ (whenever $\tau(x)$ is) and it is everywhere Lipschitz continuous (a C_L^α-like function).

Remark 3. Due to the fact that $v(0) = 0$, there exists a small enough $\overline{\tau} > 0$ (in a neighborhood of 0, $\|v(x)\|_{q,1/\mathbf{r}}$ is a monotonic increasing function, so that $\tau(x)$ is also an increasing function) such that for any $\tau^\times \in (0, \overline{\tau}]$, we have $u_{\tau^\times} : \mathcal{E}(\tau^\times) \to \mathcal{B}_{\mathbf{r}}^m(p)$, i.e., for all $x \in \mathcal{E}(\tau^\times)$, control $u_{\tau^\times}(x)$ is bounded in the $p, \mathbf{r}$-norm.

Remark 4. In the case of those systems for which $\tau^\times > 0$ can be taken as arbitrary, the proper function $E(x)$ should be chosen as the associated Lyapunov function $V(x)$ in the above control design, because the neighborhood $\mathcal{E}(\tau^\times)$ must be an invariant set. Otherwise, there could be many excursions of a closed-loop trajectory outside $\mathcal{E}(\tau^\times)$, so that the feedback control (3.18) will lose smoothness any time a trajectory crosses $\partial\mathcal{E}(\tau^\times)$. Moreover, the optimization process to maintain the control's boundedness would be "*set on*" any time such a trajectory leaves that neighborhood.

PROPOSITION 3.4. *On the basis of the hypotheses of Proposition* 2.1, *if* $\tau(x)$ *is an admissible solution to the optimization problem* (3.5), *there exists a* $\overline{\tau} > 0$ *such that for any* $\tau^\times \in (0, \overline{\tau}]$, *the control* (3.18) *satisfies* $\|u^\times(x)\|_{p,\mathbf{r}} \leq 1$ *and the closed-loop system* (1.1)–(3.18) *is GAS.*

Finally, if the Lyapunov function $V(x)$ is an $\mathcal{E}$-function, a combination of the proposed control designs $u_\theta(x)$ given in (3.11) and $u^\times(x)$ given in (3.18) yields the following BFC function. Given $\tau^\times > 0$ and $\varepsilon > 0$, small enough parameters, we define the BFC function

$$u_\theta^\times(x) := \left(r_1\, \zeta_1^\times\left(v_{\varepsilon_1}(x)\right), \ldots, r_m\, \zeta_m^\times\left(v_{\varepsilon_m}(x)\right)\right)^\top, \tag{3.19}$$

with

$$\zeta_j^\times(v_{\varepsilon_j}(x)) = \begin{cases} \zeta_j(v_{\varepsilon_j}(x)) & \text{if } x \in \mathbb{R}^n \backslash \mathcal{E}(\tau^\times), \\ \\ \zeta_j(v_{\varepsilon_j}^\times(x)) & \text{if } x \in \mathcal{E}(\tau^\times), \end{cases} \tag{3.20}$$

where $\zeta_j(v_{\varepsilon_j})$ is given in (3.12) and $v_{\varepsilon_j}^\times(x) := v_{\varepsilon_j}(x)\mid_{\tau=\tau^\times}$, for $j = 1, \ldots, m$.

Even though control $u_\theta(x)$ (3.11) is smooth (whenever $\tau(x)$ is), it has the disadvantage that the optimization process (3.5) to calculate $\tau(x)$ might be carried out indefinitely, unless an explicit solution to it can be found. Besides, using $u_\theta^\times(x)$ (3.19), the "smoothness" (C^α) requirement on $\tau(x)$ from Proposition 3.3 can be relaxed to be a C_L^α function, as will be asked for in section 6. Thus, control $u_\theta^\times(x)$ (3.19) has an advantage over control design based on $u_\theta(x)$ (3.11).

For the remainder of this section, we shall focus on the *properties of the function* $\tau(x)$. First of all, it is obvious that (i) $\tau(x)$ is well defined on $\mathbb{R}^n$ since (3.5) is proper; and (ii) $\tau(x) \geq 0$.

It should be pointed out that, as stated, the optimization problem (3.5) could be nonsmooth over the set $N = \{x \in \mathbb{R}^n : v(x) = 0\}$. That trouble is overcome if $\|v(x)\|_{q,1/\mathbf{r}}^q$ is used instead as an objective function. On the other hand, control (3.19) is smooth whenever the function $\tau(x)$ is. Thus, smoothness of $\tau(x)$ is an important property that is guaranteed if the following two items hold:

1. smoothness of $\tau^q(x)$, defined as the solution to (3.5) with $\|v(x)\|_{q,1/\mathbf{r}}^q$ used as objective function, and

2. positive definiteness of $\tau(x)$.

For the *general* optimization problem, smoothness results for a function as $\tau^q(x)$ are of *generic* nature (see [7, 8]). Besides its obvious relevance on precluding *singularities* (aside from $x = 0$) on the controllers $u_\tau(x)$ and the stability results from Propositions 3.2 and 3.4, the second item entails the importance of requiring positive definiteness of τ. Thereby, in view of its own importance, we will focus on the latter problem, and further, we shall consider that the Lyapunov function $V(x)$ is an $\mathcal{E}$-function and control $v(x)$ is defined as $v(x) = -h(x) = -(L_gV(x))^\top$.

Remark 5. Positive semidefiniteness of $\tau(x)$ can be reduced to the following equivalences: $\tau(x) = 0 \Longleftrightarrow \|h(x^*)\|_{q,1/\mathbf{r}} = 0$, where x^* is an optimal point of (3.5) $\Longleftrightarrow$ for all $x \in \partial\mathcal{E}(c)$, where $c = V(x^*) \geq 0$, we have $\|h(x)\|_{q,1/\mathbf{r}} = 0$ (since $\|h(x)\|_{q,1/\mathbf{r}} \leq \|h(x^*)\|_{q,1/\mathbf{r}} = 0$) $\Longleftrightarrow$ for all $x \in \partial\mathcal{E}(c)$, $h(x) = (L_gV(x))^\top = 0 \Longleftrightarrow$ the set of vector fields $\{g_1(x), \ldots, g_m(x)\}$ is *tangential* to the whole compact set $\partial\mathcal{E}(c)$.

Hereafter, in view of the previous remark, we will assume the following.

HYPOTHESIS H3. *Assume that* $n > 1$ *and, further,* $L_gV(x)$ *does not vanish on a whole boundary level set* $\partial\mathcal{E}(c)$ *(*$c = 0$ *being the only exception), i.e., no* $c > 0$ *exists such that for all* $x \in \partial\mathcal{E}(c)$*,* $L_gV(x) = 0$.

The above hypothesis is not too restrictive, since the set of vector fields that are *transversal* to a fixed proper function ($V(x)$ in this case) is dense and open (with respect to the *Whitney topology*) in the set of all vector fields defined on $\mathbb{R}^n$. This statement follows from the results presented in [29] for general *k-jets*. Consequently, $\tau(x)$ is positive-definite, and hence an admissible function, in a *generic sense*.

Finally, we have that the function $\tau(x)$ is proved admissible in the following particular, though important, cases. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *homogeneous of degree* β iff for any $\lambda \in \mathbb{R}$, $f(\lambda x) = \lambda^\beta f(x)$; and it is said to be an *odd (even) function* iff for all $x \in \mathbb{R}^n$, $f(-x) = -f(x)$ $(= f(x))$.

PROPOSITION 3.5. *Consider system* (1.1) *and* $n > 1$. *Then,* $\tau(x)$ *is an admissible function in the following special cases:*

(a) *if* $g(x) = B = (b_j)$*—a constant* $n \times m$ *matrix;*

(b) *if* $g_j(x)$ *for* $j = 1, \ldots, m$*, and the associated Lyapunov function* $V(x)$ *are homogeneous;*

(c) *if* $g(x) = d(x) + b$*, where* $0 \neq b \in \mathbb{R}^n$ *is a constant and* $d_j(x)$ *are odd functions for* $j = 1, \ldots, m$*, and the uncontrolled system has an even Lyapunov function* $V(x)$.

Proof. The three items can be proved by reductio ad absurdum: assume that there exists $c > 0$ such that $\tau(x)|_{\partial\mathcal{E}(c)} \equiv 0$.

Case (a). In this case, the systems $\dot{x}= g_j(x) = b_j$, $j = 1, \ldots, m$, have parallel solutions, and thus $\Delta = \bigcap_{j=1}^m \ker b_j$ is a linear subspace with $\dim \Delta < n$. On the other hand, for each $c > 0$, the level set $\mathcal{E}(c)$ satisfies $\operatorname{int}\mathcal{E}(c) \neq \emptyset$ (due to continuity of $V(x)$), and the (boundary) level set $\partial\mathcal{E}(c)$ is compact (due to properness of $V(x)$). Therefore, the set of vector fields $\{b_1, \dots, b_m\}$ cannot be tangential to a whole compact set $\partial\mathcal{E}(c)$, unless $\partial\mathcal{E}(c) \subset \Delta$, but this is a contradiction.

Case (b). In view that $V(x)$ is positive-definite and $c > 0$, then $0 \in \operatorname{int}\mathcal{E}(c) \neq \emptyset$. Due to compactness of $\mathcal{E}(c)$, a continuity argument shows that given any $y \in \mathbb{R}^n$, there exist $x \in \partial\mathcal{E}(c)$ and $\eta > 0$ such that $y = \eta x$. Therefore, since $g_1(x), \dots, g_m(x)$ and $V(x)$ are homogeneous functions of degrees γ_j and β, respectively, we should obtain that for all $y \in \mathbb{R}^n$, $-h_j(y) = -L_{g_j}V(y) = -\eta^{\gamma_j+\beta-1}L_{g_j}V(x) = 0$, $j = 1, \dots, m$. Consequently, if we assume that $\tau(x)$ vanishes in a whole boundary level set $\partial\mathcal{E}(c)$, then either $g(x) \equiv 0$ or $V(x) \equiv 0$ in the whole space $\mathbb{R}^n$.

Case (c). In this case, the set of vector fields $g_j(x) = d_j(x) + b_j$, $j = 1, \dots, m$, should be tangential to the whole symmetric set $\partial\mathcal{E}(c) = \{x \in \mathbb{R}^n : V(x) = c\}$. In fact, suppose that for all $x \in \partial\mathcal{E}(c)$, $-h_j(x) = -L_{g_j}V(x) = -\nabla V(x)(d_j(x) + b_j) = 0$, $j = 1, \dots, m$. However, by symmetry, $-x \in \partial\mathcal{E}(c)$, and thus, $-h_j(-x) = -L_{g_j}V(-x) = -2\nabla V(x)\, b_j = 0$, $j = 1, \dots, m$, but this contradicts case (a). □

4. Suboptimal properties.

In this section we identify a control problem that involves a notion of optimality related to the associated Lyapunov function. Then, we will show that the proposed control $u_\theta(x)$ given in (3.11) is a suboptimal solution to that problem.

Assume that the control-value set $U \subset \mathbb{R}^m$ is compact and $0 \in U$, and denote by $\mathcal{U}^*$ the set of all U-valued piecewise continuous functions defined on $\mathbb{R}$. Although the Lyapunov approach based on Hypotheses H1–H2 suffices to analyze global stability, the fact that $V(x)$ is not strictly decreasing along the closed-loop solutions does not give a “margin” that can be exploited for robustness analysis or achieving certain performance. An approach based on the existence of a Lyapunov function for the closed-loop system could be more convenient (see [1, 15, 3]). A C^α ($\alpha \geq 1$) function $V : \mathbb{R}^n \to \mathbb{R}$ is a Lyapunov function for system (1.1) iff it is positive definite, proper, and $L_{f+\Sigma_{j=1}^m g_j u_j}V(x(t)) < 0$, where $x(t)$ is a solution to (1.1) with $u(t) \in \mathcal{U}^*$. Since $dV/dt = L_{f+\Sigma_{j=1}^m g_j u_j}V(x(t))$ represents the decay rate of $V(x(t))$, a reasonable criterion for choosing $u(t)$ is to minimize dV/dt subject to $u \in \mathcal{U}^*$. The resulting minimal control can be represented as a feedback control: for each value $x(t)$ denote by $\omega(x(t))$ the minimal value of $u \in \mathcal{U}^*$.

Definition 4.1. *Let $V : \mathbb{R}^n \to \mathbb{R}$ be a C^α ($\alpha \geq 1$) positive-definite function. We say that $\omega : \mathbb{R}^n \to U$ is optimal with respect to $V(x)$ iff for all $x \in \mathbb{R}^n$ and all $u \in \mathcal{U}^*$*

$$L_{f+\Sigma_{j=1}^m g_j \omega_j}V(x) \leq L_{f+\Sigma_{j=1}^m g_j u_j}V(x). \tag{4.1}$$

If, further, $V(x)$ satisfies that for all $x \neq 0$, $L_{f+\Sigma_{j=1}^m g_j \omega_j}V(x) < 0$, then $V(x)$ is a Lyapunov function for the closed-loop system $\dot{x} = f(x(t)) + \sum_{j=1}^m g_j(x(t))\, \omega_j(x(t))$.

In [3], the problem of finding an optimal $\omega(x)$ has been illustrated by considering special cases of set U: the Euclidean ball of radius r, $\mathcal{B}_{\mathbf{r}}^m(\infty)$ and $\mathcal{B}_{\mathbf{r}}^m(1)$. Observe that inequality (4.1) is equivalent to

$$\min_{u \in \mathcal{U}^*} \{L_g V(x)\, u\}\,. \tag{4.2}$$

For control-value set $U = \mathcal{B}_{\mathbf{r}}^m(p)$ for $1 < p < \infty$, it can be proven (via the Hölder inequality) that the minimal value is given by $-\|h(x)\|_{q,1/\mathbf{r}}$ and it is accomplished by the BFC function $\omega_p(x)$ given in (3.14) with $v(x) = -h(x)$. Control $\omega_p(x)$ lies on $\partial\mathcal{B}_{\mathbf{r}}^m(p)$

and is the unique *optimal* BFC function related to the $p, \mathbf{r}$-norm, in the sense that it is a solution to the equation $\min_{u \in \mathcal{U}^*}\{\sum_{j=1}^m L_{g_j}V(x)\, u_j\} = \sum_{j=1}^m L_{g_j}V(x)\, \omega_{p_j}(x)$. Uniqueness follows from the strict convexity of $U = \mathcal{B}_{\mathbf{r}}^m(p)$, but it is understood modulo the set $N = \{x \in \mathbb{R}^n : L_gV(x) = 0\}$ because $\omega_p(x)|_N$ is arbitrary (a *singular control*). Hence, control $\omega_p(x)$ is *not an admissible input* with respect to the set $\mathcal{U}$ given in (1.4). Thereby, given a Lyapunov function, $\omega_p(x)$ is optimal in the sense that it maximizes the *robustness stability margin* against bounded uncertainties (see [21]).

The aforementioned optimal problem is also the basis of the CLF approach for feedback stabilization. Indeed, Artstein's theorem [1] expresses that the existence of a smooth CLF is equivalent to the existence of a continuous feedback control that renders the GAS of the resulting closed-loop system. Moreover, Artstein's theorem holds for general *convex* (possibly constrained) control-value sets. However, in view that control design derived from the proof in [1] is nonconstructive (based on partitions of unity), Sontag and coworkers have proposed *explicit "universal" formulae* for *almost smooth* feedback control laws that stabilize the system (1.1), provided an appropriate CLF is known. In the bounded control case, such formulae were obtained to fullfil specific control-value sets U: the Euclidean open unit ball in [15] and Minkowski open unit balls (with $1 < p \leq 2$) in [19]. Along this line of thought, and possibly conceived as a pure mathematical problem *per se*, in [25], Sontag presented the following as an important *open* problem: *Find universal formulas for CLF stabilization for general (convex) control-value sets U.*

Remark 6. It should be noted that feedback control $u_\theta(x)$ given in (3.11) shares a similar structure to the formula (3.14) for $\omega_p(x)$ and, from the proof of Proposition 3.2, also subestimates it. Hence, if V is also a Lyapunov function for system (1.1), $u_\theta(x)$ is a smooth approximation to the optimal control $\omega_p(x)$.

Summarizing, if Hypothesis H1 holds and V is a Lyapunov function for system (1.1), the special control $u_\theta(x)$ (3.11) is threefold: (1) it is an arbitrarily small control (i.e., lies in arbitrary $p, \mathbf{r}$-normed balls) that stabilizes system (1.1); (2) it is a smooth *suboptimal robust control*; and (3) it allows us to address the problem of input *rates* constrained to lie in prescribed sets, in the context of CLF stabilization.

5. Control design when $U = \prod_{i=1}^{\mu} \mathcal{B}_{\mathbf{r}_i}^{m_i}(p_i)$. In this section we introduce a natural extension of the control design method developed in section 3 that allows us to consider all permutations of the m control input entries arbitrarily taking values in possibly different $p, \mathbf{r}$-normed balls. Without loss of generality, this condition is accomplished by renaming the control input entries (if necessary) so that u is partitioned as $u = (\mathbf{u}_1, \ldots, \mathbf{u}_\mu)^\top$, with each *control block* $\mathbf{u}_i$ taking values in a different $p_i, \mathbf{r}_i$-normed ball. For this aim, let v be the global stabilizer considered in Proposition 2.1, and take the partition $v = (\mathbf{v}_1, \ldots, \mathbf{v}_\mu)$. Denote $\mathbf{p} = (p_1, \ldots, p_\mu)$ with $p_i \in \mathbb{Q}^*_{>1}$, and define the vector parameter $\mathbf{q} = (q_1, \ldots, q_\mu)$ in such a way that p_i and q_i satisfy (3.3). Hence, each expression $(q_i - 1)$ is a quotient of two odd numbers, so that the proposed control (see below in (5.3)) will preserve the sign of the original control block $\mathbf{v}_i(x)$. (This condition is needed for the stabilization result from Proposition 2.1.)

Assume that the control-value set is given by the following Cartesian product of $p, \mathbf{r}$-normed balls:

$$U = \mathcal{B}_{\mathbf{r}_1}^{m_1}(p_1) \times \cdots \times \mathcal{B}_{\mathbf{r}_\mu}^{m_\mu}(p_\mu) \tag{5.1}$$

for $1 \leq \mu \leq m$ and $m_1 + \cdots + m_\mu = m$.

Choose an arbitrary $0 \neq x_0 \in \mathbb{R}^n$ and set $c = E(x_0)$ (> 0). In order to obtain the stabilizing BFC function, we require the solution to the following μ parametric optimization problems:

$$\begin{gathered} \tau_i(x_0) := \max_{x\in\partial\mathcal{E}(c)} \|\mathbf{v}_i(x)\|_{q_i,1/\mathbf{r}_i}\,, \\ \text{s.t.} \\ \partial\mathcal{E}(c) := \{x \in \mathbb{R}^n : E(x) = c = E(x_0)\}\,. \end{gathered} \tag{5.2}$$

A vector of solutions $\tau(x) = (\tau_1(x), \dots, \tau_\mu(x))$ to the set of optimization problems (5.2) will be called *admissible* iff each $\tau_i(x)$ is a positive definite function.

Then, for admissible $\tau(x)$, the BFC function relative to U given in (5.1) is $u_\theta = (\mathbf{u}_{\theta_1}, \dots, \mathbf{u}_{\theta_\mu})^\top$, where formulae for $\mathbf{u}_{\theta_i}$ are analogous to (3.11), i.e.,

$$\mathbf{u}_{\theta_i} := \big(\overset{[i]}{r}_1\, \overset{[i]}{\zeta}_1\big(\overset{[i]}{v}_{\varepsilon_1}\big), \dots, \overset{[i]}{r}_{m_i}\, \overset{[i]}{\zeta}_{m_i}\big(\overset{[i]}{v}_{\varepsilon_{m_i}}\big)\big)^\top, \ \text{with} \quad \overset{[i]}{v}_{\varepsilon_j}(x) := \frac{\overset{[i]}{r}_j\, \overset{[i]}{v}_j(x)}{\varepsilon_i + \tau_i(x)}\,, \tag{5.3}$$

for $j = 1, \dots, m_i$, $\varepsilon_i > 0$, are tuning parameters for $i = [i] = 1, \dots, \mu$, and the state-functions $\overset{[i]}{\zeta}_j$ are defined

$$\begin{cases} \text{explicitly}: \quad \overset{[i]}{\zeta}_j = (\theta \overset{[i]}{v}_{\varepsilon_j})^{1/(p_i-1)} & \text{if } p_i \in \mathbb{Q}^*_{>1} \cap (1,2], \\ \\ \text{implicitly}: \quad (p_i-2)(\overset{[i]}{\zeta}_j)^{p_i-1} + \overset{[i]}{\zeta}_j = (p_i-1)\theta \overset{[i]}{v}_{\varepsilon_j} & \text{if } p_i \in \mathbb{Q}^*_{>1} \cap [2,\infty), \end{cases} \tag{5.4}$$

with $0 < \theta \approx 1$ (needed in Proposition 5.2). In the implicit case, we have that $(\overset{[i]}{\zeta}_j)'(0) = p_i - 1$.

The following result follows along the line of the proof of Proposition 3.2.

PROPOSITION 5.1. *On the basis of the hypotheses of Proposition* 2.1, *if* $\tau(x)$ *is an admissible solution to the set of optimization problems* (5.2), *then for any* $\varepsilon = (\varepsilon_1, \dots \varepsilon_\mu)$, *with* $\varepsilon_i > 0$, *the control* $u_\theta(x)$ *given in* (5.3) *satisfies* $\|\mathbf{u}_{\theta_i}(x)\|_{p_i,\mathbf{r}_i} < 1$ *for* $i = 1, \dots, \mu$, *and the closed-loop system* (1.1)–(5.3) *is GAS.*

Moreover, the proposed control design (5.3) leads also to address the problem of *globally rate-limited* actuators. Fix the control bound parameter $\mathbf{r}$, and consider that inputs are constrained to take values in the m-dimensional $\mathbf{r}$-hyperbox, $\mathcal{B}^m_{\mathbf{r}}(\infty)$. Then, given an a priori control rate bound $\kappa = (\kappa_1, \dots, \kappa_\mu)$, with $\kappa_i = (\overset{[i]}{\kappa}_1, \dots, \overset{[i]}{\kappa}_{m_i})^\top \in \mathbb{R}^{m_i}_{>0}$, there is a vector of rational numbers $\mathbf{q} = (q_1, \dots, q_\mu)$, with $q_i \in \mathbb{Q}^*_{>1}$ for $i = 1, \dots, \mu$, such that control block $\mathbf{u}_{\theta_i}$ takes values in $\mathcal{B}^{m_i}_{\mathbf{r}_i}(p_i)$ and $d\mathbf{u}_{\theta_i}/dt$ takes values in $\mathcal{B}^{m_i}_{\kappa_i}(p_i)$, with p_i and q_i satisfying (3.3).

The following result follows along the lines of the proof of Proposition 3.3.

PROPOSITION 5.2. *Consider a fixed control bound* $\mathbf{r}$ *and assume that* $d\overset{[i]}{v}_{\varepsilon_j}(x)/dt$ *have global bounds for* $j = 1, \dots, m_i$ *and* $i = 1, \dots, \mu$, *where* $\overset{[i]}{v}_{\varepsilon_j}(x)$ *is given in* (5.3). *Then, for any* $\kappa = (\kappa_1, \dots, \kappa_\mu)^\top \in \mathbb{R}^m_{>0}$, *there is a* $\mathbf{q} = (q_1, \dots, q_\mu)^\top \in (\mathbb{Q}^*_{>1})^\mu$ *(depending on* κ*), such that each control block* (5.3) *satisfies* $\|\mathbf{u}_{\theta_i}\|_{p_i,\mathbf{r}_i} < 1$ *and* $\|d\mathbf{u}_{\theta_i}/dt\|_{p_i,\kappa_i} < 1$, *with* p_i *and* q_i *satisfying* (3.3) *for* $i = 1, \dots, \mu$.

Analogously to the case of control given in (3.19), if the associated Lyapunov function $V(x)$ (from Hypothesis H1) is an $\mathcal{E}$-function, a combined control design $u_\theta(x)$ with $u^\times(x)$ given in (3.18) can be proposed. Hence, given a set of small enough parameters $\tau_i^\times > 0$, and tuning parameters $\varepsilon_i > 0$, $i = 1, \dots, \mu$, we define the BFC function $u_\theta^\times = (\mathbf{u}^\times_{\theta_1}, \dots, \mathbf{u}^\times_{\theta_\mu})^\top$ with

$$\mathbf{u}^\times_{\theta_i} := \big(\overset{[i]}{r}_1\, \overset{[i]}{\zeta}{}^\times_1\big(\overset{[i]}{v}_{\varepsilon_1}\big), \dots, \overset{[i]}{r}_{m_i}\, \overset{[i]}{\zeta}{}^\times_{m_i}\big(\overset{[i]}{v}_{\varepsilon_{m_i}}\big)\big)^\top \ \text{and} \tag{5.5}$$

$$\zeta_j^{[i]\times}(v_{\varepsilon_j}^{[i]}(x)) = \begin{cases} \zeta_j^{[i]}(v_{\varepsilon_j}^{[i]}(x)) & \text{if } x \in \mathbb{R}^n \backslash \mathcal{E}(\tau_i^{\times}), \\ \zeta_j^{[i]}(v_{\varepsilon_j}^{[i]\times}(x)) & \text{if } x \in \mathcal{E}(\tau_i^{\times}), \end{cases} \tag{5.6}$$

where $\zeta_j^{[i]}(v_{\varepsilon_j}^{[i]})$ is given in (5.4) and $v_{\varepsilon_j}^{[i]\times}(x) := v_{\varepsilon_j}^{[i]}(x) \mid_{\tau_i=\tau_i^{\times}}$ for $i = 1, \ldots, \mu$.

5.1. Control design in the m-dimensional r-hyperbox. Assuming that $\mu = m$, the control-value set $U = \mathcal{B}_{\mathbf{r}}^m(\infty) := [-r_1, r_1] \times \cdots \times [-r_m, r_m] \subset \mathbb{R}^m$ is obtained. This special constraint set is important for two main reasons: (i) among all norm-based sets and fixed $\mathbf{r}$, it is the maximal set (under $\subseteq$), so that more control magnitude is available in this way; and (ii) inputs are actually independent of each other, in the sense that any input can take values without being acquainted with magnitudes taken by the remaining inputs. In this case, formula (5.3) consists of scalar control blocks $u_\theta := (r_1\, \zeta_1^{[1]}(v_{\varepsilon_1}^{[1]}), \ldots, r_m\, \zeta_m^{[m]}(v_{\varepsilon_m}^{[m]}))^\top$, with $v_{\varepsilon_j}^{[j]}(x) = (r_j\, v_j(x))/(\varepsilon_j + \tau_j(x))$, $j = 1, \ldots, m$; and expressions $\zeta_j^{[j]}(v_{\varepsilon_j}^{[j]})$ are given in (5.4), but parameters q_j do not depend on dual formula (3.3). This control design requires us to find the solution to m parametric optimization problems (5.2) corresponding to the m scalar control blocks. Thus, a closer look to the optimization problem (3.5) in the scalar input case is crucial. The optimization problem (3.5) for scalar input is equivalent to solving the two programs

$$\tau(x_0) = \max_x |v(x)| = \max\left\{\max_x v(x), -\min_x v(x)\right\}, \tag{5.7}$$

subject to $x \in \partial\mathcal{E}(c)$, where $c = E(x_0)$. This problem can be reduced to only one optimization problem if the maximum is attained always at just one program. A sufficient condition for this to hold proceeds as follows. Suppose that there exists $x_0 \in \partial\mathcal{E}(c)$ such that $\tau(x_0) = \max_x v(x) = -\min_x v(x)$ for $x \in \partial\mathcal{E}(c)$. Then, there exist x_1^*, $x_2^* \in \partial\mathcal{E}(c)$ such that $v(x_1^*) + v(x_2^*) = 0$, where x_1^* and x_2^* are optimal points of the programs $x_1^* = \arg\max_x v(x)$ and $x_2^* = -\arg\min_x v(x)$, respectively. On the basis that $\tau(x)$ is positive-definite, it follows that $x_1^* \neq x_2^*$. Then, there might exist an open interval $I = (-\delta + c, c + \delta)$, with $\delta > 0$, such that if Σ_I denotes the associated set of optimal solutions to (5.7), then Σ_I should be contained in the corresponding set of critical points of the above programs. If this is the case, Σ_I could be a disjoint union of piecewise smooth curves (cf. [12]). Henceforth, from the perspective of implementing a program, in order to avoid undesirable "jumps" between those curves, we can assume that the function $\psi : \mathbb{R} \to \mathbb{R}$, defined by

$$\psi(c) := \max_x v(x) + \min_x v(x) \quad \text{s.t.} \quad x \in \partial\mathcal{E}(c), \tag{5.8}$$

is either nonnegative or nonpositive for all $c > 0$.

Assuming that function ψ defined by (5.8) is nonnegative, the optimization problem (5.7) reduces to solve

$$\tau(x_0) = \max_{x \in \partial\mathcal{E}(c)} v(x) \geq 0, \quad \text{where } c = E(x_0). \tag{5.9}$$

Expressing the above program in terms of the first-order necessary condition for an extremum, we have that there exists $(x^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}$, where x^* is a regular point of the constraint and $\lambda^* \in \mathbb{R}$ is a Lagrange multiplier, such that the Lagrangian

$$L_{\lambda^*}(x^*) = v(x^*) - \lambda^*(E(x) - c) \tag{5.10}$$

for fixed $c > 0$ satisfies

$$\nabla_x L_{\lambda^*}(x^*) = \nabla_x v(x^*) - \lambda^* \nabla_x E(x^*) = 0. \tag{5.11}$$

When ψ (5.8) is nonpositive, the negative version of (5.9) ($\tau(x_0) = \max_x(-v(x)) \geq 0$) must be considered. The procedure developed here will be illustrated in section 6.2.

6. A class of homogeneous systems.

In order to guarantee a constructive solution to (3.5), a sufficient condition could be that there exists a homogeneous Lyapunov function $V(x)$ associated to system (1.1) and that functions $g_j(x)$, $j = 1, \ldots, m$, are also homogeneous. If this is the case, the optimization problem is not only feasible, but the computational burden involved in solving it for each value of the parameter c is drastically reduced. The problem is solved just once.

THEOREM 6.1. *Consider the one-parameter family of optimization problems* (3.5). *Assume that* (1.1) *admits an homogeneous Lyapunov function* $V(x)$ *of even degree* $\beta \geq 2$, *and* $g_1(x), \ldots, g_m(x)$ *are homogeneous functions of degrees all equal to* γ. *Denote by* ν $(= \beta + \gamma - 1)$ *the degree of* $h_j(x) = L_{g_j} V(x)$, $j = 1, \ldots, m$, *and set* $\varsigma := \beta/\nu$. *Further assume that* $x^* \in \mathbb{R}^n$ *is an optimal point of* (3.5) *satisfying, without loss of generality,* $V(x^*) = 1/\varsigma$, *such that it is a regular point of the constraint and both* $g(x^*)$ *and* $Dh(x^*)g(x^*)$ *have full rank. Then, for any* $x \in \mathbb{R}^n$, *we have*

$$\tau(x) = \sigma\, V^{1/\varsigma}(x), \quad \text{where } \sigma = \varsigma^{1/\varsigma}\, \|h^\top(x^*)\,(Dh(x^*)\,g(x^*))^{-1}\,\|_{p,\mathbf{r}}^{-1}. \tag{6.1}$$

Proof. Choose $0 \neq x_0 \in \mathbb{R}^n$, set $c = V(x_0)$ (> 0) and let $x^* \in \mathbb{R}^n$ be an optimal point of the proper program (3.5). Then, by hypothesis and based on the Lagrange multiplier method, we have that x^* is a regular point of the constraint and that there exists $\lambda^* \in \mathbb{R}$ such that the Lagrangian

$$L^{\varsigma}_{\lambda^*}(x^*) = \|h(x^*)\|^{\varsigma}_{q,1/\mathbf{r}} - \lambda^*(V(x^*) - c) \tag{6.2}$$

for fixed $c > 0$ satisfies

$$\nabla_x L^{\varsigma}_{\lambda^*}(x^*) = \frac{\varsigma\, \tau^{\varsigma}}{q\|y^*\|^q_{q,1/\mathbf{r}}} \nabla_x(\|y^*\|^q_{q,1/\mathbf{r}}) - \lambda^* \nabla_x V(x^*) = 0, \tag{6.3}$$

where $y^* = h(x^*)$. Obviously, $\|y\|^q_{q,1/\mathbf{r}}$ is a positive-definite, convex, and homogeneous function of degree q (in the variable y). Moreover, $\|h(x^*)\|^q_{q,1/\mathbf{r}} \neq 0$, since $\tau(x^*) > 0$ (Proposition 3.5 (b)). Then, by homogeneity of $V(x)$ and $h(x)$, $\nabla_x L^{\varsigma}_{\lambda^*}(x)$ is a homogeneous function of degree $\beta - 1$. By assumption, $g(x^*)$ has full rank. Thus, postmultiply both sides of (6.3) by $g(x^*)$ to obtain

$$\nabla_x L^{\varsigma}_{\lambda^*}\, g = \frac{\varsigma\, \tau^{\varsigma}}{\tau^q} \left((r_1 h_1)^{q-1}, \ldots, (r_m h_m)^{q-1}\right) R\, Dh\, g - \lambda^* h^\top = 0, \tag{6.4}$$

where $R := \operatorname{diag}(r_1, \ldots, r_m)$. Then, in virtue that $Dh\, g$ is nonsingular at x^*, postmultiplying both sides of (6.4) by $(Dh\, g)^{-1}$ yields

$$\varsigma\, \tau^{\varsigma - 1} u_\tau^\top = \lambda^* h^\top\, (Dh\, g)^{-1}. \tag{6.5}$$

On the other hand, control function $u_\tau(x)$ given in (3.6) satisfies $\|u_\tau(x^*)\|_{p,\mathbf{r}} = 1$ for any $q > 1$. Thus, using that fact and expression (6.5), we have

$$\lambda^* = \varsigma\, \tau^{\varsigma-1}\, \|h^\top\, (Dh\, g)^{-1}\, \|^{-1}_{p,\mathbf{r}}. \tag{6.6}$$

Consider again (6.3) and postmultiply both sides by x^* to obtain

$$\nabla_x L^{\varsigma}_{\lambda^*} x^* = \frac{\varsigma}{q}\tau^{\varsigma-q}\,\nabla_x(\|y^*\|^q_{q,1/\mathbf{r}})\,x^* - \lambda^*\nabla V(x^*)\,x^* = 0. \tag{6.7}$$

Observe that $V(x)$ and $\|h(x)\|^q_{q,1/\mathbf{r}}$ are homogeneous functions of degrees β and $q\nu$, respectively. Then, by (Euler's formula) homogeneity, we have that $\nabla_x V(x)\,x = \beta\,V(x)$ and $\nabla_x(\|h(x)\|^q_{q,1/\mathbf{r}})\,x = q\nu\,\|h(x)\|^q_{q,1/\mathbf{r}}$. Then, from (6.7) and due to Proposition 3.5(b) ($\|y^*\|^q_{q,1/\mathbf{r}} \neq 0$), we obtain

$$\nabla_x L^{\varsigma}_{\lambda^*} x^* = \frac{\varsigma}{q}\tau^{\varsigma-q}\,(q\nu\;\|y^*\|^q_{q,1/\mathbf{r}}) - \beta\,\lambda^* V(x^*) = \beta\;(\tau^{\varsigma} - \lambda^* V(x^*)) = 0. \tag{6.8}$$

Thereby, we have

$$\tau^{\varsigma}(x) = \lambda^* V(x) \quad \text{for any } x \in \partial\mathcal{E}(c). \tag{6.9}$$

On the other hand, using homogeneity, we have that $\Sigma = \{x \in \mathbb{R}^n : x = \eta\,x^*,\ \eta > 0\}$ is a line of critical points of $L^{\varsigma}_{\lambda^*}$ and also a set of regular points of the constraint and $g\mid_{\Sigma}$ has full rank. Moreover, based on the proof of Proposition 3.5 (b), it follows that Σ is also a set of maximal points for the optimization program (3.5). Hence any optimal point x^* can be represented as $x^* = \eta\,x^*_0$, where, without loss of generality, x^*_0 denotes an optimal point subject to $V(x^*_0) = 1/\varsigma$. Therefore, in order to define $\tau(x)$ given in (6.9) on the whole $\mathbb{R}^n$, we proceed as follows. Given an arbitrary x_0, set $c = V(x_0)$ and let $x^* \in \partial\mathcal{E}(c)$ be the corresponding optimal point. In view that $x^* = \eta\,x^*_0$, we have that $c = V(x^*) = \eta^{\beta}/\varsigma$, and substituting this expression into (6.9) yields η as a function of τ: $\eta^{\beta}(\tau) = \varsigma\;\tau^{\varsigma}/\lambda^*$. Thus,

$$x^*(\tau) = \left(\frac{\varsigma}{\lambda^*}\,\tau^{\varsigma}\right)^{1/\beta} x^*_0. \tag{6.10}$$

Rename x^*_0 as x^*. Then, replacing x^* with $x^*(\tau)$ in (6.6), taking into account the homogeneity of the functions, and after some straightforward algebraic calculations, we obtain

$$\lambda^* = \varsigma\,\|h^{\top}(x^*)\,(Dh(x^*)\,g(x^*))^{-1}\,\|^{-\varsigma}_{p,\mathbf{r}}. \tag{6.11}$$

Finally, substituting the above expression into (6.9) yields (6.1), which is defined in $\mathbb{R}^n$. □

Remark 7. It should be worth mentioning that $\tau(x)$ given in (6.1) is also a proper Lyapunov function (i.e., it is positive-definite and radially unbounded, and $d\tau/dt < 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$), provided that Hypothesis H2 holds, and it is a homogeneous function of degree ν.

THEOREM 6.2. *Consider the affine system* (1.1) *for which Hypotheses* H1–H2 *hold. Under the additional hypotheses of the above theorem, we have that the control* $u^{\times}_{\theta}(x)$ (3.19), *where* $\upsilon_{\varepsilon} = (\upsilon_{\varepsilon_1}, \ldots, \upsilon_{\varepsilon_m})^{\top}$ *given by*

$$\upsilon_{\varepsilon_j}(x) = -\frac{r_j}{\varepsilon + \sigma\,V^{1/\varsigma}(x)}\,h_j(x), \quad j = 1, \ldots, m, \tag{6.12}$$

satisfies $\|u^{\times}_{\theta}(x)\|_{p,\mathbf{r}} < 1$ *for arbitrary* $\tau^{\times} > 0$ *and* $\varepsilon > 0$, *and renders the GAS of the closed-loop system* (1.1)–(3.19), *where* σ *is given in* (6.1) *and* x^* *is an optimal point such that* $V(x^*) = 1/\varsigma$ *(without loss of generality). Furthermore, we have an a priori* *global bounded rate on the input, i.e., given any $\kappa = (\kappa_1, \dots, \kappa_m)^\top \in \mathbb{R}^m_{>0}$, there is a $q \in \mathbb{Q}^*_{>1}$ (depending on κ), such that $\|du^\times_\theta/dt\|_{p,\kappa} < 1$, with $p \in \mathbb{Q}^*_{>1}$, for any $x \in \mathbb{R}^n$, whenever $L_fV(x) \in \mathcal{O}(\|x\|^\beta_p)$, $\|Dh(x)f(x)\|_{p,1/\mathbf{r}} \in \mathcal{O}(\|x\|^\nu_p)$ and either $g(x) = B$, an $n \times m$ constant matrix, or $g_j(x) = D_jx$, with D_j being $n \times n$ constant matrices, $j = 1, \dots, m$. The value of parameter q depends on the associated order-constants, rate bound κ, control bound $\mathbf{r}$, parameter value $\tau^\times > 0$, and tuning parameter $\varepsilon > 0$.*

Proof. The stabilization result follows from Proposition 3.2 and the above theorem. Only the last assertion on the boundedness of $\|du^\times_\theta/dt\|_{p,\mathbf{r}}$ must be proved. Departing from Proposition 3.3, it turns out that du_θ/dt has an a priori global bound, whenever $d\upsilon_\varepsilon/dt$ is globally bounded. Then, from $\upsilon_\varepsilon(x)$ (6.12) and some calculations, we obtain

$$\frac{d\upsilon_\varepsilon}{dt} = (\varepsilon+\tau)^{-2}\,\nabla_x\tau\;\dot{x}\;Rh - \frac{1}{\varepsilon+\tau}\,R\,Dh\;\dot{x}$$

$$= \frac{1}{\varepsilon+\tau}\left(\frac{\sigma}{\varsigma(\varepsilon+\tau)}\,V^{\frac{1-\varsigma}{\varsigma}}\,\nabla_xV\;\dot{x}\;Rh - R\,Dh\;\dot{x}\right)$$

$$= \frac{1}{\varepsilon+\tau}\left(\frac{\sigma^\varsigma}{(\varepsilon+\tau)\,\varsigma\tau^{\varsigma-1}}\,(L_fV + L_gV\,u_\theta)\,Rh - R\,Dh\,(f + g\,u_\theta)\right),$$

where $R = \operatorname{diag}(r_1, \dots, r_m)$. Due to the fact that for all $x \in \mathbb{R}^n$, $|\upsilon_{\varepsilon_j}(x)| < 1$, $j = 1, \dots, m$, it follows that $\|\upsilon_\varepsilon\|_p < m^{1/p}$, in any p-norm. Moreover, for any $a, b \in \mathbb{R}^m$, we have (1) $\|R^{-1}a\|_{q,1/\mathbf{r}} = \|a\|_q$ and $\|R\,a\|_p = \|a\|_{p,1/\mathbf{r}}$, and (2) $|a^\top b| \le \|a\|_{q,1/\mathbf{r}}\,\|b\|_{p,\mathbf{r}}$ (Hölder inequality). Recalling that $\|u_\theta\|_{p,\mathbf{r}} < 1$, then $\|u_\theta\|_p = \varrho\|u_\theta\|_{p,\mathbf{r}} < \varrho$, with ϱ depending on $\mathbf{r}$. Hence, from the above expression and some calculations, we obtain

$$\begin{aligned}\left\|\frac{d\upsilon_\varepsilon}{dt}\right\|_p &\le \frac{\sigma^\varsigma}{(\varepsilon+\tau)\,\varsigma\tau^{\varsigma-1}}\,(-L_fV + |L_gV\,u_\theta|)\,\|\upsilon_\varepsilon\|_p + \frac{1}{(\varepsilon+\tau)}\,\|R\,Dh\,(f+g\,u_\theta)\|_p \\ &\le \frac{\sigma^\varsigma\,m^{1/p}}{\varsigma\tau^{\varsigma-1}}\left(-\frac{1}{(\varepsilon+\tau)}\,L_fV + \|R^{-1}\upsilon_\varepsilon\|_{q,1/\mathbf{r}}\,\|u_\theta\|_{p,\mathbf{r}}\right) \\ &\quad + \frac{1}{(\varepsilon+\tau)}\,\|Dh\,(f+g\,u_\theta)\|_{p,1/\mathbf{r}}\,.\end{aligned}$$

Finally,

$$\left\|\frac{d\upsilon_\varepsilon}{dt}\right\|_p \le \frac{\sigma^\varsigma m}{\varsigma\tau^{\varsigma-1}} - \frac{\sigma^\varsigma\,m^{1/p}\,L_fV}{\varsigma\tau^{\varsigma-1}(\varepsilon+\tau)} + \frac{1}{(\varepsilon+\tau)}\,\|Dh\,f\|_{p,1/\mathbf{r}} + \frac{\varrho}{(\varepsilon+\tau)}\,\|Dh\,g\|_{p,1/\mathbf{r}}\,. \tag{6.13}$$

In virtue that by assumption $\tau(x)$ is a proper function, we have that the right-hand side of the above expression is bounded in $\mathbb{R}^n\backslash\mathcal{E}(\tau^\times)$ for arbitrary $\tau^\times > 0$, if in the latter three terms, the orders of the denominators (as functions of τ) are greater than or equal to the corresponding ones of the numerators. Thus, observing that $\tau(x)$ has (degree) order ν $(= \beta + \gamma - 1)$, we obtain that $L_fV(x) \in \mathcal{O}(\|x\|^\beta_p)$ and $\|Dh(x)\;f(x)\|_{p,1/\mathbf{r}} \in \mathcal{O}(\|x\|^\nu_p)$ are boundedness conditions for the second and third terms of (6.13). Finally, the last term contains an x-dependent induced p-norm applied to an $m \times m$ matrix with homogeneous entries of even degree equal to $\beta + 2\gamma - 2$. Thereby, we must have $\beta + \gamma - 1 \ge \beta + 2\gamma - 2$, i.e., $\gamma \le 1$, to obtain the desired result.

However, smoothness on $g(x)$ over $\mathbb{R}^n$ (specifically at $x = 0$) restricts the values of γ to be 0 or 1, and, further, $g(x)$ admits only two possibilities. Either $g(x) = B$, an $n \times m$ constant matrix, or $g_j(x) = D_j x$, where D_j are $n \times n$ constant matrices, $j = 1, \dots, m$. On the other hand, in $\mathcal{E}(\tau^\times)$, the right-hand side of (6.13) is obviously bounded. Therefore, the right-hand side of (6.13) is globally bounded, with bound depending on the associated order-constants, p norm parameter, control bound $\mathbf{r}$, parameter value $\tau^\times > 0$, and tuning parameter $\varepsilon > 0$. □

Remark 8. From (6.13), it should be noted that if $\varsigma = 1$ (e.g., as in the case of bilinear systems), then control $u_\theta(x)$ can be used instead of $u_\theta^\times(x)$.

The class of homogeneous systems considered above includes those systems with $g(x) = B$ that can be globally asymptotically stabilized using a quadratic Lyapunov function (linear systems included) and homogeneous bilinear systems. In particular, for linear and bilinear systems, since their free dynamics are linear (i.e., $f(x) = Ax$), the result on inputs subject to global bounded rate is guaranteed on the basis of the above theorem. This fact will be explicitly shown below in Propositions 6.3 and 6.5.

6.1. Globally asymptotically stabilizable systems by linear feedback. Assume that the $n \times m$ constant matrix B has full rank, and consider the class of affine systems

$$\dot{x} = f(x) + Bu. \tag{6.14}$$

The proposed method is particularly well suited for any system (6.14) that can be globally asymptotically stabilized by means of linear feedback $v(x) = -(L_B V(x))^\top = -Kx$ and quadratic Lyapunov function $V(x) = \frac{1}{2}\, x^\top P x$, where P is an $n \times n$ positive-definite symmetric matrix. For such a system the optimization problem is easily implemented, and assuming that Hypotheses H1–H2 hold, a globally stabilizing BFC function is designed. Sufficient conditions for the existence of linear controllers are available in the literature (cf. [2, 17]).

As has been pointed out above, the control constraint problem $\|u(x)\|_{p,\mathbf{r}} \leq 1$ can be solved by means of the Lagrange multiplier method. In the present case, it reduces to finding a function $c = c(\tau)$ as solution of the ellipsoidal boundary (a relation)

$$\partial\mathcal{E}(\tau) = \left\{ x \in \mathbb{R}^n : V(x) = \frac{1}{2}\, x^\top P x = c(\tau) \right\}, \tag{6.15}$$

where $P = P^\top > 0$, s.t. $\|u(x)\|_{p,\mathbf{r}} \leq 1$.

PROPOSITION 6.3. *Consider the affine system* (6.14) *for which Hypothesis* H1 *holds with a quadratic Lyapunov function $V(x)$ given in* (6.15) *and further assume, without loss of generality, that B has full rank. Furthermore, if Hypothesis* H2 *holds, then control $u_\theta^\times(x)$ given in* (3.19)*, where $v_\varepsilon(x)$ is given by*

$$v_\varepsilon(x) = -\frac{1}{\varepsilon + \sigma\sqrt{x^\top P x}}\, R\, B^\top P x, \tag{6.16}$$

with $R = \operatorname{diag}(r_1, \dots, r_m)$, satisfies $\|u_\theta^\times(x)\|_{p,\mathbf{r}} < 1$ for any $\tau^\times > 0$ and $\varepsilon > 0$ and globally asymptotically stabilizes the closed-loop system (6.14)–(3.19)*, where $\sigma = 2^{1/2}/\|(x^{*\top} PB)\ (B^\top PB)^{-1}\|_{p,\mathbf{r}}$ and x^* is an optimal point satisfying $x^{*\top} P x^* = 1$. Furthermore, if $f(x)$ is a globally Lipschitz function at the origin (i.e., there exists $L > 0$ such that for all $x \in \mathbb{R}^n$, $\|f(x)\|_2 \leq L\|x\|_2$), then given any $\kappa = (\kappa_1, \dots, \kappa_m)^\top \in \mathbb{R}^m_{>0}$, there is a $q \in \mathbb{Q}^*_{>1}$ (depending on κ), such that $\|du_\theta^\times/dt\|_{p,\kappa} < 1$, with $p \in \mathbb{Q}^*_{>1}$,* *for all $x \in \mathbb{R}^n$, where q is defined in terms of matrices B and P, Lipschitz constant L, rate bound κ, control bound* $\mathbf{r}$, *and parameter value $\tau^\times$.*

Proof. First of all, since $g(x) = B$ has full rank, it follows that $B^\top PB$ is nonsingular (indeed, it is a positive-definite matrix). Moreover, any extremum x^* is a regular point of the constraint ($\nabla_x V(x) \neq 0$ for all $x \neq 0$). In view that $\beta = 2$ and $\nu = 1$, we have $\varsigma = 2$. Thus, $\tau(x)$ is given by (6.1), that is, $\tau(x) = \sigma\sqrt{x^\top Px}$, so that it is admissible. Then, based on Theorem 6.2, the global stabilization result follows.

On the other hand, from $\upsilon_\varepsilon(x)$ (6.16) and some algebraic calculations, we obtain $\frac{d\upsilon_\varepsilon}{dt} = \frac{1}{\varepsilon+\tau}\big(\frac{\sigma^2}{2(\varepsilon+\tau)\,\tau}(x^\top Pf(x) + x^\top PB\,u_\theta)RB^\top Px - RB^\top Pf(x) + R(B^\top PB)\,u_\theta\big)$.

Hence, working along the line of the proof of Theorem 6.2, we obtain

$$\left\|\frac{d\upsilon_\varepsilon}{dt}\right\|_p = \frac{\sigma^2}{2\tau}\left(\left\|R^{-1}\upsilon_\varepsilon\right\|_{q,1/\mathbf{r}}\|u_\theta\|_{p,\mathbf{r}} - \left(\frac{1}{\varepsilon+\tau}\right)x^\top P\,f(x)\right)\|\upsilon_\varepsilon\|_p$$

$$-\frac{1}{\varepsilon+\tau}\big(\left\|RB^\top P\,f(x)\right\|_p + \left\|RB^\top PB\right\|_p\|u_\theta\|_p\big).$$

From the fact that $|\upsilon_{\varepsilon_j}| < 1$, $j = 1, \ldots, m$, we have that $\|\upsilon_\varepsilon(x)\|_p < m^{1/p}$ for any p-norm. Moreover, recalling that $\|u_\theta\|_{p,\mathbf{r}} < 1$, then $\|u_\theta\|_p = \varrho\|u_\theta\|_{p,\mathbf{r}} < \varrho$, with ϱ depending on $\mathbf{r}$. Then, from the above expression, we have

$$\left\|\tfrac{d\upsilon_\varepsilon}{dt}\right\|_p \leq \tfrac{\sigma^2 m}{2\tau}\underbrace{-\frac{\sigma^2 m^{1/p}}{2\tau(\varepsilon+\tau)}(x^\top Pf(x))}_{(a)} + \underbrace{\frac{\left\|RB^\top Pf(x)\right\|_p}{(\varepsilon+\tau)}}_{(b)} + \tfrac{\varrho\left\|RB^\top PB\right\|_p}{(\varepsilon+\tau)}. \tag{6.17}$$

The expressions (a) and (b) from (6.17) admit the following bounds.

For (a) and recalling that for all $x \in \mathbb{R}^n$, $\|f(x)\|_2 \leq L\|x\|_2$, we have

$$-\frac{\sigma^2\,m^{1/p}}{2\tau(\varepsilon+\tau)}\big(x^\top P\,f(x)\big) < \frac{m^{1/p}}{2(x^\top Px)}\,\|Px\|_2\,\|f(x)\|_2 \leq \frac{L\,m^{1/p}}{2\left\|P^{1/2}x\right\|_2^2}\,\|Px\|_2\,\|x\|_2$$

$$\leq \frac{L\,m^{1/p}}{2}\left\|P^{1/2}\right\|_2\left\|P^{-1/2}\right\|_2;$$

whereas, for (b), we obtain $\frac{1}{(\varepsilon+\tau)}\|RB^\top P\,f(x)\|_p < \frac{1}{\sigma\|P^{1/2}x\|_2}\|RB^\top P\|_p\|f(x)\|_p \leq \frac{L\,\mu}{\sigma\,\lambda_{\min}(P^{1/2})}\|B^\top P\|_{p,1/\mathbf{r}}$, where $\mu > 0$ is an appropriate constant (entailed from the equivalence of norms).

Then, from the previous expressions and (6.17) and after some algebraic calculations, we finally obtain

$$\left\|\frac{d\upsilon_\varepsilon^\times}{dt}\right\|_p \leq \flat := L\left(\tfrac{m^{1/p}\lambda_{\max}(P^{1/2})}{2\lambda_{\min}(P^{1/2})} + \tfrac{\mu\,\|B^\top P\|_{p,1/\mathbf{r}}}{\sigma\lambda_{\min}(P^{1/2})}\right) + \left(\tfrac{\sigma^2 m}{2\tau^\times} + \tfrac{\varrho\left\|B^\top PB\right\|_{p,1/\mathbf{r}}}{\varepsilon+\tau^\times}\right) \tag{6.18}$$

for all $x \in \mathbb{R}^n$. □

Remark 9. The bound $\flat$ (right-hand side of (6.18)) admits a geometrical meaning in terms of the curvature of the ellipsoid $\mathcal{E}(\tau)$ (6.15). In fact, from (6.18) observe that

$$\lambda_{\max}(P^{1/2})/\lambda_{\min}(P^{1/2}) = \chi_{\max}/\chi_{\min}, \tag{6.19}$$

where $\chi_{\min}$ and $\chi_{\max}$ are the minimal and the maximal lengths of the semiaxes of the ellipsoid $\mathcal{E}(\tau)\;|_{c=1/2}$ (6.15), respectively. Hence, the more that ellipsoid resembles a ball, the smaller the attained bound value $\flat$ becomes.

Consider the particular case when system (6.14) is linear, i.e., $f(x) = Ax$, where A is an $n \times n$ matrix satisfying Hypothesis H1 and B is an $n \times m$ matrix. In this case, Hypothesis H1 can be rephrased in terms of quadratic Lyapunov functions.

HYPOTHESIS H1′. *Suppose there exists a positive-definite symmetric matrix* P $(P = P^\top > 0)$, *solution of the Lyapunov equation*

$$A^\top P + PA = -Q, \tag{6.20}$$

where Q *is a positive semidefinite matrix* $(Q \geq 0)$. *Then, the quadratic function*

$$V(x) = \frac{1}{2}\, x^\top P x \tag{6.21}$$

defines a Lyapunov function for the open-loop system (6.14).

From Theorem 6.2, if, in addition, Hypothesis H2 holds, GAS of that system with bounded control is guaranteed.

COROLLARY 6.4. *Consider the linear system case of* (6.14) *for which Hypotheses* H1′–H2 *hold and further assume, without loss of generality, that* B *has full rank. Then, the control function* $u_\theta^\times(x)$ (3.19), *where* $\upsilon_\varepsilon(x)$, *given in* (6.16), *satisfies* $\|u_\theta^\times(x)\|_{p,\mathbf{r}} < 1$ *and renders the GAS of the closed-loop system* (6.14)–(3.19). *Further, given* $\kappa \in \mathbb{R}^m_{>0}$, *there is* $q \in \mathbb{Q}^*_{>1}$ *such that* $\|du_\theta^\times/dt\|_{p,\kappa} < 1$ *for all* $x \in \mathbb{R}^n$.

Remark 10. Another approach to the problem presented in this subsection can be found in [23]. In that paper, the control-value set was given by an S-Euclidean normed ball, $\mathcal{B}^m_S = \{u \in \mathbb{R}^m : \|u\|_S := \sqrt{u^\top S\, u} \leq 1\}$, where S is an $m \times m$ positive-definite symmetric matrix. Additional features developed in [23] were (i) addressing globally bounded rates on inputs: $\|du/dt\|_S \leq \kappa$, and (ii) the semiglobal stabilization of ANCBC linear systems via feedback control laws with magnitude and rate bounds.

6.2. Homogeneous bilinear systems. Consider the class of homogeneous bilinear systems

$$\dot{x} = Ax + \sum_{j=1}^{m} u_j D_j x, \tag{6.22}$$

where A is an $n \times n$ matrix satisfying Hypothesis H1′ and D_j are $n \times n$ constant matrices, $j = 1, \ldots, m$. In this case, from Remark 8, we have $\varsigma = 1$ $(= \beta/\nu = 2/2)$, so that $u_\theta(x)$, given in (3.11), will be used instead of $u_\theta^\times(x)$ given in (3.19).

Based on Theorem 6.2, if a quadratic Lyapunov function $V(x)$, given by (6.21), exists, then a globally bounded smooth stabilizer is designed for system (6.22).

PROPOSITION 6.5. *Consider the homogeneous bilinear system* (6.22) *for which Hypothesis* H1′ *holds with a quadratic Lyapunov function* $V(x)$ *given in* (6.21). *If Hypothesis* H2 *also holds, then control* $u_\theta(x)$, *given in* (3.11), *where* $\upsilon_\varepsilon = (\upsilon_{\varepsilon_1}, \ldots, \upsilon_{\varepsilon_m})^\top$ *is given by*

$$\upsilon_{\varepsilon_j}(x) = -\frac{r_j}{2\,(\varepsilon + \sigma\, x^\top P x)}\, x^\top (PD_j + D_j^\top P)x, \quad j = 1, \ldots, m, \tag{6.23}$$

satisfies $\|u_\theta(x)\|_{p,\mathbf{r}} < 1$ *for arbitrary* $\varepsilon > 0$, *and globally asymptotically stabilizes the closed-loop system* (6.22)–(3.11), *where* $\sigma = \|h^\top(x^*)\,(Dh(x^*)\,g(x^*))^{-1}\|^{-1}_{p,\mathbf{r}}$, x^* *is an* *optimal point such that* $x^{*\top} P x^* = 2$ *and the set* $\{D_1 x^*, \ldots, D_m x^*\}$ *is linearly independent. Moreover, given any* $\kappa \in \mathbb{R}^m_{>0}$*, there is a* $q \in \mathbb{Q}^*_{>1}$ *such that* $\|du_\theta/dt\|_{p,\kappa} < 1$ *for all* $x \in \mathbb{R}^n$*.*

Proof. First of all, in view that the set $\{D_1 x^*, \ldots, D_m x^*\}$ is linearly independent, it follows that $Dh_j(x^*)\, g_j(x^*) = x^{*\top}(D_j^\top P D_j + P D_j^\top D_j) x^*$ for $j = 1, \ldots, m$, are also linearly independent (so that $Dh(x^*)\, g(x^*)$ is nonsingular). Moreover, since $\nabla_x V(x) \neq 0$ for all $x \neq 0$, any extremum x^* is a regular point of the constraint. Set $\varsigma = 1$, since $\beta = 2$ and $\nu = 2$. Then, based on Theorem 6.1, $\tau(x) = \sigma\, x^\top P x$ and it is (obviously) admissible, and from Theorem 6.2, the global stabilization result follows.

On the other hand, due to the fact that $|\upsilon_{\varepsilon_j}| < 1$, $j = 1, \ldots, m$, it follows that $\|\upsilon_\varepsilon\|_p < m^{1/p}$ for any p-norm. Moreover, $\|u_\theta\|_p \leq \varrho\, \|u_\theta\|_{p,\mathbf{r}} < \varrho$, where ϱ depends on $\mathbf{r}$. Then, from $\upsilon_\varepsilon(x)$ (6.23) and (6.20), the corresponding expression (6.13) becomes

$$\left\| \frac{d\upsilon_\varepsilon}{dt} \right\|_p \leq \sigma\, m + \frac{1}{\varepsilon + \tau} \left(\frac{\sigma\, m^{1/p}}{2}\, (x^\top Q x) + \|Dh\, Ax\|_{p,1/\mathbf{r}} + \varrho\, \|Dh\, g\|_{p,1/\mathbf{r}} \right), \tag{6.24}$$

where $Dh = g^\top P + x^\top P D g^\top$. The result follows from Theorem 6.2 observing that the right-hand side of the above expression is bounded for all $x \in \mathbb{R}^n$ and $\varepsilon > 0$. □

For the remainder of this section, we will address the control design in $\mathcal{B}^m_{\mathbf{r}}(\infty)$. As pointed out in section 5.1, the solution to the associated optimization problem for the control design in $\mathcal{B}^m_{\mathbf{r}}(\infty)$ is achieved by solving an optimization problem with scalar control constraint (5.7). Thereby, in the case of homogeneous bilinear systems, the solution to (5.11) is reduced to finding a root of the polynomial

$$p(\lambda) = \det H_\lambda, \quad \text{where } H_\lambda := (PD + D^\top P) - \lambda P. \tag{6.25}$$

If $\lambda = \lambda^*$ is the required root, considering (5.11) and (5.7), it follows that $c(\tau) = \tau/|\lambda^*|$ and $\tau(x) = \frac{1}{2}\, |\lambda^*|\, x^\top P x$. Henceforth, the BFC function $u_\theta(x)$ is defined in (5.3), where the components $\overset{[j]}{\upsilon}_{\varepsilon_j}$ are given by

$$\overset{[j]}{\upsilon}_{\varepsilon_j}(x) = -\frac{r_j}{\varepsilon_j + |\lambda_j^*|\, (x^\top P x)}\, x^\top (P D_j + D_j^\top P) x, \quad j = 1, \ldots, m. \tag{6.26}$$

The following theorem leads to the solution of convex and nonconvex optimization problems (5.7) associated to the BFC problem for homogeneous bilinear systems with *scalar control*. Depending on the definiteness of the function ψ (5.8), the sufficient conditions are expressed in terms of λ^* and the remaining roots of the polynomial $p(\lambda)$ (6.25). It should be noted that, from homogeneity, if $\psi(c) > 0$ (or < 0) for some $c > 0$, then ψ is positive- (or negative-) definite.

THEOREM 6.6. *Assume that Hypotheses* H1′–H2 *hold. Set* $\lambda^* = \max_\lambda \arg\{p(\lambda) = 0\}$*. Then:* (i) *if the optimization problem* (5.9) *holds,* λ^* *is the unique positive root of the polynomial* $p(\lambda)$*; or* (ii) *if the negative version of* (5.9) *holds, then all roots of the polynomial* $p(\lambda)$ *are negative. In either case, GAS of the system* (6.22) *is obtained by means of the control* $u_\theta(x)$ *given in* (5.3) *with components* $\overset{[j]}{\upsilon}_{\varepsilon_j}$ *given by* (6.26).

Proof. Case (i). As is well known [18], a (second-order) sufficient condition for x^* to be a maximum of the optimization problem (5.9) consists of

(a) the necessary condition: there exists $\lambda^* \in \mathbb{R}$ such that

$$x^{*\top} H_{\lambda^*} = x^{*\top}((PD + D^\top P) - \lambda^* P) = 0, \text{ and} \tag{6.27}$$

(b) the negative semidefiniteness of the Hessian matrix H_{λ^*} (*regular pencil* of matrices (6.25)) on

$$M = \left\{z \in \mathbb{R}^n : x^{*\top} P z = 0\right\} := \ker(x^{*\top} P). \tag{6.28}$$

Then, considering (6.27) and positive definiteness of $\tau(x)$ (5.9), it follows that $\lambda^* > 0$.

The following two results, specialized to the present problem, are required (cf. [9]).

(A) There exist n real characteristic values of H_λ, which can be assumed ordered: $\lambda_1 \le \cdots \le \lambda_n$. Moreover, for every k ($1 \le k \le n$), the kth characteristic value λ_k is given as the minimum ratio of the bilinear forms

$$\lambda_k = \min_x \frac{x^\top (PD + D^\top P) x}{x^\top P x}\,, \tag{6.29}$$

provided that the variable vector x is subject to $k-1$ linear relations: there exist $k-1$ (P-)orthonormal vectors z_i ($1 \le i \le k-1$), i.e., $z_i^\top P z_j = \delta_{ij}$ ($= \{1$, if $i = j$; or 0, if $i \ne j\}$), which are (P-)orthogonal to x, that is

$$x^\top P z_1 = 0,\ x^\top P z_2 = 0, \ldots, x^\top P z_{k-1} = 0. \tag{6.30}$$

(B) There is a nonsingular transformation, $x = T\xi$, which reduces the bilinear forms $x^\top(PD + D^\top P)x$ and $x^\top P x$ simultaneously to a sum of squares

$$\xi^\top \operatorname{diag}(\lambda_1, \ldots, \lambda_n)\xi \quad \text{and} \quad \xi^\top \xi. \tag{6.31}$$

Then, λ is a characteristic value of the pencil H_λ iff it is a root of the polynomial $p(\lambda) = \det H_\lambda$.

We claim that the negative semidefiniteness sufficient condition on H_{λ^*} holds if λ^* is the unique positive root of $p(\lambda)$. In fact, since $\dim M = \dim \ker(x^{*\top} P) = n-1$, from (A), it follows that $k = n$. Based on this, taking into account (B) and recalling that, by hypothesis, $\lambda_1, \ldots, \lambda_{n-1}$ are nonpositive roots of $p(\lambda)$, we obtain that matrix H_{λ^*} (6.25) is negative semidefinite on M. Moreover, $\lambda^* = \lambda_n > 0$—the Lagrange multiplier associated to x^*—is a root of the polynomial $p(\lambda)$. Therefore, $\lambda_1 \le \cdots \le \lambda_{n-1} \le 0$ and $\lambda_n > 0$ are all the characteristic values of H_λ.

Case (ii). If ψ (5.8) is nonpositive, the negative version of (5.9) holds. Then we obtain the following necessary condition for an extremum:

$$x^{*\top} H_{\nu^*} = 0, \tag{6.32}$$

where $H_{\nu^*} = (PD + D^\top P) + \nu^* P$. From this and positive definiteness of $\tau(x)$, it follows that $\nu^* > 0$. Denoting $\lambda^* := -\nu^*$, we obtain that $\lambda^* < 0$. Then, considering $p(\lambda) = \det H_\lambda$ (6.25) and reasoning analogously to case (i) yields that all roots of $p(\lambda)$ must be negative, λ^* being the greater one. □

7. Concluding remarks. In this work, we addressed the GAS problem of a class of nonlinear systems with stable free dynamics subject to both input magnitude and rate bounds. In general, in order to derive the bounded stabilizer, the resulting procedure implies that "gains," as state-functions, are the solution to a c-parameterized nonlinear programming. On broad lines, the procedure consists of defining a function $\tau(x)$, such that it is constant along the boundary of the c-level (compact) sets of certain functions. Then, taking a sequence of those c-level sets in decreasing size (converging to the origin), it is assigned to each set the corresponding possible highest "gain" (via τ), while keeping the control bounded. For general nonlinear systems, the resulting closed-loop system could be implicitly defined in the sense that the control law is obtained from the solution to a nonlinear algebraic equation. Furthermore, we show that the proposed control is suboptimal. For an important class of homogeneous nonlinear systems (including a class of GAS systems by linear feedback and bilinear systems), it is shown that the resulting programming problem can be explicitly solved, and, further, the problem of inputs subject to global bounded rates is addressed. In many applications, the control-value set is defined as the m-dimensional $\mathbf{r}$-hyperbox $\mathcal{B}^m_{\mathbf{r}}(\infty) = \{u \in \mathbb{R}^m : |u_j| \leq r_j,\, r_j > 0,\, j = 1, \ldots, m\}$. Thereby, we presented a design approach for control laws considering more general control-value sets: Cartesian products of $p, \mathbf{r}$-normed balls.

Acknowledgment. The authors are grateful to Dr. Carlos Ibarra for his helpful suggestions that improved the original version of this paper.

REFERENCES

[1] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[2] A. Bacciotti, P. Boieri, and L. Mazzi, *Linear stabilization of nonlinear cascade systems*, Math. Control Signals Systems, 6 (1993), pp. 146–165.

[3] D. S. Bernstein, *Nonquadratic cost and nonlinear feedback control*, Internat. J. Robust Nonlinear Control, 3 (1993), pp. 211–229.

[4] Ch. I. Byrnes, A. Isidori, and J. C. Willems, *Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1228–1240.

[5] J.-M. Coron, *Linearized control systems and applications to smooth stabilization,* SIAM J. Control Optim., 32 (1994), pp. 358–386.

[6] R. Freeman and L. Praly, *Integrator backstepping for bounded controls and control rates*, IEEE Trans. Automat. Control, 43 (1998), pp. 258–262.

[7] O. Fujiwara, *Morse programs: A topological approach to smooth constrained optimization. I*, Math. Oper. Res., 7 (1982), pp. 602–616.

[8] O. Fujiwara, *A note on differentiability of global optimal values*, Math. Oper. Res., 10 (1985), pp. 612–618.

[9] F. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.

[10] V. Gavrilyako, V. Korobov, and G. Sklyar, *Designing a bounded control for dynamic systems in entire space with the aid of a controllability function*, Automat. Remote Control, 11 (1986), pp. 1484–1490.

[11] D. H. Jacobson, *Extensions of Linear-Quadratic Control, Optimization and Matrix Theory*, Academic Press, London, New York, 1977.

[12] H. Th. Jongen, P. Jonker, and F. Twilt, *One-parameter families of optimization problems: Equality constraints*, J. Optim. Theory Appl., 48 (1986), pp. 141–161.

[13] V. Jurdjevic and J. Quinn, *Controllability and stability*, J. Differential Equations, 28 (1978), pp. 381–389.

[14] W. Lin, *Global asymptotic stabilization of general nonlinear systems with stable free dynamics via passivity and bounded feedback*, Automatica J. IFAC, 32 (1996), pp. 915–924.

[15] Y. Lin and E. D. Sontag, *A universal formula for stabilization with bounded controls*, Systems Control Lett., 16 (1991), pp. 393–397.

[16] Z. Lin, *Semi-global stabilization of linear systems with position and rate limited actuators*, Systems Control Lett., 30 (1997), pp. 1–11.

[17] Z. Lin and A. Saberi, *Semi-global stabilization of minimum phase nonlinear systems in special normal form via linear high-low-gain state feedback*, Internat. J. Robust Nonlinear Control, 4 (1996), pp. 353–362.

[18] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1984.

[19] M. Malisoff and E. D. Sontag, *Universal formulas for CLF's with respect to Minkowski balls*, in Proceedings of the American Automatic Control Council's Amer. Control Conference, 1999, San Diego CA, pp. 1598–1602.

[20] F. Mazenc and L. Praly, *Adding integrators, saturated controls, and stabilization for feedforward systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1559–1578.

[21] Z. Qu, *Robust Control of Nonlinear Uncertain Systems*, Wiley, New York, 1998.

[22] J. Shewchun and E. Feron, *High performance control with position and rate limited actuators*, Internat. J. Robust Nonlinear Control, 9 (1999), pp. 617–630.

[23] J. Solís-Daun, J. Álvarez-Ramírez, and R. Suárez, *Semiglobal stabilization of linear systems: A parametric optimization approach*, Internat. J. Robust Nonlinear Control, 9 (1999), pp. 461–484.

[24] E. D. Sontag, *An algebraic approach to bounded controllability of linear systems*, Internat. J. Control, 39 (1984), pp. 181–188.

[25] E. D. Sontag, *Control-Lyapunov functions*, in Open Problems in Mathematical Systems and Control Theory, V. Blondel et al., eds., Springer-Verlag, London, UK, 1999, pp. 211–216.

[26] R. Suárez, J. Álvarez-Ramírez, and J. Solís-Daun, *Linear systems with bounded inputs: Global stabilization with eigenvalue placement*, Internat. J. Robust Nonlinear Control, 7 (1997), pp. 835–845.

[27] H. J. Sussmann, E. D. Sontag, and Y. Yang, *A general result on the stabilization of linear systems using bounded control*, IEEE Trans. Automat. Control, 39 (1994), pp. 2411–2424.

[28] A. R. Teel, *A nonlinear small gain theorem for the analysis of control systems with saturation*, IEEE Trans. Automat. Control, 41 (1996), pp. 1256–1270.

[29] Q. Zhou, *A transversality theorem for maps*, J. Math. Res. Exposition, 3 (1983), pp. 17–20 (in Chinese).

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 707–727

NONLINEAR ADAPTIVE TRACKING USING KERNEL ESTIMATORS: ESTIMATION AND TEST FOR LINEARITY*

JEAN-MICHEL POGGI† AND BRUNO PORTIER†

Abstract. We present some statistical results on nonlinear adaptive control using kernel estimators. We are concerned with a nonlinear autoregressive model of the form

$$X_{n+1} = f(X_n) + U_n + \xi_{n+1}, \qquad n \in \mathbb{N},$$

controlled using a nonparametric estimator of the unknown function f and derived from a tracking control policy. We prove an almost sure convergence result for the noise density estimator, a pointwise central limit theorem for f, and a test for linearity of the driving function f.

Key words. adaptive control, central limit theorem, discrete-time stochastic nonlinear system, kernel estimation, test for linearity

AMS subject classifications. 93C40, 62G07, 62G10, 62G20, 93C10, 93C55

PII. S0363012998349613

1. Introduction.

In a previous paper [14], the problem of adaptive control of nonlinear systems of the form

$$X_{n+1} = f(X_n) + U_n + \xi_{n+1}, \qquad n \in \mathbb{N}, \tag{1.1}$$

is considered, where X_n, U_n, and ξ_n are the output, input, and noise of the system, respectively. State X_n is observed, ξ_n is an unobservable noise, and control U_n is to be chosen. Initial conditions X_0 and U_0 are arbitrarily chosen. The main results of [14] deal with the almost sure convergence over dilating sets of a nonparametric estimator (based on a kernel method) of the unknown function f as well as the construction of an optimal adaptive tracking control law.

In this paper, we propose some new statistical results in this context. First, we propose a kernel method-based estimator of the probability density function (p.d.f.) of ξ_n, and we prove a uniform almost sure convergence result. Next, we establish a pointwise central limit theorem for the kernel estimator of function f as well as the associated joint asymptotic normality. Finally, we derive a test for linearity of f which can be used for identification purposes.

From a statistical viewpoint, the estimation of the regression function using the kernel method is widely investigated in the uncontrolled case ($U_n \equiv 0$ in model (1.1)). See Härdle [10] for a comprehensive survey, Robinson [18], Truong and Stone [26] under mixing assumptions (see Doukhan [4] for a full study of mixing notions), and Duflo [5] and Yakowitz [27] under Markov chains framework. Nonparametric statistical methods have been already used for the identification of nonlinear dynamical systems. See Greblicki and Pawlak [8], Greblicki [7], Krzyzak [13], Georgiev [6], and Hilgert, Senoussi, and Vila [11].

Let us now make some remarks about the existing literature on the different topics related to the main results of the paper.

*Received by the editors December 9, 1998; accepted for publication (in revised form) May 9, 2000; published electronically October 4, 2000.
http://www.siam.org/journals/sicon/39-3/34961.html

†Laboratoire de Mathématique – U.M.R. C 8628, "Probabilités, Statistique et Modélisation," Université Paris-Sud, Bâtiment 425, 91405 Orsay cedex, France, and IUT Paris, Université René-Descartes, 143, av. de Versailles, 75016 Paris, France (jean-michel.poggi@math.u-psud.fr, bruno.portier@math.u-psud.fr).

Many authors have studied the pointwise asymptotic normality of the kernel estimator in uncontrolled frameworks. For the regression model, Schuster [20] establishes the joint asymptotic normality of $(\widehat{f}_n(x_1), \ldots, \widehat{f}_n(x_q))$, where $x_1, \ldots, x_q$ are q distinct points. In a mixing framework, Robinson [18] gives a multivariate central limit theorem. More recently, Roussas and Tran [19] give a central limit theorem for a recursive version of the kernel estimator of the regression function of an α-mixing process. For nonlinear autoregressive models, Duflo [5] obtains the same results.

Tests of linearity of such models have been developed by several authors in the uncontrolled case. See Tong [24], Hjellvik and Tjøstheim [12], Tjøstheim [23], and Poggi and Portier [15] for reviews. A recent paper of Poggi and Portier [16] applies the ideas previously developed for time series to control system area in a nonadaptive framework.

For a practical use of the test, the estimation of the p.d.f. of the unobservable noise ξ_n is required. Up to now, the available results deal with the uncontrolled case. For a regression model, Ahmad [1] proposes a kernel estimator of the p.d.f. of the residuals of the model. A result of almost sure convergence and a central limit theorem in the scalar case are established. For model (1.1) with $U_n \equiv 0$, Senoussi [21] builds a kernel estimator of the p.d.f. of ξ_n and proves the almost sure convergence on compact sets of its estimator.

The paper is organized as follows. In section 2, we present the model and the control law. Section 3 then develops two asymptotic results (the proofs are postponed to the appendices). These results are then used in section 4 for testing the linearity of the leading function f. Finally, section 5 contains some simulation experiments. Our simulations carried out for a simple model indicate that our asymptotic results give a good approximation for moderate sample sizes. This study by simulations concerns mainly three topics: the illustration of the behavior of the adaptive control, the estimation of the noise p.d.f., and the test for linearity of the function f.

2. Framework and assumptions. Let us now further describe the model assumptions and the control law. We assume the following properties on function f, noise ξ_n, and the initial conditions of model (1.1).

Assumption [A1].

(i) Function f is Lipschitz with Lipschitz constant r_f.

(ii) $\xi = (\xi_n)_{n \geq 1}$ is a sequence of independent and identically distributed random vectors with mean 0 and unknown invertible covariance matrix Γ.

(iii) ξ_n has a finite moment of order $m > 2$ and its distribution is absolutely continuous with respect to the Lebesgue measure. Its unknown p.d.f. denoted by p is supposed to be C^1-class; $p > 0$ and p and its gradient are bounded.

(iv) X_0 has a finite moment of order m.

In order to estimate unknown function f, we use a recursive version of the classical kernel estimator of the regression function. Then, for $x \in \mathbb{R}^d$, we estimate $f(x)$ by

$$\widehat{f}_n(x) = \frac{\sum_{i=0}^{n-1} i^{\alpha d} K\left(i^{\alpha}(X_i - x)\right)(X_{i+1} - U_i)}{\sum_{i=0}^{n-1} i^{\alpha d} K\left(i^{\alpha}(X_i - x)\right)}, \tag{2.1}$$

where K is a kernel and α is a real number in $]0,\ 1/d[$, called the bandwidth parameter. We shall call a kernel a Lipschitz positive bounded function $N : \mathbb{R}^d \to \mathbb{R}_+$ satisfying $\int N(t)\,dt = 1$ and $\int \|t\|\, N(t)\,dt < \infty$, where $\|.\|$ denotes the usual norm on $\mathbb{R}^d$.

Let us give some heuristics about the control law proposed in [14] and that we use in this paper to control model (1.1). Since the control law is designed in order to track a given reference trajectory denoted by $(X_n^*)_{n\geq 1}$, an "ideal" candidate for the control is

$$U_n = -f(X_n) + X_{n+1}^*. \tag{2.2}$$

Since f is unknown, $f(X_n)$ can be replaced by its estimate $\widehat{f}_n(X_n)$, leading to the self-tuning control, which is a second "ideal" candidate,

$$U_n = -\widehat{f}_n(X_n) + X_{n+1}^*. \tag{2.3}$$

The nonparametric estimation of function f in controlled models of the form (1.1) using a kernel method involves some kind of stability of the controlled process X_n. The tracking control law (2.3) uses a local estimation feedback and that creates new problems for the analysis of the closed-loop behavior. Portier and Oulidi [14] have investigated this question and have proposed to introduce some a priori knowledge on the function to be estimated. (This kind of approach has been first experimentally used in [17].) More precisely, this a priori knowledge about function f is modelled by a function $\widetilde{f}$, supposed to be known and which satisfies the following assumption.

Assumption [A2]. Function $\widetilde{f}$ is continuous and

$$\exists\ a_f \in \left[0,\ 1/2\right[\ \exists\ A_f \in \left]0,\ \infty\right[\ \forall\ x \in \mathbb{R}^d, \quad \left\| f(x) - \widetilde{f}(x) \right\| \leq a_f \left\| x \right\| + A_f.$$

Let $(X_n^*)_{n\geq 1}$ be a given bounded deterministic tracking trajectory such that $X_n^* \underset{n\to\infty}{\longrightarrow} x^*$ with $\|x^*\| < \infty$. The adaptive tracking control with a priori knowledge is at time n defined by

$$U_n = X_{n+1}^* - \widehat{f}_n(X_n)\mathbf{1}_{E_n}(X_n) - \widetilde{f}(X_n)\mathbf{1}_{\overline{E}_n}(X_n), \tag{2.4}$$

where $\overline{E}_n$ denotes the complementary set of E_n. The set E_n, defined by

$$E_n = \{x \in \mathbb{R}^d;\ \|\widehat{f}_n(x) - \widetilde{f}(x)\| \leq b_f \left\| x \right\| + B_f\} \tag{2.5}$$

with $b_f \in \left]a_f,\ 1-a_f\right[$ and $B_f \in \left]A_f,\ \infty\right[$, is introduced to ensure the closed-loop stability of the system. Function $\widetilde{f}$ allows us to compensate for the possible lack of observations which disrupts the local estimator $\widehat{f}_n$. Under this kind of hypothesis, both stability and convergence results are obtained and the optimality of the tracking is proved. To make the paper self-contained, let us recall some previous results of [14], useful for what follows and summarized in the following theorem.

THEOREM 2.1. *Assume that* [A1] *and* [A2] *hold.*

1. *Then, for any initial law (the law of X_0),*

$$\sup_{k\leq n} \|X_k\| = o(n^{1/m}) \ \textit{and}\ \sup_n \mathbb{E}\left[\|X_n\|^m\right] < \infty.$$

2. *Let $(v_n)_{n\geq 1}$ be a sequence of positive real numbers increasing to infinity such that $v_n = O\left(n^\nu\right)$ with $\nu > 0$. Then, for a compactly supported kernel K and for a bandwidth parameter $\alpha \in \left]0,\ 1/2d\right[$, we have for any $s \in \left]1/2 + \alpha d,\ 1\right[$,*

$$\sup_{\|x\|\leq v_n} \|\widehat{f}_n(x) - f(x)\| = o\left(\frac{n^{s-1}}{m_n}\right) + O\left(\frac{n^{-\alpha}}{m_n}\right) \quad \textit{almost surely (a.s.)}, \tag{2.6}$$

where $m_n = \inf\{p(z);\ \|z\| \leq v_n + R\}$ for some constant $R < \infty$.

In addition, if $m_n^{-1} = \inf\left(o\left(n^{\alpha}\right), O\left(n^{1-s}\right)\right)$*, then we have*

$$\sum_{k=0}^{n-1} \|\pi_k\|^2 = o(n) \quad a.s., \tag{2.7}$$

$$\widehat{\Gamma}_n = \frac{1}{n}\sum_{k=1}^{n} \left(X_k - X_k^*\right)\left(X_k - X_k^*\right)^T \xrightarrow[n\to\infty]{\text{a.s.}} \Gamma, \tag{2.8}$$

where $\pi_k = f(X_k) + U_k - X_{k+1}^* = (f - \widehat{f}_k)(X_k)\,\mathbf{1}_{E_k}(X_k) + (f - \widetilde{f})(X_k)\,\mathbf{1}_{\overline{E}_k}(X_k)$*.*

Result (2.7) is the key point used to establish convergence results of an estimator of the p.d.f. p as well as a pointwise central limit theorem for $\widehat{f}_n$. Let us also mention that the results, obtained in the next sections, hold for any control law $(U_n)_{n\geq 0}$ for which it is possible to prove that

$$\sum_{k=0}^{n-1} \left\|f(X_k) + U_k - X_{k+1}^*\right\|^2 = o(n) \quad \text{a.s.} \tag{2.9}$$

For more clarity, let us introduce the following hypothesis coming from part 2 of Theorem 2.1.

Assumption [A3]. There is a sequence of positive real numbers $(v_n)_{n\geq 1}$ increasing to infinity such that $v_n = O\left(n^{\nu}\right)$ with $\nu > 0$ and

$$\left(\inf\{p(z); \|z\| \leq v_n + R\}\right)^{-1} = \inf\{o\left(n^{\alpha}\right), O\left(n^{1-s}\right)\}$$

with $s \in]1/2 + \alpha d,\ 1[$, $\alpha \in]0,\ 1/2d[$ and for some constant $R < \infty$.

For example, when ξ_n is Gaussian (a widely used noise model), this assumption is fulfilled with $v_n = C\left(\log\log n\right)^{1/2}$, where C is any positive finite constant. In addition, taking $\alpha = 1/2(d+1)$, we obtain

$$\sup_{\|x\|\leq C(\text{loglog}\, n)^{1/2}} \left\|\widehat{f}_n(x) - f(x)\right\| = O\left(n^{-\lambda}\right) \quad \text{a.s.}$$

for any $\lambda \in]0,\ 1/2(d+1)[$.

3. Main results. The first result states both the almost sure and almost sure uniform convergence of the kernel estimator of the noise p.d.f. p. Let us remark that if the tracking objective is fulfilled, then $(X_n - X_n^*)$ is close in distribution to ξ_n, leading to a natural kernel estimator of the noise p.d.f. p denoted by $\widehat{p}_n$ and defined for all $e \in \mathbb{R}^d$ by

$$\widehat{p}_n(e) = \frac{1}{n}\sum_{i=1}^{n} i^{\beta d}\, G\left(i^{\beta}(X_i - X_i^* - e)\right), \tag{3.1}$$

where G is a kernel and β is a real number in $]0,\ 1/d[$.

THEOREM 3.1. *Assume that* [A1]–[A3] *hold.*

(1) *Then, for any* $e \in \mathbb{R}^d$, $\widehat{p}_n(e) \xrightarrow[n\to\infty]{\text{a.s.}} p(e)$.

(2) *Moreover, if* $G(x) = O(\|x\|^{-\delta})$ *with* $\delta > \frac{m\,\beta\,d}{\beta\,m+1}$*, then*

$$\sup_{e\in\mathbb{R}^d} |\widehat{p}_n(e) - p(e)| \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Proof. The proof is straightforward using Corollary A.2 (see Appendix A) with $\varphi = G$, $\lambda = \beta d$, and $\gamma = \beta$. □

Remark. For the ARX model of the form

$$X_{n+1} = \theta^T \Psi_n + U_n + \varepsilon_{n+1}$$

with $\Psi_n = (X_n, \dots, X_{n-p+1})^T$ and $U_n = -\widehat{\theta}_n^T \Psi_n + X_{n+1}^*$, where $\widehat{\theta}_n$ is either the least or the weighted least square estimator of θ, Bercu [2] (see also Guo [9]) shows that (2.7) is fulfilled with $\pi_k = (\theta - \widehat{\theta}_k)^T \Psi_k$. Therefore, in this context, the previously defined $\widehat{p}_n$ is also a kernel estimator of the p.d.f. of ε_n and the results of Theorem 3.1 hold.

We present in the next theorem a pointwise central limit theorem and a result of joint asymptotic normality for two slightly different estimators of function f whose status will appear more clearly after the theorem. The first estimator is given by (2.1) where K is a compactly supported kernel and the second one, denoted by $\widehat{F}_n$, is given by

$$\widehat{F}_n(x) = \frac{\sum_{i=0}^{n-1} i^{\beta d} G\left(i^{\beta}(X_i - x)\right)(X_{i+1} - U_i)}{\sum_{i=0}^{n-1} i^{\beta d} G\left(i^{\beta}(X_i - x)\right)}, \tag{3.2}$$

where G is a kernel and β is a real number in $]0, 1/d[$.

THEOREM 3.2. *Assume that* [A1]–[A3] *hold, that function f is $\mathcal{C}^2$-class with bounded derivatives of order* 2 *and that kernels K and G satisfy $\int \|t\|^2 N(t)\,dt < \infty$ and for $j = 1, \dots, d$, $\int t_j\, N(t)\,dt = 0$, where t_j is the jth component of t and where N stands for K and for G.*

1) *For $\alpha \in\,]1/(d+4), 1/2d[$ and $x \in \mathbb{R}^d$,*

$$Z_n^{K,\alpha}(x) = n^{(1-\alpha d)/2}(\widehat{f}_n(x) - f(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0,\ \frac{\|K\|_2^2}{(1+\alpha\, d)p(x - x^*)}\,\Gamma\right) = Z^{K,\alpha}(x).$$

2) *For $\beta \in\,]1/(d+4), 1/d[$ and $x \in \mathbb{R}^d$,*

$$Z_n^{G,\beta}(x) = n^{(1-\beta d)/2}(\widehat{F}_n(x) - f(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0,\ \frac{\|G\|_2^2}{(1+\beta\, d)p(x - x^*)}\,\Gamma\right) = Z^{G,\beta}(x).$$

3) *Besides, for $x_1, \dots, x_q$, q distinct points of $\mathbb{R}^d$,*

$$(Z_n(x_1), \dots, Z_n(x_q)) \xrightarrow[n\to\infty]{\mathcal{L}} (Z(x_1), \dots, Z(x_q)),$$

where $Z(x_1), \dots, Z(x_q)$ are independent and where Z_n (respectively, Z) stands for $Z_n^{K,\alpha}$ or $Z_n^{G,\beta}$ (respectively, $Z^{K,\alpha}$ or $Z^{G,\beta}$), with the associated constraints on α and β.

Proof. The proof is given in Appendix B. □

Remark 3.3. In part 1 of Theorem 3.2, the range of admissible values for α comes from two sources. First, $\alpha \in\,]0, 1/2d[$ coming from Theorem 2.1 and ensuring that the key property $\sum_{k=0}^{n-1} \|\pi_k\|^2 = o(n)$ holds; second, $\alpha \in\,]1/(d+4), 1/d[$ coming from the proof of the central limit theorem (see Appendix B). So, these conditions on α imply that $d \leq 3$. To obtain a central limit theorem for an estimator of f when $d \geq 4$, we consider two different estimators of f. The first one is $\widehat{f}_n$, designed with respect to the control objective and defined using the compactly supported kernel K and the bandwidth parameter α within $]0,\, 1/2d[$. The second one is $\widehat{F}_n$, designed for statistical purposes. It is defined using the kernel G and the bandwidth parameter β within $]1/(d+4),\, 1/d[$. The parameters α and β have only to match their own constraints, leading to a general central limit theorem with respect to the state-space dimension (part 2 of Theorem 3.2).

4. Application to testing for linearity. In this section, we are concerned with a test for linearity of function f. Let us consider within model (1.1) the following hypotheses.

H_0 : $\ll f$ is linear of the form $f(x) = A_\ell^T\, x + B_\ell \gg$.
H_1 : $\ll f$ is nonlinear $\gg$.

The idea of the test statistic introduced by Poggi and Portier [15] in a slightly different framework is to compare two distinct estimators of function f. The first one is well suited under H_1 and the second one is convenient for the linear case. More precisely, this statistic of nonlinearity captures the quadratic deviations between the nonparametric estimator $\widehat{f}_n(x)$ and the least squares estimator $(\widehat{A}_n^T\, x + \widehat{B}_n)$ over q points, weighted by the estimated density at these points. The least squares estimators (LSE) $\widehat{A}_n$ and $\widehat{B}_n$ are defined by

$$\widehat{A}_n = S_{n-1}^{-1} \left(\sum_{k=0}^{n-1} X_k\, Y_k^T - \frac{1}{n} \left(\sum_{k=0}^{n-1} X_k \right) \left(\sum_{k=0}^{n-1} Y_k \right)^T \right), \tag{4.1}$$

$$S_{n-1} = \sum_{k=0}^{n-1} X_k\, X_k^T - \frac{1}{n} \left(\sum_{k=0}^{n-1} X_k \right) \left(\sum_{k=0}^{n-1} X_k \right)^T + \lambda\, I_d, \tag{4.2}$$

$$\widehat{B}_n = \frac{1}{n} \sum_{k=0}^{n-1} \left(Y_k - \widehat{A}_n^T\, X_k \right), \tag{4.3}$$

where $Y_k = X_{k+1} - U_k$, the real number λ is > 0, and I_d denotes the identity matrix of order d. The additional term $\lambda\, I_d$, which plays no role in the asymptotics, ensures that matrix S_n is always invertible.

We introduce the test statistic $T_q(n)$ defined by

$$T_q(n) = \frac{(1+\alpha d)\, n^{1-\alpha d}}{\|K\|_2^2} \sum_{j=1}^{q} \widehat{p}_n(x_j - x^*) \left\| \widehat{f}_n(x_j) - \widehat{A}_n^T\, x_j - \widehat{B}_n \right\|_{\widehat{\Gamma}_n^{-1}}^2, \tag{4.4}$$

where $x_1, \dots, x_q$ are q distinct points of $\mathbb{R}^d$, $\widehat{p}_n$ and $\widehat{\Gamma}_n$ are the estimators of p and Γ, respectively, and $\|y\|_{\widehat{\Gamma}_n^{-1}}^2 = y^T\, (\widehat{\Gamma}_n)^{-1}\, y$ for $y \in \mathbb{R}^d$.

For more clarity, we have used the nonparametric estimator $\widehat{f}_n$ in the construction of the test statistic $T_q(n)$. This choice leads to the constraint $d \leq 3$ (see Remark 3.3). However, the results on the test statistic (see Theorem 4.2) also hold true when we use $\widehat{F}_n(x)$ instead of $\widehat{f}_n(x)$, replacing K by G and α by β, with the associated constraints.

First let us state some convergence results for $\widehat{A}_n$ and $\widehat{B}_n$.

THEOREM 4.1. *Assume that* [A1]–[A3] *hold and assume that* $m > 4$.

(1) *Then,*

$$\widehat{A}_n \xrightarrow[n\to\infty]{\text{a.s.}} A = \Gamma^{-1}\, \mathbb{E}\left(\xi_1\, f(\xi_1 + x^*)^T \right),$$
$$\widehat{B}_n \xrightarrow[n\to\infty]{\text{a.s.}} B = \mathbb{E}\left(f(\xi_1 + x^*) \right) - A^T\, x^*.$$

(2) *Under* H_0*,* $A = A_\ell$*,* $B = B_\ell$*, and we have*

$$\left\|\widehat{A}_n - A\right\| = O\left(\frac{\log\log n}{n}\right)^{1/2} \quad a.s.,$$
$$\left\|\widehat{B}_n - B\right\| = O\left(\frac{\log\log n}{n}\right)^{1/2} \quad a.s.$$

Part 1 of Theorem 4.1 shows that the LSE always converge to finite limits even if the true model is nonlinear. Part 2 indicates that, when the true model is linear, despite the fact that the control law uses a nonparametric estimator, the convergence rates of the LSE are the same as those obtained with the standard linear state feedback control law with appropriate excitation (see Bercu [2] and Guo [9]).

Proof. (1) Using (1.1), let us rewrite (4.1) and (4.3) under the form

$$\begin{aligned}\frac{1}{n} S_{n-1}\widehat{A}_n = {} & \frac{1}{n}\sum_{k=0}^{n-1} X_k\, f(X_k)^T + \frac{1}{n}\sum_{k=0}^{n-1} X_k\, \xi_{k+1}^T \\ & - \left(\frac{1}{n}\sum_{k=0}^{n-1} X_k\right)\left(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) + \frac{1}{n}\sum_{k=0}^{n-1}\xi_{k+1}\right)^T,\end{aligned} \tag{4.5}$$

$$\widehat{B}_n = \frac{1}{n}\sum_{k=0}^{n-1} f(X_k) + \frac{1}{n}\sum_{k=0}^{n-1}\xi_{k+1} - \frac{1}{n}\sum_{k=0}^{n-1}\widehat{A}_n^T X_k. \tag{4.6}$$

First, let us study the convergence of S_n. Let $u \in \mathbb{R}^d$. Lemma A.3 gives

$$\frac{1}{n}\sum_{k=0}^{n-1} u^T X_k \xrightarrow[n\to\infty]{\text{a.s.}} u^T\mathbb{E}\left(\xi_1 + x^*\right) = u^T x^*, \tag{4.7}$$

and

$$\frac{1}{n}\sum_{k=0}^{n-1} u^T X_k\, X_k^T u \xrightarrow[n\to\infty]{\text{a.s.}} u^T\left(\Gamma + x^*(x^*)^T\right)u. \tag{4.8}$$

Therefore, combining these two results, we derive that

$$\frac{1}{n}\, u^T S_{n-1}\, u \xrightarrow[n\to\infty]{\text{a.s.}} u^T\Gamma\, u. \tag{4.9}$$

This result holds for any $u \in \mathbb{R}^d$. So we have, for any initial law,

$$\frac{1}{n}\, S_{n-1} \xrightarrow[n\to\infty]{\text{a.s.}} \Gamma, \tag{4.10}$$

and since Γ is supposed to be invertible,

$$\lim_{n\to\infty}\frac{\lambda_{\min}(S_{n-1})}{n} = \lambda_{\min}(\Gamma) > 0 \ \text{ a.s.}, \tag{4.11}$$

where $\lambda_{\min}(M)$ denotes the minimum eigenvalue of the matrix M.

Since $m > 2$, we have

$$\left\|\frac{1}{n}\sum_{k=0}^{n-1}\xi_{k+1}\right\| = O\left(\frac{\text{loglog } n}{n}\right)^{1/2} \quad \text{a.s.} \tag{4.12}$$

Since $\sup_n \mathbb{E}\left[\|X_n\|^m\right] < \infty$ (part 1 of Theorem 2.1), for any $n \geq 1$, $\sum_{k=0}^{n-1} X_k\, \xi_{k+1}^T$ is a square integrable martingale. In addition, since $\frac{1}{n}\sum_{k=0}^{n-1} X_k X_k^T \xrightarrow[n\to\infty]{\text{a.s.}} \Gamma + x^*(x^*)^T$ and $\sum_{n=1}^{\infty}(\frac{\|X_n\|}{\sqrt{n}})^\gamma < \infty$ for $\gamma \in]2,\, m]$, we obtain using a law of the iterated logarithm for the martingales (cf. Bercu [2] adapting an original result of Stout [22]; see also Touati [25]), that

$$\left\|\sum_{k=0}^{n-1} X_k\, \xi_{k+1}^T\right\| = O\left(n\ \text{loglog}\, n\right)^{1/2} \quad \text{a.s.} \tag{4.13}$$

Let $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$. Using once again Lemma A.3, we deduce that

$$\frac{1}{n}\sum_{k=0}^{n-1} u^T X_k\, f(X_k)^T v \xrightarrow[n\to\infty]{\text{a.s.}} u^T\mathbb{E}\left((\xi_1 + x^*)\, f(\xi_1 + x^*)^T\right) v, \tag{4.14}$$

$$\frac{1}{n}\sum_{k=0}^{n-1} u^T f(X_k) \xrightarrow[n\to\infty]{\text{a.s.}} u^T\mathbb{E}\left(f(\xi_1 + x^*)\right). \tag{4.15}$$

These results hold for all $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$. Hence, combining (4.10) and (4.12) to (4.15), we prove that $\widehat{A}_n \xrightarrow[n\to\infty]{\text{a.s.}} A$, and, therefore, $\widehat{B}_n \xrightarrow[n\to\infty]{\text{a.s.}} B$.

(2) Under H_0, we have, for any $x \in \mathbb{R}^d$, $f(x) = A^T x + B$, and thus, rewriting (4.5) and (4.6), we derive that

$$\begin{aligned}\|\widehat{A}_n - A\| = O\Bigg(\left(\frac{\lambda_{\min}(S_{n-1})}{n}\right)^{-1}\Bigg(\frac{\lambda}{n}\|A\| + \left\|\frac{1}{n}\sum_{k=0}^{n-1} X_k\, \xi_{k+1}^T\right\| \\ + \left\|\frac{1}{n}\sum_{k=0}^{n-1} X_k\right\| \left\|\frac{1}{n}\sum_{k=0}^{n-1}\xi_{k+1}\right\|\Bigg)\Bigg),\end{aligned} \tag{4.16}$$

$$\|\widehat{B}_n - B\| = O\left(\left\|\frac{1}{n}\sum_{k=0}^{n-1}\xi_{k+1}\right\| + \|\widehat{A}_n - A\|\left\|\frac{1}{n}\sum_{k=0}^{n-1} X_k\right\|\right). \tag{4.17}$$

Then, part 2 of Theorem 4.1 is easily obtained using (4.7) and (4.11)–(4.13). □

Now, collecting results of Theorems 3.1, 3.2, and 4.1, we derive the following theorem leading to an asymptotic test for linearity of f.

THEOREM 4.2. *Assume that hypotheses of Theorem* 3.2 *are fulfilled and that* $m > 4$. *Then, for* $\alpha \in]1/(d+4),\, 1/2d[$ *and for* q *distinct points* $x_1, x_2, \ldots, x_q$ *of* $\mathbb{R}^d$, *we have the following.*

(1) *Under* H_0, $T_q(n) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(dq)$.

(2) *Under* H_1 *and if there is* $x \in \{x_1, x_2, \ldots, x_q\}$ *such that* $f(x) \neq A^T x + B$, *then*

$$\liminf_{n\to\infty} n^{-1+\alpha d}\, T_q(n) > 0, \ \textit{a.s.}$$

Proof. For any $x \in \mathbb{R}^d$, we have the following decomposition:

$$\begin{aligned}\widehat{f}_n(x) - \widehat{A}_n^T x - \widehat{B}_n = (\widehat{f}_n(x) - f(x)) + \left(f(x) - A^T x - B\right) \\ -(\widehat{A}_n - A)^T x - (\widehat{B}_n - B).\end{aligned} \tag{4.18}$$

From Theorem 3.1, part 2 of Theorem 3.2, and the consistency of $\widehat{\Gamma}_n$, we show that

$$\frac{n^{(1-\alpha d)/2}\sqrt{1+\alpha d}}{\|K\|_2}\left(\sqrt{\widehat{p}_n(x_j-x^*)}\;\widehat{\Gamma}_n^{-1/2}(\widehat{f}_n(x_j)-f(x_j))\right)_{1\leq j\leq q}\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\,(0,\,I_{dq})\,.$$

Under H_0, $\|\widehat{A}_n-A\|+\|\widehat{B}_n-B\|\overset{\text{a.s.}}{=}O(\frac{\log\log n}{n})^{1/2}$ (see Theorem 4.1). Thus, it turns out that these linear terms are not contributing to the asymptotics. Therefore, we derive that under H_0,

$$\frac{n^{(1-\alpha d)/2}\sqrt{1+\alpha d}}{\|K\|_2}\left(\sqrt{\widehat{p}_n(x_j-x^*)}\,\widehat{\Gamma}_n^{-1/2}(\widehat{f}_n(x_j)-\widehat{A}_n^T\,x_j-\widehat{B}_n)\right)_{1\leq j\leq q}\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\,(0,\,I_{dq})\,,$$

and then taking the square norm of this vector, we obtain the $\chi^2(dq)$ limiting distribution.

In addition, using the results of pointwise almost sure convergence of $\widehat{p}_n$ and $\widehat{f}_n$, the consistency of the LSE $\widehat{A}_n$ and $\widehat{B}_n$, and the almost sure limit of $\widehat{\Gamma}_n$, we easily derive that

$$\frac{n^{\alpha d-1}\,\|K\|_2^2}{1+\alpha d}\,T_q(n)\xrightarrow[n\to\infty]{\text{a.s.}}\sum_{j=1}^{q}p(x_j-x^*)\,\big\|f(x_j)-A^T\,x_j-B\big\|_{\Gamma^{-1}}^2\,,$$

and under H_1, the limit is strictly positive if there is $x\in\{x_1,\cdots,x_q\}$ such that $f(x)\neq A^T\,x+B$, which closes the proof of part 2.

If $\widehat{F}_n(x)$ is used in the definition of the test statistic $T_q(n)$ instead of $\widehat{f}_n(x)$, we proceed similarly using Theorem 3.2 and Remark B.1 of Appendix B to justify the different needed convergences of $\widehat{F}_n(x)$. □

These asymptotic results make it possible to construct a test of linearity for function f. Part 1 of Theorem 4.2 gives the null distribution, and part 2 guarantees that the asymptotic power of the test is equal to 1 since the test statistic explodes a.s.

5. Simulation study. Since only asymptotic results are available, this section is devoted to the following three main topics: the illustration of the behavior of the adaptive tracking control, the estimation of the noise p.d.f. and the results of the test of linearity for moderate sample size realizations.

In addition, since the control strategy can be used to focus around a special state-space location or, equivalently, to explore different operating points of the process, we deeply examine the real-valued simulated nonlinear unstable open-loop model defined by

$$X_{n+1}=\big(1.4+0.5\sin(X_n/3)\exp\big(-(X_n-118)^2/50\big)\big)\,X_n+U_n+\xi_{n+1},$$

where $\xi_n\sim\mathcal{N}(0,\,(1.2)^2)$, $X_0=5$, and $U_0=0$.

Let us denote by f the function defined by

$$f(x)=\big(1.4+0.5\sin(x/3)\exp\big(-(x-118)^2/50\big)\big)\,x.$$

The graph of f is given by the dotted line on Figure 5.3. This function is linear for large x and highly nonlinear for small x, and hence fits the theory maximally well. It also shows how the test for linearity can work in spite of "almost linearity" or "almost nonlinearity."

For the estimation of f, we take the bandwidth parameter $\alpha = 1/2$ and we use the Gaussian kernel with the usual normalization σ_K equal to the empirical standard deviation based on the last observations $X_{n-51}, \ldots, X_n$. The tracking trajectory is defined as follows:

$$X_n^* = x^* - (x^* - X_0^*) \exp(-\tau\, n) \quad \text{with } \tau = -(1/100) * \log(0.05).$$

This kind of tracking trajectory is usual and is such that the deviation between X_n^* and x^* is of 5% when $n = 100$. We choose the control law

$$U_n = -\widehat{f}_n(X_n)\mathbf{1}_{E_n}(X_n) - \widetilde{f}(X_n)\mathbf{1}_{\overline{E}_n}(X_n) + X_{n+1}^*,$$

where $\widetilde{f}(x) = x$ and $E_n = \{x \in \mathbb{R}^d;\ \|\widehat{f}_n(x) - \widetilde{f}(x)\| \leq 0.4\|x\| + 300\}$.

The first topic is the behavior of the adaptive control law as well as the closed-loop tracking performance.

5.1. Study of the adaptive control law. We study two typical situations depending on the value of x^*. With the first value of $x^* = 113$ (see Figures 5.1 to 5.4), the controlled process mainly explores the nonlinear part of the function f. On the other hand, with the second one, $x^* = 170$ (see Figures 5.5 to 5.8), the linear part of function f captures the data essentially.

Let us examine the "nonlinear" situation. In Figure 5.1, we can see the controlled process X_n superimposed with the tracking trajectory. We distinguish two periods: the transient part on the time interval [0, 200] and the near stationary part after time 200. The transient part can be divided again in two distinct periods: the starting period on the time interval [0, 100] and the overshoot period [100, 200]. During the starting period the closed-loop system is close to an unstable AR(1) process with coefficient 1.4. Since X_n^* goes sufficiently slowly towards its limit, the value of X_{n+1} is always in a small neighborhood of the last state-space locations leading to a local nonparametric estimator of good quality. On the other hand, during the overshoot period, the open-loop system is highly unstable and X_{n+1} is often far from the previous observations of the process, and therefore the local nonparametric estimator explores the nonlinear domain of the function f with only a few observations leading to a crude estimation. Nevertheless, the number of observations around x^* increases, and, therefore, the estimation of the function f becomes better and better and U_n matches the control objective.

In Figure 5.2, we see that the control effort is moderate on the time interval [0, 100] since the open-loop system is close to a linear system that is easy to be controlled. The control effort is very high afterward, since the open-loop system is locally highly unstable leading to the control burden (large slope). Nevertheless, this behavior is expected.

On Figures 5.3 and 5.4, one can appreciate the quality of the functional estimation of f explaining the good quality of the tracking performance after the learning period ending at time 200. The lack of proper fit in Figure 5.3 is a consequence of "lack of excitation" which is well known in a control context when tracking is the goal (see [3]).

Let us now examine the same plots but for $x^* = 170$. In contrast to the previous situation, the control effort is more regular since the local slope of function f is always the same, except for the nonlinear part corresponding to the process values within the time interval [100, 130]. As can be seen in Figure 5.5 and 5.6, both overshoots

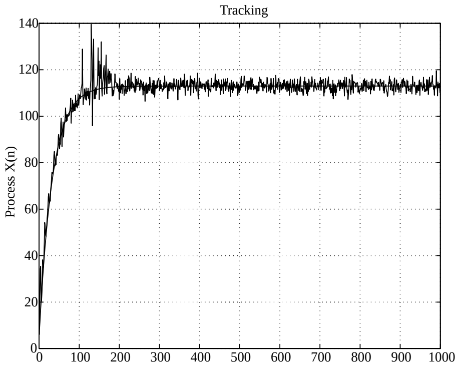


FIG. 5.1. *The process X_n superimposed with the tracking trajectory for $x^* = 113$.*

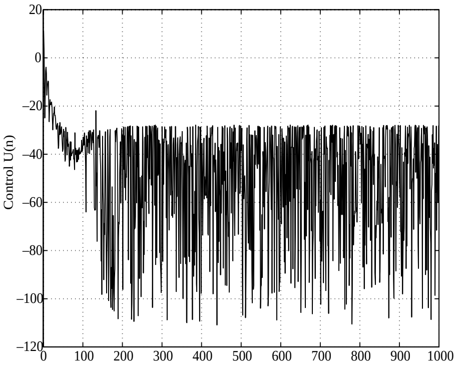


FIG. 5.2. *The corresponding adaptive control U_n.*

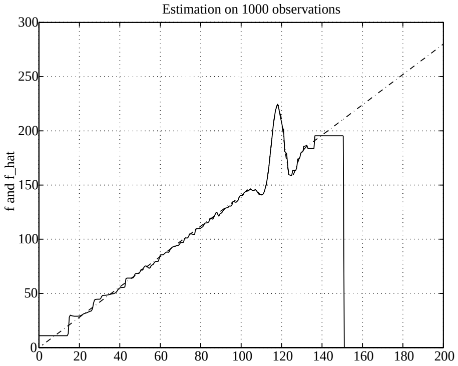


FIG. 5.3. *The true function (dashed line) superimposed with its nonparametric estimator. $x^* = 113$.*

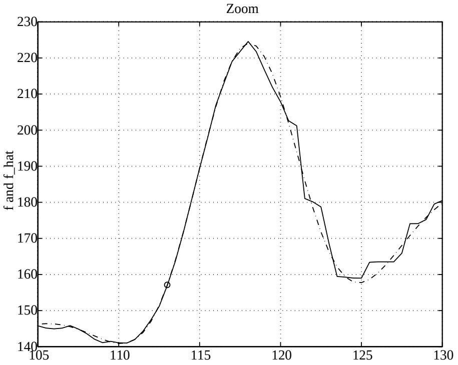


FIG. 5.4. *Zoom in Figure 5.3 showing the nonlinear part around the working point $x^* = 113$.*

for X_n and large values for U_n occur only when the data lie in this nonlinear part of function f. As in the above case, the nonparametric estimator behaves well.

A natural measure of the tracking performance is given by $\widehat{\Gamma}_n$. The result of Portier and Oulidi [14, Theorem 4.2], states that asymptotically, $\widehat{\Gamma}_n$ is equal to the noise variance Γ. Let us evaluate the quantity $(1/750)\sum_{k=251}^{1000}(X_k - X_k^*)^2$, which gives a good idea of Γ. For $x^* = 113$ and $x^* = 170$, we find 1.45 and 1.51, respectively, to be compared to the noise variance equal to 1.44.

Another way to illustrate the satisfactory behavior of the functional estimate of function f is to examine the estimation of the noise p.d.f. One can find in Figures 5.9 and 5.10 the estimates of $p(x - x^*)$ for $x^* = 113$ and $x^* = 170$, respectively, to compare with the p.d.f. of $\mathcal{N}(x^*, (1.2)^2)$. In both cases, these estimations are of good quality.

5.2. Test results for simulated data. The test defined in the previous section is global since H_0 assumes the function to be globally linear. Nevertheless, since n is finite and small, the noise ξ_n can be considered as bounded, and since the tracking objective is to put the process X_n near x^*, we can consider that we test the local linearity of the function f only on a neighborhood of x^* which contains almost all the values of the controlled process X_n. Then, we investigate various values of x^* in

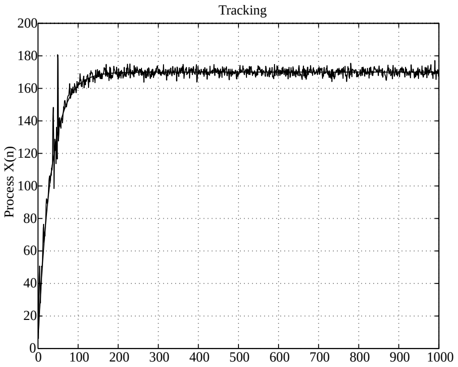


FIG. 5.5. *The process X_n superimposed with the tracking trajectory for $x^* = 170$.*

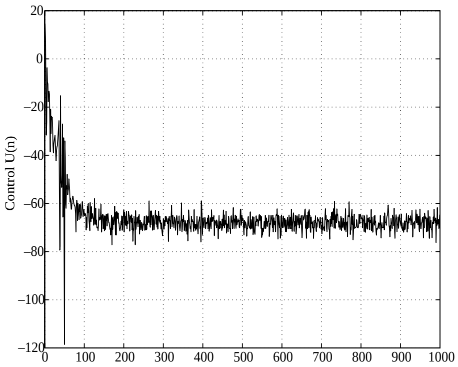


FIG. 5.6. *The corresponding adaptive control U_n.*

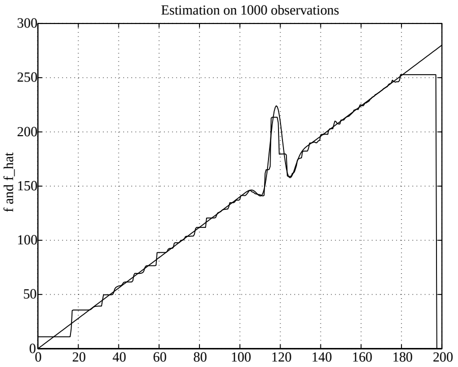


FIG. 5.7. *The true function (dashed line) superimposed with its nonparametric estimator. $x^* = 170$.*

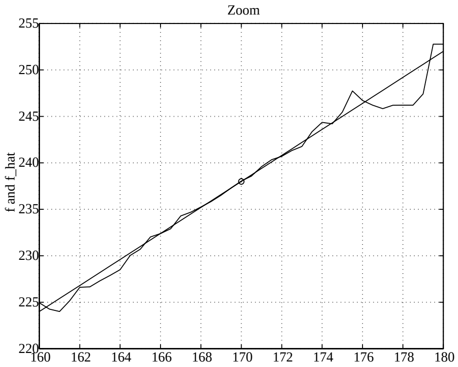


FIG. 5.8. *Zoom in Figure* 5.7 *showing the nonlinear part around the working point $x^* = 170$.*

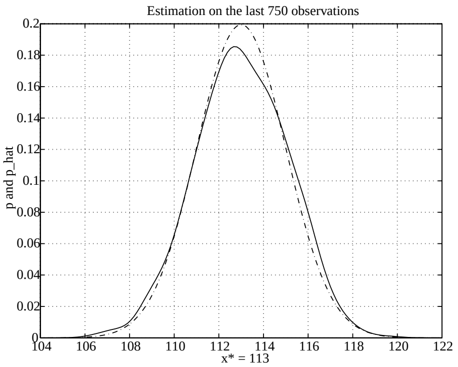


FIG. 5.9. *The true Gaussian density (dashed line) superimposed with its nonparametric estimator (dotted line). $x^* = 113$.*

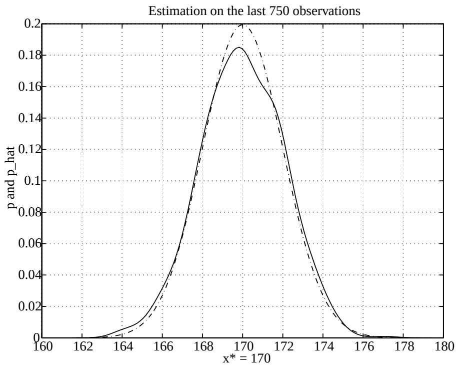


FIG. 5.10. *The true Gaussian density (dashed line) superimposed with its nonparametric estimator (dotted line). $x^* = 170$.*

order to illustrate the behavior of the test both under H_0 and under H_1. For $x^* = 70$ and $x^* = 170$, the actual system is close to a linear one, and for $x^* \in [100, 130]$, the actual system is nonlinear. One can find the selected working points in Figure 5.11.

Table 5.1
Test results under H_0. Level 5%. Percentage of rejections.

Working point	n	200	500	1000
$x^* = 70$	*Reject (%)*	2%	3.5%	3%
$t_{learn} = 200$	K.S	0.083	0.083	0.056
$x^* = 170$	*Reject (%)*	2%	4%	4%
$t_{learn} = 200$	K.S	0.087	0.086	0.088

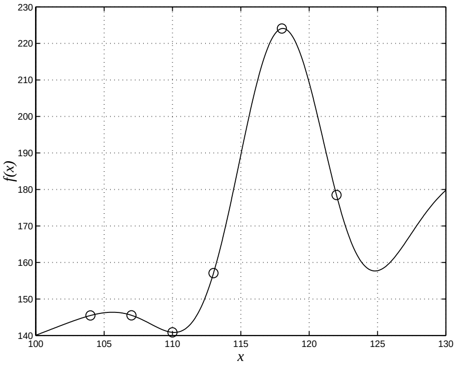


Fig. 5.11. *Function f around its nonlinear part and some working points of interest marked by a circle.*

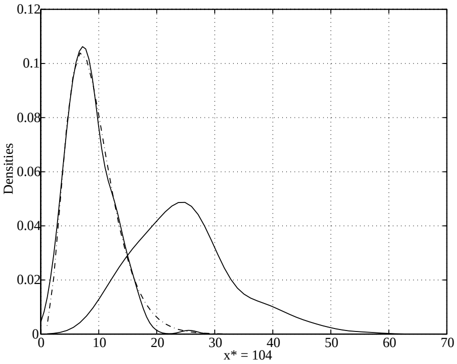


Fig. 5.12. *On the left, densities of $\sum_{i=1}^{10} Z_f^2(x_i)$ and $\chi^2(10)$, and on the right, density of $T_{10}(500)$.*

For each model, $m = 200$ independent realizations of length $n = 200, 500$, and 1000 are generated. We are interested in the empirical level under H_0 and the empirical power under H_1 as well as in the closeness between the simulated distribution of the test statistic and the corresponding theoretical distribution. We use the design points selection rule proposed in [15]. We take $q = 7$, 10, and 14 for $n = 200$, 500, and 1000, respectively.

Let us first examine the "linear" case. One can find in Table 5.1 the results obtained for $x^* = 70$ and $x^* = 170$. The closeness between the simulated distribution of the test statistic and the theoretical distribution is given by the Kolmogorow–Smirnoff (K.S.) statistic and is underlined if the test is rejected at the 5% level (the critical value is equal to 0.096). Only the data X_n obtained after time $n > t_{learn}$ are used for the test. The time t_{learn} corresponds to the end of the learning period needed to stabilize both X_n and $\widehat{f}_n$.

The empirical levels are close to the 5% theoretical level, and the simulated distribution of $T_q(n)$ is close to the $\chi^2(q)$ distribution even for moderate sample sizes.

Let us now examine the "nonlinear" case. Two problems are considered: the empirical power and the convergence in distribution under the alternative hypothesis.

In order to examine the last point, it is useful to check empirical power validity by measuring the closeness between the simulated distribution of $\sum_{i=1}^{q} Z_f^2(x_i)$ (since function f is known in the simulation case), and the $\chi^2(q)$ distribution, according to the following convergence result: $\sum_{i=1}^{q} Z_f^2(x_i) \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \chi^2(q)$, where

$$Z_f(x) = \sqrt{\frac{n^{1-\alpha}(1+\alpha)}{\|K\|_2^2\,\widehat{\Gamma}_n}}\sqrt{\widehat{p}_n(x-x^*)}\left(\widehat{f}_n(x) - f(x)\right). \tag{5.1}$$

TABLE 5.2

Test results under H_1. $t_{learn} = 200$. Percentage of correct decisions.

Working point	n	200	500	1000
$x^* = 104$	*Reject (%)*	19.5%	55%	92%
	K.S	0.052	0.049	0.034
$x^* = 107$	*Reject (%)*	63%	95%	100%
	K.S	0.095	0.087	0.076
$x^* = 110$	*Reject (%)*	100%	100%	100%
	K.S	0.42	0.316	0.163
$x^* = 113$	*Reject (%)*	100%	100%	100%
	K.S	0.037	0.095	0.101
$x^* = 118$	*Reject (%)*	100%	100%	100%
	K.S	0.141	0.1621	0.305
$x^* = 122$	*Reject (%)*	100%	100%	100%
	K.S	0.107	0.063	0.104

TABLE 5.3

Test results under H_1. $t_{learn} = 500$. Percentage of correct decisions.

Working point	n	200	500	1000
$x^* = 110$	*Reject (%)*	99.5%	99%	99%
	K.S	0.082	0.033	0.090
$x^* = 118$	*Reject (%)*	100%	100%	100%
	K.S	0.034	0.033	0.092

The closeness between the two distributions is quantified by the classical K.S. statistic underlined in Table 5.2 if rejected at the 5% level.

For $x^* = 104$ and $n = 500$, these two densities (both centered at 10), and the density distribution of the test statistic $T_{10}(500)$ are shown in Figure 5.12, illustrating both reasonable convergence and the power of the test in this special case.

For various working points (see Figure 5.11), one can find in Table 5.2 the percentages of correct decisions obtained using a short learning period defined by $t_{learn} = 200$. The convergence results are quite satisfactory for $x^* = 104$, 107, 113, and 122 (see the K.S. statistic which is less than or near to the critical value) but remain unstable due to this too short learning period. For $x^* = 104$ and $x^* = 107$, the percentage of correct decisions is small for short sample sizes since the nonlinearity of function f is hard to detect. Nevertheless, this percentage increases with n and is close to 100% when $n = 1000$. So the test behaves well.

For the other working points $x^* = 110$ and 118, the learning period is too short in order to consider that the stationarity of the closed-loop model is sufficiently accurate. In Table 5.3, for a larger learning period of length $t_{learn} = 500$, the K.S. statistics become correct and the percentages of rejections are about 100%.

Appendix A. We give in this first appendix two useful results and their proofs. In what follows, the notation *cte* denotes any positive constant and $\mathcal{F} = (\mathcal{F}_n)_{n\geq 1}$ with $\mathcal{F}_n = \sigma(X_0, U_0, \xi_1, \ldots, \xi_n)$.

LEMMA A.1. *Assume that assumptions* [A1]–[A3] *hold. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a positive, Lipschitz, bounded function satisfying $\int \varphi(t)\,dt < \infty$, $\int \|t\|\,\varphi(t)\,dt < \infty$, and let γ and λ be two positive real numbers such that $\gamma \in\,]0,\, 1/d[$ and $\lambda \geq \gamma d$.*

Let us denote $L_n(x) = \sum_{i=1}^{n} i^{\lambda}\,\varphi\left(i^{\gamma}(X_i - x)\right)$, $\lambda' = \lambda - \gamma d + 1$, and $\overline{\varphi} = \int \varphi(t)\,dt$.

(1) *Then, for any $x \in \mathbb{R}^d$, $\lambda' n^{-\lambda'} L_n(x) \underset{n\to\infty}{\xrightarrow{\text{a.s.}}} \overline{\varphi}\, p(x - x^*)$.*

(2) *In addition, if* $\varphi(x) = O(\|x^{-\delta}\|)$ *with* $\delta > \frac{m\,\gamma\,d}{\gamma\,m+1}$, *then*

$$\sup_{x\in\mathbb{R}^d} \left|\lambda'\, n^{-\lambda'}\, L_n(x) - \overline{\varphi}\; p(x-x^*)\right| \;\xrightarrow[n\to\infty]{\text{a.s.}}\; 0.$$

Proof. Let us rewrite

$$L_n(x) - n^{\lambda'}\,\frac{\overline{\varphi}}{\lambda'}\,p(x-x^*) = M_n^L(x) + (L_n^c(x) - J_n(x)) + R_n(x), \tag{A.1}$$

where

$$M_n^L(x) = L_n(x) - L_n^c(x), \tag{A.2}$$

$$\begin{aligned} L_n^c(x) &= \sum_{i=1}^{n} i^{\lambda}\,\mathbb{E}\left[\varphi\left(i^{\gamma}(X_i - x)\right)\,/\,\mathcal{F}_{i-1}\right] \\ &= \sum_{i=1}^{n} \int i^{\lambda-\gamma d}\,\varphi(t)\,p\left(i^{-\gamma}t - \pi_{i-1} + x - X_i^*\right)\,dt, \end{aligned} \tag{A.3}$$

$$J_n(x) = \overline{\varphi}\;\sum_{i=1}^{n} i^{\lambda-\gamma d}\,p\left(x - \pi_{i-1} - X_i^*\right), \tag{A.4}$$

$$R_n(x) = J_n(x) - n^{\lambda'}\,\frac{\overline{\varphi}}{\lambda'}\,p(x-x^*). \tag{A.5}$$

Since $\lambda'\,n^{-\lambda'}\sum_{i=1}^{n}\; i^{\lambda-\gamma d} \xrightarrow[n\to\infty]{} 1$, $\overline{\varphi} < \infty$, $\int \|t\|\,\varphi(t)\,dt < \infty$, and $\|Dp\|_\infty < \infty$, where Dp denotes the gradient of p,

$$|R_n(x)| \le \overline{\varphi}\;\|Dp\|_\infty \sum_{i=0}^{n-1} i^{\lambda-\gamma d}\,\left(\|\pi_i\| + \|X_i^* - x^*\|\right) \tag{A.6}$$

$$\left|L_n^c(x) - J_n(x)\right| \le \|Dp\|_\infty \left(\int \|t\|\,\varphi(t)\,dt\right) \sum_{i=1}^{n}\; i^{\lambda-\gamma-\gamma d}. \tag{A.7}$$

Therefore, since $X_n^* \xrightarrow[n\to\infty]{} x^*$, $\sum_{k=0}^{n-1}\|\pi_k\|^2 \overset{\text{a.s.}}{=} o(n)$, and $\gamma > 0$, we derive that

$$\sup_{x\in\mathbb{R}^d} \left|\lambda' n^{-\lambda'}\; L_n^c(x) - \overline{\varphi}\; p(x-x^*)\right| \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{A.8}$$

Now, let us study $M_n^L(x)$. For any $x \in \mathbb{R}^d$ and any $n \ge 1$, $M_n^L(x)$ is a square integrable martingale adapted to $\mathcal{F}$ for which we have, if we set $M_0^L(x) = 0$ and $\Delta_i(x) = M_i^L(x) - M_{i-1}^L(x)$,

$$\begin{aligned} \langle M^L(x)\rangle_n &= \sum_{i=1}^{n}\; \mathbb{E}\left[\Delta_i(x)^2/\mathcal{F}_{i-1}\right] \\ &\le \|p\|_\infty \left(\int \varphi^2(t)\,dt\right) \sum_{i=1}^{n} i^{2\lambda-\gamma d} = O\left(n^{1+2\lambda-\gamma d}\right). \end{aligned} \tag{A.9}$$

Then, from a strong law of large numbers for the martingales (for example, Duflo [5, Theorem 1.3.17, p. 21]), we obtain that for any $x \in \mathbb{R}^d$,

$$n^{-\lambda'}\; M_n^L(x) \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{A.10}$$

Combining (A.8) and (A.10) gives part 1 of Lemma A.1.

To prove part 2, let us first establish the uniform almost sure convergence over dilating sets of $M_n^L(x)$. From (A.9), we have

$$\langle M(0)\rangle_n \leq \text{cte } n^{1+2\lambda-\gamma d}, \tag{A.11}$$

and since φ is supposed to be bounded,

$$|\Delta_n(0)| \leq \text{cte } n^{\lambda}. \tag{A.12}$$

In addition, since φ is supposed to be Lipschitz, we have for any $x, y \in \mathbb{R}^d$ and for any $\eta \in]0,\, 1[$,

$$\begin{aligned} \text{(A.13)}\ |\Delta_n(x) - \Delta_n(y)| &\leq \text{cte } n^{\lambda(1-\eta/2)}\, |\Delta_n(x) - \Delta_n(y)|^{\eta/2} \\ &\leq \text{cte } \|x-y\|^{\eta/2}\ n^{\lambda+\gamma\eta/2}, \\ \langle M^L(x) - M^L(y)\rangle_n &\leq \sum_{i=1}^n i^{2\lambda}\, \mathbb{E}\left[\left(\varphi\left(i^\gamma(X_i - x)\right) - \varphi\left(i^\gamma(X_i - y)\right)\right)^2/\mathcal{F}_{i-1}\right] \\ \text{(A.14)}\qquad &\leq \|p\|_\infty \sum_{i=1}^n i^{2\lambda-\gamma d} \int \left(\varphi(t) - \varphi(t + i^\gamma(x-y))\right)^{(2-\eta)+\eta}\, dt \\ &\leq \text{cte } \|x-y\|^\eta\ n^{1+2\lambda-\gamma d+\gamma\eta}. \end{aligned}$$

Then, with (A.11)–(A.14), the assumptions of Theorem 6.4.34 of Duflo [5] are fulfilled, and we derive that for any $A < \infty$, $\nu > 0$, and $t > \lambda - \frac{\gamma d}{2} + \frac{1}{2}$,

$$\sup_{\|\,x\,\| \leq A\, n^\nu} \left|M_n^L(x)\right| = o(n^t), \quad \text{a.s.,} \tag{A.15}$$

and, in particular, since $\gamma \in]0,\, 1/d[$, we can choose $t = \lambda'$. Then, combining (A.8) and (A.15), we obtain that for any $A < \infty$ and $\nu > 0$

$$\sup_{\|\,x\,\| \leq A\, n^\nu} \left|\lambda'\, n^{-\lambda'}\, L_n(x) - \overline{\varphi}\, p(x - x^*)\right| \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{A.16}$$

Besides, since $\varphi(t) = O(\|t\|^{-\delta})$, then $L_n(x) \overset{\text{a.s.}}{=} O(\sum_{i=1}^n i^{\lambda-\gamma\delta}\, \|X_i - x\|^{-\delta})$.

Since $\sup_{k\leq n} \|X_k\| = o\left(n^{1/m}\right)$ a.s., there is an integer n^* such that for $n \geq n^*$, we have $\|X_n\| < n^{1/m}$. Hence, for x such that $\|x\| > 2\, n^{1/m}$, we have $\|X_n - x\| > n^{1/m}$ and

$$\sup_{\|x\| > 2\, n^{1/m}} n^{-\lambda'}\, L_n(x) = O(n^{\gamma d - \gamma\delta - \delta/m}) \ \text{ a.s.} \tag{A.17}$$

In addition, since $p > 0$ and $m > 2$, there is $\tau > 0$ such that $p(x) = o(\|x\|^{-\tau})$ and

$$\sup_{\|x\| > 2\, n^{1/m}} p(x - x^*) = o(n^{-\tau/m}), \tag{A.18}$$

and therefore, since $\tau > 0$ and $\delta > \frac{m\gamma d}{\gamma m + 1}$, we derive from (A.17) and (A.18) that

$$\sup_{\|x\| > 2\, n^{1/m}} \left|\lambda'\, n^{-\lambda'}\, L_n(x) - \overline{\varphi}\, p(x - x^*)\right| \xrightarrow[n\to\infty]{\text{a.s.}} 0, \tag{A.19}$$

and we close the proof of Lemma A.1 combining (A.19) and (A.16) with $A = 2$ and $\nu = 1/m$. □

Corollary A.2. *If we set* $L_n(x) = \sum_{i=1}^n i^\lambda \, \varphi\left(i^\gamma (X_i - X_i^* - x)\right)$, *then we obtain the results of Lemma* A.1 *with* $p(x)$ *instead of* $p(x - x^*)$.

Now we give a result of a strong law of large numbers for the process $(X_n)_{n\geq 1}$.

Lemma A.3. *Assume that* [A1]–[A3] *hold and that* $m > 4$. *Let* $g : \mathbb{R}^d \to \mathbb{R}$ *be a function such that*

$$|g(x) - g(y)| \leq \text{cte } \|x - y\| \left(1 + \|x\| + \|y\|\right). \tag{A.20}$$

Then, $\frac{1}{n} \sum_{k=1}^n g(X_k) \underset{n\to\infty}{\xrightarrow{\text{a.s.}}} \mathbb{E}\left(g(\xi_1 + x^*)\right)$.

Proof. Let us rewrite $\sum_{k=1}^n g(X_k) = M_n + \sum_{k=1}^n \mathbb{E}\left[g(X_k) \,/\, \mathcal{F}_{k-1}\right]$, where

$$M_n = \sum_{k=1}^n \left(g(X_k) - \mathbb{E}\left[g(X_k) \,/\, \mathcal{F}_{k-1}\right]\right).$$

Since

$$\mathbb{E}\left[g(X_k) \,/\, \mathcal{F}_{k-1}\right] - \mathbb{E}\left(g(\xi_1 + x^*)\right) = \int \left(g\left(u + \pi_{k-1} + X_k^*\right) - g(u + x^*)\right) p(u)\, du,$$

$$\begin{aligned}&\left|g\left(u + \pi_{k-1} + X_k^*\right) - g(u + x^*)\right| \\ &\quad \leq \text{cte } \|\pi_{k-1} + (X_k^* - x^*)\| \left(1 + \|u\| + \|\pi_{k-1}\| + \|X_k^* - x^*\|\right),\end{aligned}$$

and $\int \|u\| \; p(u)\, du < \infty$, we derive that

$$\left|\frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[g(X_k) \,/\, \mathcal{F}_{k-1}\right] - \mathbb{E}\left(g(\xi_1 + x^*)\right)\right| = O\left(\frac{1}{n}\sum_{k=1}^n \left(\|\pi_{k-1}\|^2 + \|X_k^* - x^*\|^2\right)\right).$$

Therefore, since $X_n^* \underset{n\to\infty}{\longrightarrow} x^*$ and $\sum_{k=1}^n \|\pi_{k-1}\|^2 = o(n)$, a.s., then

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[g(X_k) \,/\, \mathcal{F}_{k-1}\right] \underset{n\to\infty}{\xrightarrow{\text{a.s.}}} \mathbb{E}\left(g(\xi_1 + x^*)\right). \tag{A.21}$$

In addition, since $\sup_n \mathbb{E}\left[\|X_n\|^m\right] < \infty$ with $m > 4$, for any $n \geq 1$, M_n is a square integrable martingale adapted to $\mathcal{F}$ for which we have, if we set $M_0 = 0$,

$$\begin{aligned}\langle M\rangle_n &= \sum_{k=1}^n \mathbb{E}\left[\left(M_k - M_{k-1}\right)^2 /\mathcal{F}_{k-1}\right] \leq \sum_{k=1}^n \mathbb{E}\left[g^2(X_k)/\mathcal{F}_{k-1}\right]. \\ &= O\left(\sum_{k=1}^n \mathbb{E}\left[\left(1 + \|X_k\|^4\right)/\mathcal{F}_{k-1}\right]\right).\end{aligned} \tag{A.22}$$

Besides, since $\|X_k\|^4 = O(1 + \|\pi_{k-1}\|^4 + \|\xi_k\|^4)$ and $\mathbb{E}(\|\xi_1\|^4) < \infty$, we obtain that

$$\langle M\rangle_n = O\left(\sum_{k=1}^n (1 + \|\pi_{k-1}\|^4)\right) \;=\; o(n^{1+2/m}) \text{ a.s.} \tag{A.23}$$

Finally, since $m > 2$, we deduce from a strong law of large numbers for martingales that $n^{-1} M_n \underset{n\to\infty}{\xrightarrow{\text{a.s.}}} 0$. Lemma A.3 is established by combining this last result with (A.21). □

Appendix B. This appendix is concerned with the proof of Theorem 3.2. Without any restriction, we focus our attention on $\widehat{F}_n$, since the proofs of the results concerning $\widehat{F}_n$ and $\widehat{f}_n$ are the same. Indeed, in both cases, as mentioned in Remark 3.3, $\widehat{f}_n$ is used for the control law to ensure that the key property $\sum_{k=0}^{n-1} \|\pi_k\|^2 = o(n)$ holds.

For $x \in \mathbb{R}^d$, we can write

$$\text{(B.1)}\quad n^{(1-\beta d)/2}\left(\widehat{F}_n(x) - f(x)\right) = \frac{n}{H_{n-1}(x)}\left(n^{-(1+\beta d)/2}(R_{n-1}(x) + M_n(x))\right),$$

where

$$\text{(B.2)}\qquad H_{n-1}(x) = \sum_{i=0}^{n-1} i^{\beta d} G\left(i^{\beta}(X_i - x)\right),$$

$$\text{(B.3)}\qquad R_{n-1}(x) = \sum_{i=0}^{n-1} i^{\beta d} G\left(i^{\beta}(X_i - x)\right)(f(X_i) - f(x)),$$

$$\text{(B.4)}\qquad M_n(x) = \sum_{i=0}^{n-1} i^{\beta d} G\left(i^{\beta}(X_i - x)\right)\xi_{i+1}.$$

Let us study the convergence of the three terms $n^{-1}H_{n-1}(x)$, $n^{-(1+\beta d)/2}R_{n-1}(x)$, and $n^{-(1+\beta d)/2}M_n(x)$, respectively.

Applying Lemma A.1 with $\varphi = G$, $\lambda = \beta d$, and $\gamma = \beta$ gives for $\beta \in \,]0, 1/d[$

$$\text{(B.5)}\qquad \frac{1}{n} H_{n-1}(x) \underset{n\to\infty}{\overset{\text{a.s.}}{\longrightarrow}} p(x - x^*).$$

For $x \in \mathbb{R}^d$, $M_n(x)$ is a square integrable martingale adapted to $\mathcal{F}$ for which we have

$$\text{(B.6)}\qquad \begin{aligned}\langle M(x)\rangle_n &= \sum_{i=1}^{n} \mathbb{E}\left[(M_i(x) - M_{i-1}(x))(M_i(x) - M_{i-1}(x))^T \;/\mathcal{F}_{i-1}\right]\\ &= \left(\sum_{i=0}^{n-1} i^{2\beta d}\, G^2\left(i^{\beta}(X_i - x)\right)\right)\Gamma.\end{aligned}$$

Since G is Lipschitz and bounded, G^2 is also Lipschitz and bounded. Then, by Lemma A.1 used with $\varphi = G^2$, $\lambda = 2\beta d$, and $\gamma = \beta$,

$$\text{(B.7)}\qquad n^{-(1+\beta d)}\langle M(x)\rangle_n \underset{n\to\infty}{\overset{\text{a.s.}}{\longrightarrow}} \frac{\|G\|_2^2}{1+\beta d}\, p(x - x^*)\,\Gamma.$$

Now to apply the central limit theorem for the martingales to $n^{-(1+\beta d)/2}M_n(x)$ (see, for example, Duflo [5, Theorem 2.1.9, p. 46]), Lindeberg's condition remains to be proved. Let us denote $\tau_i = i^{\beta d}G(i^{\beta}\,(X_i - x))\xi_{i+1}$ and $\Phi(c) = \mathbb{E}[\|\xi_1\|^2\, \mathbf{1}_{\{\|\,\xi_1\,\| \,\geq\, c\}}]$.

For $\varepsilon > 0$, we have

$$\text{(B.8)}\qquad \begin{aligned}\rho_n(\varepsilon) &= n^{-(1+\beta d)} \sum_{i=0}^{n-1} \mathbb{E}\left[\|\tau_i\|^2\, \mathbf{1}_{\left\{\|\tau_i\| \geq n^{(1+\beta d)/2}\varepsilon\right\}}/\mathcal{F}_i\right]\\ &\leq n^{-(1+\beta d)}\; \Phi\left(\frac{n^{(1-\beta d)/2}\varepsilon}{\|G\|_\infty}\right)\sum_{i=0}^{n-1} i^{2\beta d}G^2\left(i^{\beta}(X_i - x)\right).\end{aligned}$$

Besides, for $i \geq 1$,

$$\mathbb{E}\left[i^{2\beta d} G^2\left(i^\beta (X_i - x)\right) / \mathcal{F}_{i-1}\right] \leq i^{\beta d} \|p\|_\infty \|G\|_2^2, \tag{B.9}$$

and thus,

$$\mathbb{E}\left(\rho_n(\varepsilon)\right) \leq \text{cte}\left(n^{-(1+\beta d)} \sum_{i=0}^{n-1} i^{\beta d}\right) \Phi\left(\frac{n^{(1-\beta d)/2}\varepsilon}{\|G\|_\infty}\right). \tag{B.10}$$

Since $n^{-(1+\beta d)} \sum_{i=0}^{n-1} i^{\beta d}$ has a finite limit and since $\lim_{c\to\infty} \Phi(c) = 0$ (since ξ has a finite moment of order $m > 2$), we derive that for $\beta \in]0,\, 1/d[$,

$$\forall \varepsilon > 0, \quad \rho_n(\varepsilon) \xrightarrow[n\to\infty]{\mathrm{P}} 0, \tag{B.11}$$

and Lindeberg's condition is fulfilled. Therefore, for $x \in \mathbb{R}^d$ and $\beta \in]0,\, 1/d[$,

$$n^{-(1+\beta d)/2} M_n(x) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0,\, \frac{\|G\|_2^2}{(1+\beta d)}\, p(x - x^*)\, \Gamma\right). \tag{B.12}$$

Now, let us show that for $x \in \mathbb{R}^d$ and $\beta > 1/(d+4)$,

$$n^{-(1+\beta d)/2}\, R_{n-1}(x) \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{B.13}$$

First, let us rewrite $R_n(x)$ under the form

$$R_n(x) = \left(R_n(x) - R_n^c(x)\right) + \left(R_n^c(x) - Q_n(x)\right) + Q_n(x), \tag{B.14}$$

where

$$\begin{aligned} R_n^c(x) &= \sum_{i=1}^n i^{\beta d} \mathbb{E}\left[G\left(i^\beta (X_i - x)\right) \left(f(X_i) - f(x)\right) / \mathcal{F}_{i-1}\right] \\ &= \sum_{i=1}^n \int G(t) \left(f(i^{-\beta} t + x) - f(x)\right) p\left(i^{-\beta} t + x - f(X_{i-1}) - U_{i-1}\right) dt, \\ Q_n(x) &= \sum_{i=1}^n p\left(x - f(X_{i-1}) - U_{i-1}\right) \int G(t)\, \left(f(i^{-\beta} t + x) - f(x)\right) dt. \end{aligned}$$

Since f is Lipschitz, $\|Dp\|_\infty < \infty$, and $\int \|t\|^2 G(t)\, dt < \infty$, we easily show that

$$\sup_{x\in\mathbb{R}^d} \|R_n^c(x) - Q_n(x)\| = O\left(n^{1-2\beta}\right) \quad \text{a.s.} \tag{B.15}$$

Let us denote f_ℓ the ℓth component of f. Since f is supposed to be $\mathcal{C}^2$−class, for $1 \leq \ell \leq d$, we have the expansion

$$f_\ell(i^{-\beta} t + x) - f_\ell(x) = \sum_{j=1}^d i^{-\beta}\, t_j\, \frac{\partial f_\ell}{\partial x_j}(x) + \frac{1}{2} \sum_{j=1}^d \sum_{k=1}^d i^{-2\beta}\, t_j\, t_k\, \frac{\partial^2 f_\ell}{\partial x_j \partial x_k}(z)$$

with $z \in \left[x,\, x + i^{-\beta} t\right]$. Thus, since $\int t_j G(t)\, dt = 0$ for any $j = 1, \ldots, d$, we derive that

$$\left\| \int G(t)\, \left(f(i^{-\beta} t + x) - f(x)\right) dt \right\| = O\left(i^{-2\beta}\right), \tag{B.16}$$

and since p is bounded, then we have

$$\sup_{x\in\mathbb{R}^d} \|Q_n(x)\| = O\left(n^{1-2\beta}\right) \quad \text{a.s.} \tag{B.17}$$

For $x \in \mathbb{R}^d$, $M_n^R(x) = R_n(x) - R_n^c(x)$ is a square integrable martingale adapted to $\mathcal{F}$ for which we have

$$\sum_{i=1}^{n} \mathbb{E}\left[\left\|M_i^R(x) - M_{i-1}^R(x)\right\|^2 / \mathcal{F}_{i-1}\right] = O\left(\sum_{i=1}^{n} i^{\beta d - 2\beta}\right) \quad \text{a.s.,} \tag{B.18}$$

since f is Lipschitz, G is bounded, $\|p\|_\infty < \infty$, and $\int \|t\|^2 G(t)\, dt < \infty$. Therefore, using a result of a strong law of large numbers for the martingales, we obtain that for $\delta > 0$,

$$M_n^R(x) = o(n^{(1+\beta d)/2-\beta} (\log n)^{(1+\delta)/2}) \quad \text{a.s.} \tag{B.19}$$

Then, using results (B.15), (B.17), and (B.19), we prove (B.13) as soon as $\beta > 1/(d+4)$. Finally, combining (B.5), (B.12), and (B.13), we obtain that for $x \in \mathbb{R}^d$, $\beta < 1/d$, and $\beta > 1/(d+4)$,

$$n^{(1-\beta d)/2}(\widehat{F}_n(x) - f(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\|G\|_2^2}{(1+\beta d)\, p(x-x^*)}\, \Gamma\right). \tag{B.20}$$

Now following Duflo [5], let us study the joint asymptotic normality. Taking the previous results into account, it suffices to prove that for q distinct points of $\mathbb{R}^d$, denoted $x_1, \ldots, x_q$, the vector $n^{-(1+\beta d)/2} (M_n(x_1), \ldots, M_n(x_q))$ converges in distribution to a centered Gaussian vector with independent components. We easily verify this by remarking that for $x \neq y$,

$$n^{-(1+\beta d)} \sum_{i=1}^{n-1} i^{2\beta d}\, G\left(i^\beta(X_i - x)\right)\, G\left(i^\beta(X_i - y)\right) \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

This completes the proof of Theorem 3.2.

Remark B.1. From (B.6) it follows easily that

$$\mathbb{E}\left[\|\langle M(x)\rangle_n\|\right] = O\left(n^{1+\beta d}\right). \tag{B.21}$$

Then, from a strong law of large numbers, we derive that $n^{-1} M_n(x) \xrightarrow[n\to\infty]{\text{a.s.}} 0$ for all $\beta \in]0,\, 1/d[$, and combining this result with (B.5), (B.15), (B.17), and (B.19) gives

$$\widehat{F}_n(x) \xrightarrow[n\to\infty]{\text{a.s.}} f(x). \tag{B.22}$$

This result is useful to prove part 2 of Theorem 4.2 when $\widehat{F}_n(x)$ is used in the construction of the test statistic $T_q(n)$ instead of $\widehat{f}_n(x)$.

REFERENCES

[1] I. A. Ahmad, *Residuals density estimation in nonparametric regression*, Statist. Probab. Lett., 14 (1992), pp. 133–139.

[2] B. Bercu, *Central limit theorem and law of iterated logarithm for least squares algorithms in adaptive tracking*, SIAM J. Control Optim., 36 (1998), pp. 910–928.

[3] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser Boston, Boston, MA, 1991.

[4] P. Doukhan, *Mixing: Properties and Examples*, Lecture Notes in Statist. 85, Springer-Verlag, Berlin, New York, Heidelberg, 1994.

[5] M. Duflo, *Random Iterative Models*, Springer-Verlag, Berlin, New York, Heidelberg, 1997.

[6] A. Georgiev, *Nonparametric system identification by kernel methods*, IEEE Trans. Automat. Control, 29 (1984), pp. 356–358.

[7] W. Greblicki, *Nonparametric identification of Wiener systems*, IEEE Trans. Inform. Theory, 38 (1992), pp. 1487–1493.

[8] W. Greblicki and M. Pawlak, *Nonparametric identification of Hammerstein systems*, IEEE Trans. Inform. Theory, 35 (1989), pp. 409–418.

[9] L. Guo, *Convergence and logarithm laws of self-tuning regulators*, Automatica J. IFAC, 31 (1995), pp. 435–450.

[10] W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK, 1990.

[11] N. Hilgert, R. Senoussi, and J.-P. Vila, *Nonparametric identification of controlled nonlinear time varying processes*, SIAM J. Control Optim., to appear.

[12] V. Hjellvik and D. Tjøstheim, *Nonparametric tests of linearity for time series*, Biometrika, 82 (1995), pp. 351–368.

[13] A. Krzyzak, *On estimation of a class of nonlinear systems by the kernel regression estimate*, IEEE Trans. Inform. Theory, 36 (1990), pp. 141–152.

[14] B. Portier and A. Oulidi, *Nonparametric estimation and adaptive control of functional autoregressive models*, SIAM J. Control Optim., 39 (2000), pp. 411–432.

[15] J.-M. Poggi and B. Portier, *A test of linearity for functional autoregressive models*, J. Time Ser. Anal., 18 (1997), pp. 615–640.

[16] J.-M. Poggi and B. Portier, *A test of linearity for NARX models*, European J. Control, 1 (1998), pp. 298–305.

[17] B. Portier and G. Oppenheim, *Adaptive control of nonlinear dynamic systems: Study of the nonparametric estimator*, J. Syst. Eng., 1 (1993), pp. 40–50.

[18] P. M. Robinson, *Nonparametric estimators for time series*, J. Time Ser. Anal., 4 (1983), pp. 185–207.

[19] C. G. Roussas and L. T. Tran, *Asymptotic normality of the recursive kernel regression estimate under dependence conditions*, Ann. Statist., 20 (1992), pp. 98–120.

[20] E. F. Schuster, *Joint asymptotic distribution of the estimated regression function at a finite number of distinct points*, Ann. Math. Statist., 43 (1972), pp. 84–88.

[21] R. Senoussi, *Uniform iterated logarithm laws for martingales and their application to functional estimation in controlled Markov chains*, Stochastic Process. Appl., to appear.

[22] W. F. Stout, *A martingale analogue of Kolmogorov's law of the iterated logarithm*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 15 (1970), pp. 279–290.

[23] D. Tjøstheim, *Non-linear time series: A selective review*, Scand. J. Statist., 21 (1994), pp. 97–130.

[24] H. Tong, *Nonlinear Time Series: A Dynamical Approach*, Oxford Science Publications, Oxford, UK, 1990.

[25] A. Touati, *Vitesse de convergence en loi de l'estimateur des moindres carrés d'un modèle autorégressif (cas mixte)*, Ann. Inst. H. Poincaré Probab. Statist., 32 (1996), pp. 211–230.

[26] Y. K. Truong and C. J. Stone, *Nonparametric function estimation involving time series*, Ann. Statist., 20 (1992), pp. 77–97.

[27] S. Yakowitz, *Nonparametric density and regression*, J. Multivariate Anal., 30 (1989), pp. 124–136.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 728–756

CONTROLLABILITY OF SYSTEMS DESCRIBED BY CONVOLUTIONAL OR DELAY-DIFFERENTIAL EQUATIONS*

P. VETTORI† AND S. ZAMPIERI†

Abstract. In this paper, infinite dimensional systems described by convolutional equations and, in particular, by delay-differential equations are considered. Different controllability notions for this class of linear time-invariant systems are discussed and compared. A characterization of spectral controllability for a system whose trajectories satisfy a homogeneous system of independent convolutional equations is given. This result extends an analogous result which was known to hold for difference or differential equations. Finally, for a particular class of systems, including systems in the state space form, it is shown that a well-known theorem, which states the equivalence of spectral controllability and the existence of an image representation, holds true. An example is presented showing that this result is false for generic delay-differential systems as soon as there are two noncommensurate delays.

Key words. convolutional equations, delay-differential systems, linear systems, behavioral approach, functional equations, distributed delays, controllability

AMS subject classifications. 34K35, 39B62, 46E10, 47N70, 93B05, 93B25, 93B28, 93C30

PII. S0363012999359718

1. Introduction.

In solving control and optimization problems, the analysis of system controllability is always an important preliminary step. Differently from what happens for linear finite dimensional systems, for linear infinite dimensional systems controllability can be defined in various ways, and, moreover, its characterization is often a difficult task. For the class of infinite dimensional systems described by linear differential equations with delays, something more can be said, in particular, when the delays are commensurate, i.e., when they are multiples of a single delay. In [7] various definitions of controllability are proposed and compared for this class of systems.

The notion of controllability that has been analyzed in more detail in literature is the so-called spectral controllability, because it plays an important role in the spectrum assignment problem. For systems with commensurate delays, this controllability notion has been interpreted in terms of trajectories steering in [9, 23]. The use of the behavioral approach and, in particular, the notion of behavioral controllability which arises in this framework has been essential for obtaining these results.

For this reason, in this paper we will follow this approach in the analysis of controllability of delay systems and we will try to extend the results in [9, 23] to systems with noncommensurate delays and, more in general, to systems described by convolutional equations.

Now we briefly recall the notion of a dynamical system in the behavioral approach and the definition of controllability. According to this approach a dynamical system is defined as a triple

$$\Sigma = (T, W, \mathcal{B}),$$

where T denotes the time set, W denotes the system alphabet, and $\mathcal{B}$, which is a subset of the set W^T of all the trajectories, represents the set of trajectories which

*Received by the editors July 30, 1999; accepted for publication (in revised form) March 21, 2000; published electronically October 4, 2000.

http://www.siam.org/journals/sicon/39-3/35971.html

†Dipartimento di Elettronica e Informatica, Università di Padova, via Gradenigo 6/a, 35131 Padova, Italy (p.vettori@dei.unipd.it, zampi@dei.unipd.it).

are allowed by the dynamical constraints imposed by the system. This is called the behavior of the system.

In this paper we will consider only systems with time set $T = \mathbb{R}$ and with time invariant behavior $\mathcal{B}$, i.e., such that for any $w \in \mathcal{B}$ we have $\sigma_\tau w \in \mathcal{B}$ for every $\tau \in \mathbb{R}$. (σ_τ is the forward shift operator which will be defined in section 2.1.) These systems will be called continuous time-invariant systems.

Differently from what happens in the classical systems theory, controllability is defined as an intrinsic property of the system rather than a property of a state space representation of the system.

Definition 1.1. *A continuous time-invariant system $\Sigma = (\mathbb{R}, W, \mathcal{B})$ is controllable if for every $w_1, w_2 \in \mathcal{B}$ there exists a trajectory $\bar{w} \in \mathcal{B}$ and a $T > 0$ such that*

$$\bar{w}(t) = \begin{cases} w_1(t) & \textit{if } t \le 0, \\ w_2(t-T) & \textit{if } t \ge T. \end{cases} \tag{1}$$

Loosely speaking, given two trajectories of a controllable behavior, there exists a trajectory in the behavior that shares the “past” with the first one and the “future” with the second one.

In the first papers devoted to the behavioral approach [22, 27], a rather complete theory for systems described by linear differential equations has been developed. Moreover, a complete characterization of controllability has been proposed. More specifically, the class of continuous systems that has been analyzed in [22] has time set $T = \mathbb{R}$, system alphabet $W = \mathbb{R}^q$, and behavior

$$\mathcal{B} = \ker R\left(\tfrac{d}{dt}\right) \triangleq \{w \in C^\infty(\mathbb{R}, \mathbb{R}^q) \text{ such that } R\left(\tfrac{d}{dt}\right) w = 0\},$$

where $R\left(\frac{d}{dt}\right) \in \mathbb{R}\left[\frac{d}{dt}\right]^{p\times q}$ is a polynomial differential operator acting from $C^\infty(\mathbb{R}, \mathbb{R}^q)$ to $C^\infty(\mathbb{R}, \mathbb{R}^p)$. The polynomial matrix provides a so-called kernel representation of the system behavior $\mathcal{B}$. It may happen that the behavior $\mathcal{B}$ also admits an image representation. This means that there exists a polynomial matrix $M\left(\frac{d}{dt}\right) \in \mathbb{R}\left[\frac{d}{dt}\right]^{q\times d}$ such that

$$\mathcal{B} = \operatorname{im} M\left(\tfrac{d}{dt}\right) \triangleq \{w = M\left(\tfrac{d}{dt}\right) v \in C^\infty(\mathbb{R}, \mathbb{R}^q) \text{ with } v \in C^\infty(\mathbb{R}, \mathbb{R}^d)\}.$$

One of the main results shown in [22] is that the existence of an image representation is equivalent to the controllability of the system. More precisely, we have the following theorem [22, Thm. 5.2.5, 6.6.1].

Theorem 1.2. *Given the matrix $R\left(\frac{d}{dt}\right) \in \mathbb{R}\left[\frac{d}{dt}\right]^{p\times q}$ and the behavior $\mathcal{B} = \ker R\left(\frac{d}{dt}\right)$, the following conditions are equivalent.*

1. *$R(\lambda)$ has constant rank for all $\lambda \in \mathbb{C}$.*
2. *$\mathcal{B}$ is controllable.*
3. *$\mathcal{B}$ admits an image representation: $\mathcal{B} = \operatorname{im} M\left(\frac{d}{dt}\right)$, where $M\left(\frac{d}{dt}\right) \in \mathbb{R}\left[\frac{d}{dt}\right]^{q\times d}$.*

Condition 1 corresponds to a generalized Popov–Belevitch–Hautus (PBH) test condition, which is also called spectral controllability.

In the discrete case [27, Thm. V.2, V.3], a similar result holds true. Actually, all the conditions in the previous theorem are shown [26, Prop. 4.3] to be equivalent to the fact that

$$\mathcal{B} = \overline{\mathcal{B}_{\text{cs}}},$$

where $\mathcal{B}_{\mathrm{cs}}$ means the subset of $\mathcal{B}$ constituted of trajectories which have compact support and where $\bar{\cdot}$ means closure with respect to the pointwise convergence topology in the space of all sequences. It can be shown that a similar equivalence holds true also in the continuous time case if we take on $C^\infty(\mathbb{R}, \mathbb{R}^q)$ the standard Fréchet topology of the uniform convergence of derivatives of any order on compacts.

Consider now the situation in which we take behaviors that are kernel of polynomial matrices in the derivative and in one delay operator. More precisely, consider continuous time systems whose behaviors satisfy the relation

$$\mathcal{B} = \{w \in C^\infty(\mathbb{R}, \mathbb{R}^q) \ : \ R\left(\tfrac{d}{dt}, \sigma\right) w = 0\} = \ker R\left(\tfrac{d}{dt}, \sigma\right),$$

where σ is the unitary shift operator $(\sigma f)(t) = f(t-1)$, and where $R\left(\frac{d}{dt}, \sigma\right) \in \mathbb{R}\left[\frac{d}{dt}, \sigma\right]^{p \times q}$ is a polynomial matrix in two variables.

These systems are called differential systems with one delay or with commensurate delays. They have been analyzed in [9, 23], in which the following result has been shown.

THEOREM 1.3. *Given the matrix* $R\left(\frac{d}{dt}, \sigma\right) \in \mathbb{R}\left[\frac{d}{dt}, \sigma\right]^{p \times q}$ *and the behavior* $\mathcal{B} = \ker R\left(\frac{d}{dt}, \sigma\right)$*, the following conditions are equivalent.*

1. $R(\lambda, e^{-\lambda})$ *has constant rank for all* $\lambda \in \mathbb{C}$.
2. $\mathcal{B}$ *is controllable.*
3. $\mathcal{B}$ *admits an image representation* $\mathcal{B} = \operatorname{im} M\left(\frac{d}{dt}, \sigma\right)$*, where* $M\left(\frac{d}{dt}, \sigma\right) \in \mathbb{R}\left[\frac{d}{dt}, \sigma\right]^{q \times d}$.

The above result is similar to Theorem 1.2. Condition 1 is the generalization of spectral controllability for delay-differential systems. It is possible to prove that condition $\mathcal{B} = \overline{\mathcal{B}_{\mathrm{cs}}}$ is equivalent to controllability even in this case. This is a direct consequence of the results we present in this contribution (see Remark 4.8).

In this paper we will present some results on controllability for linear continuous time systems with several noncommensurate delays. We will show to what extent the results presented in [22, 26, 27] and extended in the one-delay case in [9, 23] continue to hold true also in the multidelay case.

More specifically, in this paper we will prove that Theorem 1.3 can be extended to the multidelay case as well as to more generic convolutional operators only partially. We will present a theorem showing what extension is possible in general and a counterexample showing that a complete extension does not hold. Finally, we will specify an interesting subclass of systems for which a complete extension is possible.

Our interest in studying differential systems with several noncommensurate delays is not only of theoretical nature. It is quite clear that a family of noncommensurate delays can be approximated arbitrarily well by a family of delays which are multiples of a single delay. Therefore, one could argue that any differential system with several noncommensurate delays can be approximated by a differential system with only one delay. This is correct in principle. However, we must pay attention to a delicate point: the analysis of systems with noncommensurate delays is important just because sometimes the approximating differential system with one delay may have different structural properties with respect to the original system to be approximated. Example 3.4 describes a differential system with two delays: the system is spectrally controllable if and only if the delays are noncommensurate. However, it admits an image representation only if the ratio of the delays is not a Liouville irrational number (see Proposition 2.6).

Furthermore, observe that even if an approximation would make sense, the better the approximation is, the higher the degrees of the delay-differential operators describing the approximating system have to be, thus giving rise to computational complexity problems.

2. Notation and preliminaries.

In this section we introduce the notation which will be used in this paper. We recall the most important properties of the function spaces and operator algebras which will be considered in the following sections.

2.1. Function spaces and functionals.

We will consider only behaviors that are subsets of smooth functions. More precisely, if we denote by $\mathcal{E}$ the set of infinitely differentiable functions $C^\infty(\mathbb{R},\mathbb{R})$ equipped with the standard Fréchet topology (uniform convergence of derivatives of any order on compacts), then behaviors considered in this paper are always subspaces of $\mathcal{E}^q$ for some positive $q \in \mathbb{N}$.

The subspace of smooth functions having compact support (test functions) is denoted by $\mathcal{D}$, and the set of holomorphic functions on $\mathbb{C}$ is denoted by $\mathcal{O}$.

These spaces have topological duals, i.e., the sets of continuous linear functionals. It is well known that the topological dual $\mathcal{D}'$ of $\mathcal{D}$ coincides with the space of Schwartz distributions, while the topological dual $\mathcal{E}'$ of $\mathcal{E}$ consists in that subset of $\mathcal{D}'$ constituted by compact support distributions [24, Thm. 24.2].

The value of $\alpha \in \mathcal{E}'$ at $w \in \mathcal{E}$ is denoted by $\langle \alpha, w \rangle \in \mathbb{R}$. The shift operator σ_τ, with delay $\tau \in \mathbb{R}$, is defined for every $w \in \mathcal{E}$ as $(\sigma_\tau w)(t) \triangleq w(t-\tau)$. This operator can be used to show how every distribution $\alpha \in \mathcal{E}'$ acts on $\mathcal{E}$ by mapping $w \in \mathcal{E}$ into the convolution $(\alpha \star w)(t) \triangleq \langle \alpha, \sigma_t w \rangle$. It can be shown that convolutions are still in $\mathcal{E}$ and that in this way $\mathcal{E}'$ is isomorphic to the set of continuous linear operators on $\mathcal{E}$ that commute with σ_τ for every $\tau \in \mathbb{R}$ (see [25, Thm. 2.16]).

2.2. Paley–Wiener functions.

It is possible to define the Laplace transforms of distributions in $\mathcal{E}'$ [24, Exercise 25.20]. The Laplace transform maps a distribution with compact support $\alpha \in \mathcal{E}'$ into $\hat{\alpha}(s) \triangleq \langle \alpha, e^{-st} \rangle$, where α acts on e^{-st}, regarded as a function of t with parameter $s \in \mathbb{C}$. It can be seen that $\hat{\alpha}(s)$ is always a holomorphic function [24, Prop. 29.1].

If $a(s) \in \mathcal{O}$ is the Laplace transform of $\alpha \in \mathcal{E}'$ and $w \in \mathcal{E}$ is a smooth function, convolutions will be denoted also as

$$a(s)w \triangleq \alpha \star w \in \mathcal{E}.$$

It can be seen that, given $\alpha, \beta \in \mathcal{E}'$, there is a unique distribution denoted by $\alpha \star \beta \in \mathcal{E}'$, called the convolution of α and β, such that $\alpha \star (\beta \star w) = (\alpha \star \beta) \star w$ for every $w \in \mathcal{E}$; moreover, the Laplace transform of a convolution is the product of the transforms $\widehat{\alpha \star \beta}(s) = \hat{\alpha}(s)\hat{\beta}(s)$. Actually this is a classical and fundamental result. The Paley–Wiener theorem [2, p. 27–28] states that the convolutional algebra of distributions with compact support is isomorphic, via Laplace transform, to the multiplicative algebra of Paley–Wiener functions so defined:

$$(2) \qquad \mathcal{A} \triangleq \{a(s) \in \mathcal{O} \text{ such that } \exists A, B, C > 0,\ |a(s)| \le A(1+|s|)^B e^{C|\operatorname{Re} s|}\ \forall s \in \mathbb{C}\}.$$

In this context, every matrix $M(s) = [m_{ij}(s)] \in \mathcal{A}^{p \times q}$ represents two distinct operators. On one hand, it acts as a convolutional operator on (column) vectors with entries in $\mathcal{E}$, i.e.,

$$M(s) : \mathcal{E}^q \to \mathcal{E}^p,\ w \mapsto v = M(s)w, \text{ i.e., } v_i = \sum_{j=1}^{q} m_{ij}(s) w_j.$$

On the other hand, it can be seen as a linear transformation on (row) vectors with entries in $\mathcal{A}$. To distinguish the different meanings of the same matrix, we use in this case the following notation:

$$\circ M(s) : \mathcal{A}^p \to \mathcal{A}^q,\ a(s) \mapsto b(s) = a(s)M(s),\ \text{i.e., } b_j(s) = \sum_{i=1}^{p} a_i(s)m_{ij}(s).$$

Therefore, the kernel and image of a matrix $M(s) \in \mathcal{A}^{p\times q}$ depend on the operator it represents. When it operates "on the right," usually on $\mathcal{E}$, we have

$$\begin{aligned} \ker_{\mathcal{E}} M(s) &\triangleq \{w \in \mathcal{E}^q \text{ such that } M(s)w = 0\}, \\ \operatorname{im}_{\mathcal{E}} M(s) &\triangleq \{M(s)w \in \mathcal{E}^p \text{ with } w \in \mathcal{E}^q\}. \end{aligned}$$

Conversely, if $M(s)$ represents the linear transformation acting "on the left" on vectors, whose entries are usually in $\mathcal{A}$, we get

$$\begin{aligned} \ker_{\mathcal{A}} \circ M(s) &\triangleq \{a(s) \in \mathcal{A}^p \text{ such that } a(s)M(s) = 0\}, \\ \operatorname{im}_{\mathcal{A}} \circ M(s) &\triangleq \{a(s)M(s) \in \mathcal{A}^q \text{ with } a(s) \in \mathcal{A}^p\} = \mathcal{A}^p M(s). \end{aligned}$$

We remark that, with this convention, $\circ M(s)$ is the adjoint of $M(s)$, and, if we define orthogonals of subsets $E \subseteq \mathcal{E}^q$ or $A \subseteq \mathcal{A}^q$ as

$$E^\perp \triangleq \{a(s) \in \mathcal{A}^q : a(s)w = 0 \ \ \forall w \in E\}, \quad A^\perp \triangleq \{w \in \mathcal{E}^q : a(s)w = 0 \ \ \forall a(s) \in A\},$$

then we can write in a simple way a fundamental result about duality between adjoint operators. (For the proof see, e.g., [24, p. 388], which states the proposition in terms of "polars" [24, p. 196], which in our context coincide with orthogonals.)

PROPOSITION 2.1. *For every $M(s) \in \mathcal{A}^{p\times q}$,*

$$\begin{aligned} \ker_{\mathcal{E}} M(s) &= \operatorname{im}_{\mathcal{A}} \circ M(s)^\perp, & \ker_{\mathcal{E}} M(s)^\perp &= \overline{\operatorname{im}_{\mathcal{A}} \circ M(s)}, \\ \ker_{\mathcal{A}} \circ M(s) &= \operatorname{im}_{\mathcal{E}} M(s)^\perp, & \ker_{\mathcal{A}} \circ M(s)^\perp &= \overline{\operatorname{im}_{\mathcal{E}} M(s)}. \end{aligned}$$

Another useful property of adjoint operators is the following proposition (see [16, Prop. 21.9]).

PROPOSITION 2.2. *For every $M(s) \in \mathcal{A}^{p\times q}$, $\operatorname{im}_{\mathcal{E}} M(s)$ is closed if and only if $\operatorname{im}_{\mathcal{A}} \circ M(s)$ is closed.*

2.3. Operator subalgebras of $\mathcal{A}$.

The notation we have introduced so far permits us to treat differential or delay-differential polynomial operators in a unified manner. Indeed, the action on smooth functions of both the derivative operator $\frac{d}{dt}$ and of the shift operator σ_τ corresponds to a convolution with the distributions in $\mathcal{E}'$ having Laplace transforms s and $e^{-s\tau}$, respectively.

This fact explains why polynomials in $\frac{d}{dt}$ and σ are also called exponential polynomials, being actually isomorphic, through Laplace transform, to elements in the ring $\mathbb{R}\left[s, e^{-s}\right]$ which is in turn isomorphic to the ring of polynomials in two variables $\mathbb{R}[z_0, z_1]$.

Consider now equations involving more delays: if the delays belong to the set $T = \{t_j,\, j = 1, \dots, M\}$, the ring of polynomials in σ_{t_j} is still isomorphic to the ring of polynomials in e^{-st_j}, but not to the ring of polynomials in M indeterminates. However, consider the $\mathbb{Z}$-module generated by $t_1, \dots, t_M$. Being free, it has a basis

$\tau_1, \dots, \tau_m$ composed by positive elements that are also $\mathbb{Q}$-independent. Thus they are called noncommensurate. Observe that in this case polynomials in the delays σ_{τ_i}, polynomials in $e^{-s\tau_i}$, and polynomials in m variables are isomorphic [15, Thm. 1], i.e.,

$$\mathbb{R}[e^{-s\tau_1}, \dots, e^{-s\tau_m}] \cong \mathbb{R}[z_1, \dots, z_m]. \tag{3}$$

However, in order to represent the operators $\sigma_{t_1}, \dots, \sigma_{t_M}$ as a linear combination of powers of $\sigma_{\tau_1}, \dots, \sigma_{\tau_m}$, we have to consider polynomials also with negative powers in σ_{τ_i}, called Laurent polynomials. It can be seen that

$$\mathcal{R}_m \triangleq \mathbb{R}[e^{-s\tau_1}, \dots, e^{-s\tau_m}, e^{s\tau_1}, \dots, e^{s\tau_m}] \cong \mathbb{R}[e^{-st_1}, \dots, e^{-st_M}, e^{st_1}, \dots, e^{st_M}],$$

while a similar result does not hold if we consider the ring of standard polynomials. Observe that, by virtue of (3), we obtain that

$$\mathcal{R}_m \cong \mathbb{R}[z_1, \dots, z_m, z_1^{-1}, \dots, z_m^{-1}] \cong \mathbb{R}[\sigma_{\tau_1}, \dots, \sigma_{\tau_m}, \sigma_{\tau_1}^{-1}, \dots, \sigma_{\tau_m}^{-1}].$$

Delay–differential polynomial operators with time delays belonging to T are thus expressible as polynomials in $\frac{d}{dt}$ and in σ_{τ_i}, i.e., they are elements of the ring $\mathcal{R}_m[s]$. We note that the use of operators like $\sigma_{\tau_i}^{-1}$, which are well defined, only introduces finite anticipative (noncausal) operators. Anyway, their use is the standard practice within behavioral theory of discrete time (i.e., difference) behaviors.

A generalization of the ring of delay-differential polynomial operators was first introduced in [9] in case of commensurate delays. We give here its definition in the general case of m noncommensurate delays.

Proposition 2.3. *The set of holomorphic fractions of exponential polynomials with m delays*

$$\mathcal{H}_m \triangleq \left\{ \frac{n(s)}{d(s)} \text{ such that } n(s), d(s) \in \mathcal{R}_m[s] \right\} \cap \mathcal{O} \tag{4}$$

satisfies the following properties.

1. *$\mathcal{H}_m$ coincides with the set of Laurent polynomials in $e^{-s\tau_i}$, with coefficients in $\mathbb{R}(s)$, that are holomorphic. This means that every element in $\mathcal{H}_m$ can be written as the quotient of a Laurent exponential polynomial in $\mathcal{R}_m[s]$ and a polynomial in $\mathbb{R}[s]$.*
2. *Whenever $n(s), d(s) \in \mathcal{H}_m$, and $\frac{n(s)}{d(s)} \in \mathcal{O}$, then $\frac{n(s)}{d(s)} \in \mathcal{H}_m$.*
3. *If $n(s) \in \mathcal{A}$, $d(s) \in \mathcal{H}_m$, and $\frac{n(s)}{d(s)} \in \mathcal{O}$, then $\frac{n(s)}{d(s)} \in \mathcal{A}$.*
4. *$\mathcal{H}_m$ is a subalgebra of $\mathcal{A}$: every $a(s) \in \mathcal{H}_m$ is an operator of $\mathcal{E}$ unto itself.*
5. *Every nonzero $a(s) \in \mathcal{H}_m$ is a surjective operator: $\operatorname{im}_{\mathcal{E}} a(s) = \mathcal{E}$.*
6. *If $n(s), d(s) \in \mathcal{H}_m$, and $\frac{n(s)}{d(s)} \in \mathcal{H}_m$, then $\ker_{\mathcal{E}} d(s) \subseteq \ker_{\mathcal{E}} n(s)$.*

Proof. Property 1 was proved in [1] for more generic polynomials in $\mathbb{C}^n$; see [11, App. A] for an alternative proof. Properties 5 [6] and 6 [18, pp. 282, 318] are classical results in functional analysis; property 2 trivially follows from the definition of $\mathcal{H}_m$, while properties 3 and 4 can be deduced using the fifth property and [2, Thm. 2.7, 2.8]. □

The way in which elements in $\mathcal{H}_m$ (as well as more generic fractions with denominator in $\mathcal{H}_m$, which were introduced in Proposition 2.3-3) operate on $\mathcal{E}$ is the following: any $a(s) \in \mathcal{H}_m$ is the holomorphic fraction of two exponential polynomials $n(s)$ and $d(s)$ in $\mathcal{R}_m[s]$. Given $v \in \mathcal{E}$, we can find a function $x \in \mathcal{E}$ such that $d(s)x = v$

by surjectivity of $d(s)$, which is ensured by Proposition 2.3-5. Then let $w = a(s)v \in \mathcal{E}$, which is always well defined by Proposition 2.3-4, be given by $w = n(s)x$. Indeed, since $n(s) = a(s)d(s)$ and convolution is an associative operation, we obtain that $a(s)v = a(s)d(s)x = n(s)x = w$.

Observe that x is not uniquely determined. However, the generic solution $\bar{x} = x + k$, with $k \in \ker_{\mathcal{E}} d(s)$, gives $w = n(s)\bar{x}$ since $n(s)k = 0$ by Proposition 2.3-6.

Example 2.4. Let $a(s) = \frac{1-e^{-s}}{s}$. Actually, $a(s)$ is a holomorphic function, since the unique zero of its denominator $d(s) = s$ is common to its numerator $n(s) = 1-e^{-s}$, and thus it is canceled. Therefore $a(s) \in \mathcal{H}_m$ by definition (4).

The function $a(s)$ acts as follows: for every $v \in \mathcal{E}$, $a(s)v = w$ if and only if there is an $x \in \mathcal{E}$ such that $d(s)x = v$ and $n(s)x = w$. The first equation is equivalent to $d(s)x = sx = \frac{d}{dt}x = v$, i.e., $x(t) = \int_0^t v(\tau)\,d\tau + C$, with C an arbitrary constant. Finally, $w = n(s)x = (1 - e^{-s})x = (1 - \sigma)x$, i.e.,

$$w = \frac{1 - e^{-s}}{s} v \;\Leftrightarrow\; w(t) = \int_{t-1}^{t} v(\tau)\,d\tau.$$

2.4. Bézout equations. Bézout equations constitute a fundamental tool in algebraic systems theory. A finite set of elements $r_1, \dots, r_l$ in a ring $\mathcal{R}$ satisfies a Bézout equation if there are $x_i \in \mathcal{R}$ such that $\sum_1^l x_i r_i = 1$ or, equivalently, if the ideal generated by $r_1, \dots, r_l$ coincides with $\mathcal{R}$. The ring of polynomials in one variable, which is the operator ring employed within the theory of linear differential (or difference) systems, is a Bézout domain, i.e., every finitely generated ideal is principal. In this case polynomials without common zeros always satisfy a Bézout equation and thus constitute a set of generators of the whole ring.

The ring $\mathcal{H}_1$ is a Bézout domain, as it was shown in [9, Thm. 3.2]. This property, along with the fact that every matrix admits the Smith form [9, Thm. 3.5-5], permits to extend many theorems on differential systems to systems with one delay. Also $\mathcal{O}$ is a Bézout domain [12, Thm. 9], while this is not the case for the rings $\mathcal{H}_m$ and $\mathcal{A}$, as we will show in the proposition which follows. An analogous example can be found in [19]. For $\mathcal{H}_m$, a different counterexample is proposed in [11, Ex. 5.13], and Gröbner basis theory is there employed to achieve the claimed result.

We need the following preliminary result.

Lemma 2.5. *Given an element $a(s) \in \mathcal{A}$, the following three facts are equivalent.*

(i) *$a(s)$ is invertible in $\mathcal{A}$.*

(ii) *$a(s)$ has no zeros.*

(iii) *$a(s) = ke^{s\tau}$ for some $\tau, k \in \mathbb{R}$, $k \neq 0$.*

Proof. It is clear that exponentials have no zeros. Moreover, a function in $\mathcal{A}$ with no zeros, being of exponential type by definition (2), has the form $e^{h(s)}$ with $h(s)$ a polynomial of degree at most one by Hadamard's theorem (see [17, p. 24]), and therefore it has inverse $e^{-h(s)}$, which is still in $\mathcal{A}$.

We have to show that every invertible element in $\mathcal{A}$ is an exponential like $ke^{s\tau}$. Since $\mathcal{A} \cong \mathcal{E}'$, this corresponds to prove that the invertible distributions are $k\delta_\tau$, where δ_τ is the Dirac measure at τ ($\langle \delta_\tau, f \rangle = f(\tau)$), whose Laplace transform is indeed $e^{-s\tau}$.

Let $\alpha \in \mathcal{E}'$ be an invertible distribution. Then there is a $\beta \in \mathcal{E}'$ such that $\delta = \alpha\star\beta$. If we denote by $[\alpha]$ the smallest closed interval that contains the support of α, by the Titchmarsh–Lions theorem on supports [18, p. 277], $\{0\} = [\delta] = [\alpha \star \beta] = [\alpha] + [\beta]$. This implies that the supports of α and β have to be $\{\tau\}$ and $\{-\tau\}$ for some $\tau \in \mathbb{R}$. Hence α must be a linear combination of δ_τ and its derivatives [24, Thm. 24.6], i.e., its Laplace transform is $\hat{\alpha}(s) = a(s)e^{-s\tau}$, $a(s) \in \mathbb{R}[s]$. In the same way, $\hat{\beta}(s) = b(s)e^{s\tau}$,

$b(s) \in \mathbb{R}[s]$, and hence $1 = \hat{\alpha}(s)\hat{\beta}(s) = a(s)b(s)$. The proof is complete since the only invertible polynomials are constant. □

PROPOSITION 2.6. *The rings $\mathcal{H}_m$ with $m > 1$ and $\mathcal{A}$ are not Bézout domains.*

Proof. We choose the following two functions in $\mathcal{H}_2 \subseteq \mathcal{A}$:

$$r_1(s) = \frac{1 - e^{-s}}{s} \text{ and } r_2(s) = 1 - e^{-s\tau}. \tag{5}$$

We want to show that, for particular values of $\tau \in \mathbb{R}$, the ideal in $\mathcal{A}$ generated by $r_1(s)$ and $r_2(s)$ is not principal. This would imply also that the ideal in $\mathcal{H}_2$ generated by $r_1(s)$ and $r_2(s)$ cannot be principal.

Let $\tau \notin \mathbb{Q}$. So the functions $r_1(s)$ and $r_2(s)$ have no common zeros. Observe preliminarly that, if the ideal in $\mathcal{A}$ generated by $r_1(s)$ and $r_2(s)$ was principal, then the single generator of this ideal should be an element in $\mathcal{A}$ with no zeros and so invertible by Lemma 2.5. This would imply that $r_1(s)$ and $r_2(s)$ should satisfy a Bézout equation. We show now that this is not possible.

Let τ be a Liouville number, i.e., for all $C \in \mathbb{N}$ there are infinitely many $\frac{n}{d} \in \mathbb{Q}$ such that

$$\left|\tau - \frac{n}{d}\right| \leq d^{-1-C}. \tag{6}$$

Suppose, moreover, that $r_i(s)$ satisfy a Bézout equation in $\mathcal{A}$, i.e., that there exist $a_1(s), a_2(s) \in \mathcal{A}$ such that

$$a_1(s)r_1(s) + a_2(s)r_2(s) = 1. \tag{7}$$

Note that $r_2(s) = -2ie^{-\frac{s\tau}{2}} \sin\left(\frac{is\tau}{2}\right)$. So if we evaluate (7) in $s = 2d\pi i$, with $d \in \mathbb{N}\backslash\{0\}$, since $r_1(2d\pi i) = 0$, we get $a_2(2d\pi i)2ie^{-\tau d\pi i} \sin \tau d\pi = 1$. Using basic trigonometry and relation (6), we have $|\sin \tau d\pi| = |\sin(\tau d\pi - n\pi)| \leq |\tau d\pi - n\pi| \leq \pi d^{-C}$. Therefore,

$$1 = |a_2(2d\pi i)|\,|2ie^{-\tau d\pi i}|\,|\sin \tau d\pi| \leq |a_2(2d\pi i)|2\pi d^{-C}.$$

But $a_2(s) \in \mathcal{A}$ and therefore, by definition (2), simplified since $\operatorname{Re} s = 0$, there exist $A, B > 0$ such that $|a_2(2d\pi i)| \leq A(1 + 2d\pi)^B$. Thus, employing the above equation, we obtain

$$1 \leq |a_2(2d\pi i)|2\pi d^{-C} \leq A2\pi(1 + 2d\pi)^B d^{-C}.$$

If we choose $C > B$ and let $d \to \infty$, we get a contradiction. □

Nevertheless, even if functions in $\mathcal{A}$ that do not have common zeros do not generate the whole ring, the ideal they generate is always dense in $\mathcal{A}$, as the following theorem states.

THEOREM 2.7. *Let $a_i(s) \in \mathcal{A}$, $i = 1, \dots, l$ be Paley–Wiener functions without common zeros. Then the closure of the ideal they generate over $\mathcal{A}$ coincides with $\mathcal{A}$.*

Proof. This theorem is a direct consequence of an important result due to Schwartz, called the spectral analysis theorem [2, Thm. 2.11]: it states that if $a_i(s)$ is the Laplace transform of $\alpha_i \in \mathcal{E}'$, then the unique solution $w \in \mathcal{E}$ of $\alpha_i \star w = 0$ for every $i = 1, \dots, l$ is $w = 0$. Therefore, if $R(s) \in \mathcal{A}^{n \times 1}$ is the column containing $a_i(s)$, by the spectral analysis theorem $\ker_{\mathcal{E}} R(s) = \{0\}$. By Proposition 2.1, $\operatorname{im}_{\mathcal{A}} \circ R(s) = \{0\}^{\perp} = \mathcal{A}$, which is exactly what is claimed. □

2.5. Generalized inverses of matrices. Bézout equations are important for our purposes due to their connection with particular properties of matrices. Actually, it is known that a full row rank matrix is left invertible if and only if its maximal minors satisfy a Bézout equation. This result may be extended to matrices which are not full rank.

To this aim we introduce a terminology which will be used throughout this paper. Given a matrix R of rank r, $r \times r$ minors of R will be called rank minors of R.

Definition 2.8. *Given a domain $\mathcal{S}$ and a matrix $R \in \mathcal{S}^{p\times q}$, we say that $G \in \mathcal{S}^{q\times p}$ is the* 1*-inverse of R if $RGR = R$. It is a* 1, 2*-inverse if it is a* 1*-inverse and $GRG = G$.*

Since a matrix admits a 1-inverse if and only if it admits a $1,2$-inverse (it is easy to check that if G is a 1-inverse, then GRG is a $1,2$-inverse), we will consider only $1,2$-inverses and call them simply generalized inverses.

The following theorem [4, Thm. 8] generalizes the aforementioned equivalence of invertibility and the existence of a Bézout equation to matrices which are not full rank.

Theorem 2.9. *Every matrix with entries in a domain has a generalized inverse if and only if its rank minors satisfy a Bézout equation. The knowledge of the coefficients of the Bézout equation permits us to construct the generalized inverse.*

Remark 2.10. A matrix $R(s)$ with entries in $\mathcal{O}$ admits a generalized inverse in $\mathcal{O}$ if and only if

$$\operatorname{rank}_{\mathbb{C}} R(\lambda) = r \text{ for every } \lambda \in \mathbb{C}.$$

Indeed, this happens if and only if its rank minors have no common zeros. Since $\mathcal{O}$ is a Bézout domain, this is equivalent to satisfy a Bézout equation with coefficients in $\mathcal{O}$.

We report now two other facts on matrices that admit a generalized inverse.

Lemma 2.11. *Let $\mathcal{S}$ be a subring of $\mathcal{O}$. If $R(s) \in \mathcal{S}^{p\times q}$ admits a generalized inverse $G(s) \in \mathcal{S}^{q\times p}$, then*

$$\ker_{\mathcal{S}} \circ R(s) = \operatorname{im}_{\mathcal{S}} \circ (I - R(s)G(s)), \quad \operatorname{im}_{\mathcal{S}} \circ R(s) = \ker_{\mathcal{S}} \circ (I - G(s)R(s)). \tag{8}$$

If, moreover, the elements of $\mathcal{S}$ are operators on $\mathcal{E}$,

$$\operatorname{im}_{\mathcal{E}} R(s) = \ker_{\mathcal{E}} (I - R(s)G(s)), \qquad \ker_{\mathcal{E}} R(s) = \operatorname{im}_{\mathcal{E}} (I - G(s)R(s)). \tag{9}$$

Proof. We prove only (9). The proof of (8) follows similarly. Since $R(s) = R(s)G(s)R(s)$, we have $(I - R(s)G(s))R(s) = 0$ and $R(s)(I - G(s)R(s)) = 0$, and so $\operatorname{im}_{\mathcal{E}} R(s) \subseteq \ker_{\mathcal{E}}(I - R(s)G(s))$ and $\operatorname{im}_{\mathcal{E}}(I - G(s)R(s)) \subseteq \ker_{\mathcal{E}} R(s)$.

Conversely, if $w \in \ker_{\mathcal{E}}(I - R(s)G(s))$, then $0 = (I - R(s)G(s))w = w - R(s)G(s)w$, and so $w = R(s)G(s)w \in \operatorname{im}_{\mathcal{E}} R(s)$. Thus $\ker_{\mathcal{E}}(I - R(s)G(s)) \subseteq \operatorname{im}_{\mathcal{E}} R(s)$. Even more trivially, we obtain that $w \in \ker_{\mathcal{E}} R(s)$ implies $w = (I - G(s)R(s))w$ and therefore $w \in \operatorname{im}_{\mathcal{E}}(I - G(s)R(s))$. □

The following theorem plays a fundamental role in many propositions that we will prove in the following sections.

Theorem 2.12. *Let $\mathcal{S}$ be a subring of $\mathcal{O}$. Suppose that $R(s) \in \mathcal{S}^{p\times q}$ has rank r and admits a generalized inverse over $\mathcal{O}$. Then for some $d \in \mathbb{N}$ there is a matrix $M(s) \in \mathcal{S}^{q\times d}$, with rank $q - r$, such that*

$$\ker_{\mathcal{O}} R(s)\circ = \operatorname{im}_{\mathcal{O}} M(s)\circ \text{ and } \operatorname{im}_{\mathcal{O}} \circ R(s) = \ker_{\mathcal{O}} \circ M(s). \tag{10}$$

Furthermore, $M(s)$ also admits a generalized inverse over $\mathcal{O}$.

Since the proof is quite involved (it needs a rather cumbersome notation), it is postponed in Appendix A. Observe that $M(s)$ can be constructed employing only the rank minors of $R(s)$.

3. Controllability of convolutional systems.

In this section we aim to introduce a set of conditions related to the concept of controllability for the class of dynamical systems described by homogeneous convolutional equations. In the rest of the paper we will show the relations which occur between them.

Definition 3.1. *Assume that $R(s) \in \mathcal{A}^{p\times q}$ is a matrix of Paley–Wiener functions and consider the behavior $\mathcal{B} = \ker_{\mathcal{E}} R(s)$. We will be concerned with the following properties of $\mathcal{B}$.*

SC *$\mathcal{B}$ is* spectrally controllable*: $\mathrm{rank}_{\mathbb{C}}\, R(\lambda) = r$ for every $\lambda \in \mathbb{C}$ or, equivalently, the rank minors of $R(s)$ have no common zeros (see Remark* 2.10*). This is a direct generalization of the PBH controllability test for differential systems.*

BC *$\mathcal{B}$ is controllable in the behavioral sense of Definition* 1.1.

IR *$\mathcal{B}$ admits an* image representation*: $\mathcal{B} = \mathrm{im}_{\mathcal{E}}\, M(s)$ for some matrix of operators $M(s) \in \mathcal{A}^{q\times d}$.*

DCS *The subset of $\mathcal{B}$ consisting of the trajectories having* compact support*,i.e., $\mathcal{B}_{cs} \triangleq \mathcal{B} \cap \mathcal{D}^q = \ker_{\mathcal{D}} R(s)$ is* dense *in $\mathcal{B}$: $\mathcal{B} = \overline{\mathcal{B}_{cs}}$.*

DIR *There is a* dense image representation *of $\mathcal{B}$: $\mathcal{B} = \overline{\mathrm{im}_{\mathcal{E}}\, M(s)}$ for some matrix of operators $M(s) \in \mathcal{A}^{q\times d}$.*

GI *There exists a* generalized inverse *of $R(s)$: there is a matrix $G(s) \in \mathcal{A}^{q\times p}$ such that $R(s)G(s)R(s) = R(s)$.*

Note that **SC** and **GI** can be reformulated within the module theoretic framework proposed by Fliess and Mounier [8, 20]. Actually, spectral controllability and existence of a generalized inverse are “algebraic,” while the other are more “analytic” conditions.

The relation between spectral controllability of $\mathcal{B} = \ker_{\mathcal{E}} R(s)$, existence of a generalized inverse of $R(s)$ and null controllability, i.e., the possibility to steer to zero every trajectory of $\mathcal{B}$, was investigated in [21] within a different framework.

Remark 3.2. Definition 3.1 is stated for general behaviors which are kernels of convolutional operators. Matrices over the ring $\mathcal{A}$ are more difficult to handle than matrices over the delay-differential operator ring $\mathcal{R}_m[s]$ or its “fraction” ring $\mathcal{H}_m$. However, this general framework is necessary to obtain some important results, even if the starting point is a matrix $R(s)$, which has delay-differential entries.

Indeed, theorems establishing the existence of an image representation in $\mathcal{A}$ can be used as a preliminary step for obtaining an image representation in $\mathcal{R}_m[s]$ or $\mathcal{H}_m$. This happens when $\mathcal{B}$ is the kernel of a matrix $R(s)$ with entries in $\mathcal{H}_m$, as stated in the following proposition.

Proposition 3.3. *Let $R(s) \in \mathcal{H}_m^{p\times q}$ have rank r. Suppose that $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ admits a (dense) image representation. Then it admits a (dense) image representation in $\mathcal{R}_m[s]$, i.e., there exists a full column rank matrix $M(s) \in \mathcal{R}_m[s]^{q\times(q-r)}$ such that $\mathcal{B}$ is the (closure of the) image of $M(s)$.*

Proof. Let us suppose that $\mathcal{B} = \overline{\mathrm{im}_{\mathcal{E}}\, N(s)}$, where $N(s) \in \mathcal{A}^{q\times d}$. As we shall show in Theorem 3.5, this assumption implies that $\mathcal{B}$ is spectrally controllable and therefore, by Theorem 2.12, we can construct $L(s) \in \mathcal{H}_m^{q\times e}$ such that $\ker_{\mathcal{O}} R(s)^{\circ} = \mathrm{im}_{\mathcal{O}}\, L(s)^{\circ}$. Since $R(s)N(s) = 0$, the columns of $N(s)$ belong to the image over $\mathcal{O}$ of $L(s)$, i.e., there exists an $X(s) \in \mathcal{O}^{e\times d}$ such that $N(s) = L(s)X(s)$.

Furthermore, if we denote by $\mathcal{K}_m$ the field of fractions of $\mathcal{H}_m$, by standard linear algebra arguments there exists a basis of $\ker_{\mathcal{K}_m} R(s)^{\circ}$, which means that there exists a full column rank matrix $\bar{M}(s) \in \mathcal{K}_m^{q\times(q-r)}$ such that $\ker_{\mathcal{K}_m} R(s)\circ = \operatorname{im}_{\mathcal{K}_m} \bar{M}(s)\circ$. Since also $R(s)L(s) = 0$, $L(s) = \bar{M}(s)\bar{Y}(s)$, where $\bar{Y}(s) \in \mathcal{K}_m^{(q-r)\times e}$, and, multiplying both members by an appropriate element $l(s) \in \mathcal{H}_m$, we get the relation

$$l(s)L(s) = M(s)Y(s),$$

where both $M(s)$ and $Y(s)$ have entries in $\mathcal{R}_m[s]$.

Since $\mathcal{K}_m$ is necessarily a Bézout domain, by Theorem 2.9 we can find a generalized inverse $\bar{G}(s) \in \mathcal{K}_m^{e\times q}$ of $L(s)$ such that $L(s)\bar{G}(s)L(s) = L(s)$ or, multiplying by a suitable $g(s) \in \mathcal{H}_m$, a matrix $G(s) \in \mathcal{H}_m^{e\times q}$ such that

$$L(s)G(s)L(s) = g(s)L(s).$$

So, putting together the equations, we found

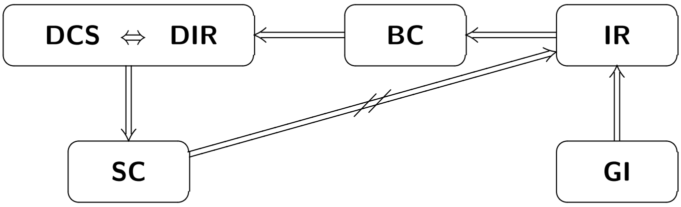
$$l(s)g(s)N(s) = l(s)g(s)L(s)X(s) = l(s)L(s)G(s)L(s)X(s) = M(s)Y(s)G(s)N(s).$$

Since $l(s)g(s) \in \mathcal{H}_m$ is a surjective operator by Proposition 2.3-5, we obtain that

$$\begin{aligned} \operatorname{im}_{\mathcal{E}} N(s) &= \operatorname{im}_{\mathcal{E}} l(s)g(s)N(s) = \operatorname{im}_{\mathcal{E}} M(s)Y(s)G(s)N(s) \\ &\subseteq \operatorname{im}_{\mathcal{E}} M(s) \subseteq \ker_{\mathcal{E}} R(s) = \overline{\operatorname{im}_{\mathcal{E}} N(s)}. \end{aligned}$$

So, if $\operatorname{im}_{\mathcal{E}} N(s)$ is closed, it is equal to $\operatorname{im}_{\mathcal{E}} M(s)$. In any case, the closures of these sets are equal. $\square$

3.1. The general case. The true and false implications described in the following scheme hold without any further assumption.



As our first step, we prove that **SC** $\not\Rightarrow$ **IR**, which shows that the properties we listed above are not equivalent in general. In particular, note that the following example shows that Theorem 1.3 cannot be generalized to $\mathcal{H}_m$ and to $\mathcal{A}$.

In fact, we will present a matrix $R(s)$ with entries in $\mathcal{H}_m$ such that $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ is spectrally controllable, i.e., $\operatorname{rank}_{\mathbb{C}} R(\lambda)$ does not depend on $\lambda \in \mathbb{C}$, but $\mathcal{B}$ does not admit an image representation, i.e., there is no matrix $M(s)$ with entries in $\mathcal{A}$ such that $\mathcal{B} = \operatorname{im}_{\mathcal{E}} M(s)$.

Example 3.4. Let $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ be the kernel representation of the behavior $\mathcal{B}$, where the entries of $R(s)$ were defined in Proposition 2.6.

$$R(s) \triangleq [r_2(s) \;\; -r_1(s)] = \left[1 - e^{-s\tau} \quad -\frac{1-e^{-s}}{s} \right]. \tag{11}$$

As we have already shown in Proposition 2.6, if τ is a Liouville number, $\mathcal{B}$ is spectrally controllable, but $r_1(s)$ and $r_2(s)$ do not satisfy any Bézout equation. Assume by contradiction that there exists $N(s) = [n_{ij}(s)] \in \mathcal{A}^{2\times d}$ such that $\ker_{\mathcal{E}} R(s) = \operatorname{im}_{\mathcal{E}} N(s)$. Since $R(s)N(s) = 0$, it must be

$$r_1(s)n_{2j}(s) = r_2(s)n_{1j}(s) \text{ for every } j = 1, \dots, d. \tag{12}$$

The above equation shows that, if s_0 is a zero of $r_i(s)$ with multiplicity μ, then it is also a zero of $n_{ij}(s)$ with multiplicity greater than or equal to μ. Actually, this is a consequence of the **SC** condition: we know that if $r_1(s_0) = 0$ for some $s_0 \in \mathbb{C}$, then $r_2(s_0) \neq 0$ and so $n_{1j}(s_0) = 0$. Therefore,

$$y_j(s) \triangleq \frac{n_{1j}(s)}{r_1(s)} = \frac{n_{2j}(s)}{r_2(s)} \in \mathcal{O}.$$

By Proposition 2.3-3, $y_j(s) \in \mathcal{A}$ and, if we let $M(s) \triangleq [r_1(s)\ r_2(s)]^\top$ and $Y(s) \triangleq [y_j(s)] \in \mathcal{A}^d$, then $R(s)M(s) = 0$ and $N(s) = M(s)Y(s)$. This implies that

$$\ker_{\mathcal{E}} R(s) = \operatorname{im}_{\mathcal{E}} N(s) = \operatorname{im}_{\mathcal{E}} M(s)Y(s) \subseteq \operatorname{im}_{\mathcal{E}} M(s) \subseteq \ker_{\mathcal{E}} R(s).$$

This equation shows that $\operatorname{im}_{\mathcal{E}} M(s)$ is closed, being equal to $\ker_{\mathcal{E}} R(s)$, and therefore, by Proposition 2.2, $\operatorname{im}_{\mathcal{A}} \circ M(s)$ is closed also. But this is the ideal generated by $r_1(s)$ and $r_2(s)$ and hence, by Theorem 2.7, $\operatorname{im}_{\mathcal{A}} \circ M(s) = \mathcal{A}$, i.e., there are two elements $a_i(s) \in \mathcal{A}$ such that $a_1(s)r_1(s) + a_2(s)r_2(s) = 1$, which is in contradiction with what was obtained in Proposition 2.6.

We present now the main theorem of this section in which we prove the implications shown in the previous scheme.

THEOREM 3.5. *Given any behavior in kernel representation $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ with $R(s) \in \mathcal{A}^{p\times q}$, the following chain of implications always holds:* **GI** $\Rightarrow$ **IR** $\Rightarrow$ **BC** $\Rightarrow$ **DCS** $\Leftrightarrow$ **DIR** $\Rightarrow$ **SC**.

Proof. We will prove each implication separately.

GI $\Rightarrow$ **IR**. It is a direct consequence of Lemma 2.11.

IR $\Rightarrow$ **BC**. If $\mathcal{B} = \operatorname{im}_{\mathcal{E}} M(s)$, then every $w \in \mathcal{B}$ is the image of some smooth function $v \in \mathcal{E}^d$, i.e., $w = M(s)v$. Since $M(s)$ is the Laplace transform of a distribution with compact support, $w(t)$ does not depend on $v(\tau)$ if $\tau \notin [t+a, t+b]$ for some $a \leq b$, both depending only on $M(s)$.

Therefore, if we choose a $T > b - a$, we have that $t + T + a > t + b$. Thus, if $w_i = M(s)v_i$ for $i = 1, 2$ and we define $v(\tau) \triangleq v_1(\tau)$ for $\tau < b$, $v(\tau) \triangleq v_2(\tau - T)$ for $\tau > T+a$ and completing $v(\tau)$ smoothly in the interval $\tau \in [b, T-a]$, then $w = M(s)v$ is such that $w \in \mathcal{B}$, $w(t) = w_1(t)$ as $t \leq 0$ and $w(t) = w_2(t - T)$ as $t > T$.

BC $\Rightarrow$ **DCS**. First, we prove that if $\mathcal{B}$ is controllable, it is possible to steer every trajectory to zero also in the past, i.e. for every $w \in \mathcal{B}$ there exists a $\tilde{w} \in \mathcal{B}$ and a $T > 0$ such that

$$\tilde{w}(t) = 0 \ \ \forall t < -T \text{ and } \tilde{w}(t) = w(t) \ \ \forall t > 0.$$

Actually, given $w_1 = 0$ and $w_2 = w$, by Definition 1.1 there is a $\tau > 0$ and a $\bar{w} \in \mathcal{B}$ such that $\bar{w}(t) = 0$ for $t \leq 0$ and $\bar{w}(t) = \sigma_\tau w(t)$ for $t \geq \tau$. If we take $T \triangleq \tau$, then $\tilde{w} \triangleq \sigma_{-\tau}\bar{w}$ is the desired function.

Next, given any $w \in \mathcal{B}$, consider an increasing sequence of compact intervals

$$[t_i, \tau_i] = K_i \subset K_{i+1} \text{ such that } \cup_i K_i = \mathbb{R}.$$

For every i we can find a trajectory $u_i \in \mathcal{B}$ that is equal to $w(t)$ for $t \geq t_i$ and zero in the "past"; we can also find a trajectory $v_i \in \mathcal{B}$ that is equal to $u_i(t)$ for $t \leq \tau_i$ and zero in the "future." Clearly,

$$v_i \in \mathcal{B}_{\text{cs}} \text{ and } v_i(t) = w(t) \text{ for every } t \in K_i.$$

The sequence v_i converges to w in the topology of $\mathcal{E}$ and therefore w is a limit point of $\mathcal{B}_{\mathrm{cs}}$. Therefore $\mathcal{B} \subseteq \overline{\mathcal{B}_{\mathrm{cs}}}$.

The converse inclusion is easier to establish: since $\mathcal{B}$ is closed and $\mathcal{B}_{\mathrm{cs}} \subseteq \mathcal{B}$, $\overline{\mathcal{B}_{\mathrm{cs}}} \subseteq \mathcal{B}$ also.

DCS $\Rightarrow$ **SC**. It is well known that matrices of holomorphic functions admit the Smith form [13]. A direct consequence of this fact (see, e.g., [25, Ch. 4.1.1]) is the following: since $R(s)$ admits a Smith form, then there exist two matrices $U_1(s) \in \mathcal{O}^{r\times p}$ and $U_2(s) \in \mathcal{O}^{(p-r)\times p}$ such that $U(s) \triangleq \left[\begin{smallmatrix} U_1(s) \\ U_2(s)\end{smallmatrix}\right] \in \mathcal{O}^{p\times p}$ is a square and invertible matrix in $\mathcal{O}$. Thus

$$\operatorname{im}_{\mathbb{C}} \circ U_1(s_0) \oplus \operatorname{im}_{\mathbb{C}} \circ U_2(s_0) = \mathbb{C}^p \text{ for every } s_0 \in \mathbb{C}, \tag{13}$$

and, moreover,

$$\ker_{\mathcal{O}} \circ R(s) = \operatorname{im}_{\mathcal{O}} \circ U_2(s) = \mathcal{O}^{p-r} U_2(s). \tag{14}$$

Note that the dimension of $\ker_{\mathbb{C}} \circ R(s_0)$ is greater than or equal to the dimension of $\ker_{\mathcal{O}} \circ R(s)$. In particular, this inequality holds in a strict sense if and only if $\operatorname{rank}_{\mathbb{C}} R(s_0) < \operatorname{rank}_{\mathcal{O}} R(s)$, i.e., if the behavior $\mathcal{B}$ is not spectrally controllable. Therefore, considering (13) and (14), we can derive the following equivalence:

$$\textbf{SC} \Leftrightarrow \ker_{\mathbb{C}} \circ R(s_0) \cap \operatorname{im}_{\mathbb{C}} \circ U_1(s_0) = \{0\} \text{ for every } s_0 \in \mathbb{C}. \tag{15}$$

We want to prove that **DCS** implies the second condition in (15): let us take $s_0 \in \mathbb{C}$ and

$$c \in \ker_{\mathbb{C}} \circ R(s_0) \cap \operatorname{im}_{\mathbb{C}} \circ U_1(s_0) \tag{16}$$

and construct the operator (well defined by Proposition 2.3-3 because $cR(s_0) = 0$)

$$a(s) \triangleq \frac{1}{s - s_0} cR(s) \in \mathcal{A}^q.$$

We aim to show that $\mathcal{B} \subseteq \ker_{\mathcal{E}} a(s)$.

Indeed, take a $w \in \mathcal{B}$. The way in which $a(s)$ operates implies that if $v = a(s)w$, then $(s - s_0)v = cR(s)w = 0$. This means that $\frac{d}{dt}v = s_0 v$, i.e., $v(t) = v(0)e^{s_0 t}$. Consequently, the image $a(\mathcal{B})$ of $\mathcal{B}$ through $a(s)$ consists only in exponentials.

If we take $w \in \mathcal{B}_{\mathrm{cs}} \subseteq \mathcal{B}$ and $v = a(s)w$, then $v(\tau) = 0$ for $|\tau|$ sufficiently large, since it is the convolution between a distribution and a function both with compact support. Since v is an exponential, it must be the zero function. Employing the hypothesis that $\mathcal{B} = \overline{\mathcal{B}_{\mathrm{cs}}}$, by (18) we can argue that

$$a(\mathcal{B}) = a(\overline{\mathcal{B}_{\mathrm{cs}}}) \subseteq \overline{a(\mathcal{B}_{\mathrm{cs}})} = \overline{\{0\}} = \{0\} \Rightarrow \mathcal{B} = \ker_{\mathcal{E}} R(s) \subseteq \ker_{\mathcal{E}} a(s).$$

This result implies, by a theorem of Malgrange [18, p. 282], that there exists an $x(s) \in \mathcal{O}^p$ such that $a(s) = x(s)R(s)$. By definition of $a(s)$, we also have that $(s - s_0)a(s) = cR(s)$ and hence $(c - (s - s_0)x(s))R(s) = 0$. This implies, by (14) that there exists an $y(s) \in \mathcal{O}^{p-r}$ such that $c - (s - s_0)x(s) = y(s)U_2(s)$.

We deduce that $c = y(s_0)U_2(s_0)$ and so $c \in \operatorname{im}_{\mathbb{C}} \circ U_2(s_0)$. But, as assumed in (16), we also have $c \in \operatorname{im}_{\mathbb{C}} \circ U_1(s_0)$, and hence, by (13), we can argue that $c = 0$. By (15) this implies that the system is spectrally controllable.

DCS $\Rightarrow$ **DIR**. Condition **DCS** implies **SC** and therefore, by theorem 2.12, there exists a matrix $M(s) \in \mathcal{A}^{q\times d}$ with rank $q - r$ such that $R(s)M(s) = 0$. We shall show that $\mathcal{B} = \overline{\mathrm{im}_{\mathcal{E}}\, M(s)}$ or, equivalently, by Proposition 2.1, that $\overline{\mathrm{im}_{\mathcal{A}} \circ R(s)} = \mathcal{B}^{\perp} = \ker_{\mathcal{A}} \circ M(s)$. Since the inclusion $\mathcal{B}^{\perp} \subseteq \ker_{\mathcal{A}} \circ M(s)$ is straightforward, we need to prove only the other one.

First observe that, using hypothesis **DCS** and [24, Cor. 1, 2, p. 363], we obtain that $\mathcal{B}^{\perp} = (\overline{\mathcal{B}_{\mathrm{cs}}})^{\perp} = (\mathcal{B}_{\mathrm{cs}}^{\perp\perp})^{\perp} = \mathcal{B}_{\mathrm{cs}}^{\perp}$. Thus, our aim is to prove that $\ker_{\mathcal{A}} \circ M(s) \subseteq \mathcal{B}_{\mathrm{cs}}^{\perp}$.

Let $x(s) \in \ker_{\mathcal{A}} \circ M(s)$. Denoting by $\mathcal{F}$ the fraction field of $\mathcal{A}$, we have that the subspaces $\mathrm{im}_{\mathcal{F}} \circ R(s) \subseteq \ker_{\mathcal{F}} \circ M(s)$ have the same dimension r and thus coincide. Since $x(s) \in \ker_{\mathcal{F}} \circ M(s)$, there exists a $\bar{y}(s) \in \mathcal{F}^p$ such that $x(s) = \bar{y}(s)R(s)$. Since $\bar{y}(s)$ is a quotient of elements in $\mathcal{A}$, there must be an $a(s) \in \mathcal{A}$ such that $y(s) = a(s)\bar{y}(s) \in \mathcal{A}^p$, and so $a(s)x(s) = y(s)R(s)$.

As an immediate consequence, for every $w \in \mathcal{B}_{\mathrm{cs}}$ we have

$$a(s)x(s)w = y(s)R(s)w = 0. \tag{17}$$

But w has compact support, and hence it admits that the Laplace transform $\hat{w}(s)$ and the convolutional equation (17) can be rewritten as product of Laplace transforms:

$$a(s)x(s)\hat{w}(s) = 0.$$

Since $\mathcal{A}$ is a domain, the last equation holds if and only if $x(s)\hat{w}(s) = 0$, and this implies that $x(s) \in \mathcal{B}_{\mathrm{cs}}^{\perp}$.

DIR ⇒ **DCS**. The only difficult thing to prove is that $\mathcal{B} \subseteq \overline{\mathcal{B}_{\mathrm{cs}}}$. One of the equivalent definitions of continuity is the following [5, Thm. III.8.3]. A function $f : V \to W$ is continuous if and only if for every subset $U \subseteq V$,

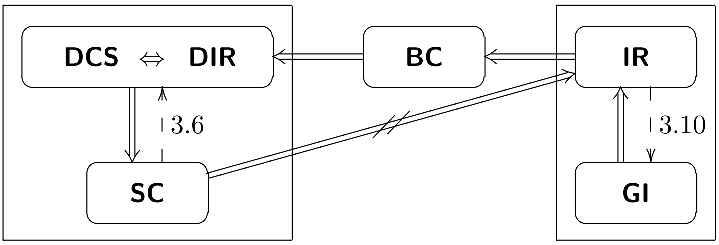
$$\mathrm{im}_{\overline{U}}\, f \subseteq \overline{\mathrm{im}_U\, f}. \tag{18}$$

Since $\mathcal{D}$ is dense in $\mathcal{E}$, then $\mathrm{im}_{\mathcal{E}}\, M(s) = \mathrm{im}_{\overline{\mathcal{D}}}\, M(s) \subseteq \overline{\mathrm{im}_{\mathcal{D}}\, M(s)}$.

We note now that $\mathrm{im}_{\mathcal{D}}\, M(s) \subseteq \ker_{\mathcal{D}} R(s)$. Therefore, taking the closure of the equation written above, we get

$$\mathcal{B} = \overline{\mathrm{im}_{\mathcal{E}}\, M(s)} \subseteq \overline{\mathrm{im}_{\mathcal{D}}\, M(s)} \subseteq \overline{\ker_{\mathcal{D}} R(s)} = \overline{\mathcal{B}_{\mathrm{cs}}}. \qquad \square$$

3.2. The full row rank case: Regular behaviors. In this section we will show that, under a rather mild assumption, conditions **SC**, **DCS**, and **DIR** are equivalent and, adding another stronger hypothesis, even the equivalence **IR** ⇔ **GI** holds true. The situation is presented in the following picture. The numbers over the dashed lines refer to the theorems that prove the equivalences of the conditions in the boxed regions and that also contain the assumptions under which such equivalences hold.



Before going into the detailed analysis of these facts, it is useful to say something more about the *density* conditions **DCS** and **DIR**.

Kernels of continuous operators are always closed, if we do not consider pathological topological spaces, while images need not be closed. Actually, even the image $\operatorname{im}_{\mathcal{E}} a(s)$ of an element $a(s) \in \mathcal{A}$ may be not closed (see [2, Thm. 2.7, 2.8]). Observe that this cannot occur if $a(s) \in \mathcal{H}_m$, since in this case $a(s)$ is surjective by Proposition 2.3-5. However, matrices with entries in $\mathcal{H}_m$ may have an image which is not closed. This can be seen taking the matrix $M(s) \in \mathcal{H}_2^{2\times 1}$ which has been defined in Example 3.4.

As suggested in [22, p. 208], we could define directly the image representation of a behavior as the closure of the image of a convolutional operator. This would make sense, especially from a practical point of view, because it is difficult to understand the concrete relevance of the system trajectories which belong to the closure of the image but not to the image itself. However, employing this definition and therefore identifying conditions **IR** and **DIR**, the existence of an image representation would have in general no relation at all with behavioral controllability, as far as we know.

The hypothesis we are assuming in this section is that $\mathcal{B} = \ker_{\mathcal{E}} R(s)$, where we suppose that the matrix $R(s) \in \mathcal{A}^{p\times q}$ has full row rank, i.e., that the convolutional equations that constitute the homogeneous system whose set of solutions is $\mathcal{B}$, are linearly independent. In this case the behavior is called *regular*. This is, in general, a weak assumption, and if $R(s)$ is in $\mathcal{H}_1$, it is known [9] that there always exists a full row rank matrix whose kernel is $\mathcal{B}$ or, in other words, the behavior of a delay-differential system with commensurate delays is always regular. However, it is unknown whether this property continues to hold for delay-differential systems with noncommensurate delays.

THEOREM 3.6. *Let $R(s) \in \mathcal{A}^{p\times q}$ be full row rank and $\mathcal{B} = \ker_{\mathcal{E}} R(s)$. If $\mathcal{B}$ is spectrally controllable, then there exists a matrix $M(s) \in \mathcal{A}^{q\times d}$ such that $\mathcal{B} = \overline{\operatorname{im}_{\mathcal{E}} M(s)}$.*

Proof. By Theorem 2.12 we know that there is a matrix $M(s) \in \mathcal{A}^{q\times d}$ such that

$$\operatorname{im}_{\mathcal{O}} \circ R(s) = \ker_{\mathcal{O}} \circ M(s). \tag{19}$$

Since this implies that $R(s)M(s) = 0$, we have that also

$$\operatorname{im}_{\mathcal{A}} \circ R(s) \subseteq \ker_{\mathcal{A}} \circ M(s). \tag{20}$$

We want to prove that $\overline{\operatorname{im}_{\mathcal{A}} \circ R(s)} \supseteq \ker_{\mathcal{A}} \circ M(s)$.

Let $x(s) \in \ker_{\mathcal{A}} \circ M(s)$. By (19) there is a $y(s) \in \mathcal{O}^p$ such that

$$x(s) = y(s)R(s). \tag{21}$$

Let us suppose, without loss of generality, that $R(s) = [R_1(s)\ R_2(s)]$, where $R_1(s)$ is a square full rank submatrix of $R(s)$. Let $\operatorname{adj} R_1(s)$ be the adjoint matrix of $R_1(s)$ and let $x(s)$ be partitioned in the same way as $R(s)$ so that $x_1(s) = y(s)R_1(s)$. We obtain

$$x_1(s) \operatorname{adj} R_1(s) = y(s)R_1(s) \operatorname{adj} R_1(s) = y(s) \det R_1(s). \tag{22}$$

We need now the following important result [18, p. 308]. If we define

$$H(d) \triangleq \left\{ a(s) \in \mathcal{A} \text{ such that } a(s)\frac{n(s)}{d(s)} \in \mathcal{A}, \text{ whenever } \frac{n(s)}{d(s)} \in \mathcal{O} \text{ and } n(s) \in \mathcal{A} \right\}, \tag{23}$$

then for any nonzero Paley–Wiener function $d(s) \in \mathcal{A}$ we have that

$$\overline{H(d)} = \mathcal{A}. \tag{24}$$

In other words, in general, if $n(s)$ and $d(s)$ are Paley–Wiener functions and $h(s) \triangleq n(s)/d(s)$ is holomorphic, then $h(s)$ does not necessarily belong to $\mathcal{A}$. However, even if $h(s) \notin \mathcal{A}$, the functions $a(s) \in H(d)$, as, for instance, $d(s)$ itself, are such that $a(s)h(s) \in \mathcal{A}$. Equation (24) tells us that the set $H(d)$ of functions $a(s)$ that map, by multiplication, *every* holomorphic fraction with denominator $d(s)$ into $\mathcal{A}$, is not only nontrivial, but also dense in $\mathcal{A}$.

If we let $d(s) \triangleq \det R_1(s)$ and $n(s) \triangleq x_1(s) \operatorname{adj} R_1(s) \in \mathcal{A}^p$, then (22) becomes

$$n(s) = y(s)d(s) \tag{25}$$

and so $n(s)/d(s) \in \mathcal{O}^p$. Thus, by definition (23) of $H(d)$, we have that

$$a(s) \in H(d) \;\Rightarrow\; a(s)\frac{n(s)}{d(s)} \in \mathcal{A}^p. \tag{26}$$

Although $H(d)$ was defined only for scalar equations, we remark that $H(d)$ depends only on $d(s)$, and hence (26) is true componentwise.

Since $H(d)$ is dense in $\mathcal{A}$, by (24) there must be a sequence $a_n(s) \in H(d)$ converging to $1 \in \mathcal{A}$. From (25) and (26) we obtain that

$$a_n(s)\frac{n(s)}{d(s)} = a_n(s)y(s) \in \mathcal{A}^p.$$

If we multiply by $a_n(s)$ both members of (21), we obtain

$$a_n(s)x(s) = a_n(s)y(s)R(s) \in \mathcal{A}^p R(s).$$

Therefore, $a_n(s)x(s)$ is a sequence in $\mathcal{A}^p R(s)$ and its limit lies in the closure of this set. Since $a_n(s)$ converges to 1, $a_n(s)x(s) \to x(s) \in \overline{\mathcal{A}^p R(s)}$. In other words,

$$\ker_{\mathcal{A}} \circ M(s) \subseteq \overline{\operatorname{im}_{\mathcal{A}} \circ R(s)}.$$

Considering also (20), if we take the orthogonals that invert the inclusions [24, p. 195], we obtain

$$\overline{\operatorname{im}_{\mathcal{A}} \circ R(s)}^{\perp} \subseteq \ker_{\mathcal{A}} \circ M(s)^{\perp} \subseteq \operatorname{im}_{\mathcal{A}} \circ R(s)^{\perp}.$$

By Proposition 2.1, $\overline{\operatorname{im}_{\mathcal{A}} \circ R(s)}^{\perp} = \ker_{\mathcal{E}} R(s)^{\perp\perp} = \operatorname{im}_{\mathcal{A}} \circ R(s)^{\perp\perp\perp}$, and the last term is equal [24, Cor. 1, p. 363] to $\operatorname{im}_{\mathcal{A}} \circ R(s)^{\perp} = \ker_{\mathcal{E}} R(s)$. Therefore,

$$\mathcal{B} = \ker_{\mathcal{E}} R(s) = \ker_{\mathcal{A}} \circ M(s)^{\perp} = \overline{\operatorname{im}_{\mathcal{E}} M(s)}. \qquad \square$$

3.2.1. Single input behaviors. In this section we concentrate our attention on the relation between conditions **IR** and **GI**, i.e., between the existence of an image representation and the existence of a generalized inverse of the matrix providing the kernel representation. We have already shown that **GI** ⇒ **IR** and that there exists an image representation in $\mathcal{H}_m$ whenever $R(s) \in \mathcal{H}_m^{p\times q}$.

In this section we will show that also the converse is true under a rank condition that can be better understood if we introduce the concept of input/output representation of a behavior.

DEFINITION 3.7. *An input/output representation of a behavior $\mathcal{B} \subseteq \mathcal{E}^q$ is a partition of its variables $w \in \mathcal{B}$ into input variables $u \in \mathcal{E}^m$, $m = q - r$, and output variables $y \in \mathcal{E}^r$, such that the input is the following.*

Free. *For every $u \in \mathcal{E}^m$ there is a $y \in \mathcal{E}^r$ such that the trajectory w consisting of y and u is in $\mathcal{B}$.*

Maximal. *Once u has been fixed, y does not contain other free variables.*

Note that we are disregarding properness issues for the sake of simplicity. For more details, see [22, ch. 3.3].

In general, a behavior $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ with $R(s) \in \mathcal{A}^{p\times q}$ may not admit an input/output representation. However, delay-differential systems always have such a representation.

THEOREM 3.8. *Let $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ with $R(s) \in \mathcal{H}_m^{p\times q}$ with rank r. Then $\mathcal{B}$ admits an input/output representation with r outputs and $q - r$ inputs.*

Proof. After a suitable permutation of its columns, we can partition $R(s)$ as

$$R(s) = [P(s) \ \ -Q(s)],$$

and $w \in \mathcal{B}$ consistently in the following way: $w^\top = [y^\top \ u^\top]$. So $P(s) \in \mathcal{H}_m^{p\times r}$ is a full column rank matrix. The description of $\mathcal{B}$ becomes

$$w \in \mathcal{B} \ \Leftrightarrow \ R(s)w = 0 \ \Leftrightarrow \ P(s)y = Q(s)u. \tag{27}$$

We prove that this is an input/output representation. Indeed, there is a matrix $C(s) \in \mathcal{H}_m^{r\times p}$ such that

$$C(s)P(s) = r(s)I, \tag{28}$$

where $r(s) \in \mathcal{H}_m$ is nonzero. Moreover, by hypothesis every column of $Q(s)$ is not independent from the columns of $P(s)$. This means that there is a scalar $a(s) \in \mathcal{H}_m$ and a matrix $F(s) \in \mathcal{H}_m^{r\times(q-r)}$ such that $a(s)Q(s) = P(s)F(s)$. This relation implies that, since $\mathcal{H}_m$ is a domain, $\ker_{\mathcal{H}_m} \circ P(s) \subseteq \ker_{\mathcal{H}_m} \circ Q(s)$. Therefore, if we multiply (28) by $P(s)$ on the left, we get

$$P(s)C(s)P(s) - P(s)r(s) = (P(s)C(s) - r(s)I)P(s) = 0,$$

and so even $(P(s)C(s) - r(s)I)Q(s) = 0$. Thus

$$P(s)C(s)Q(s) = Q(s)r(s). \tag{29}$$

Now, to see that (27) is an input/output representation, we have to show first of all that the input u is free. Fix any $u \in \mathcal{E}^{q-r}$. Since $r(s)$ is surjective on $\mathcal{E}$ by Proposition 2.3-5, there exists $v \in \mathcal{E}^{q-r}$ such that $r(s)v = u$, and so, if we let $y \triangleq C(s)Q(s)v$ and use identity (29), we obtain that

$$P(s)y = P(s)C(s)Q(s)v = r(s)Q(s)v = Q(s)u.$$

Finally, fixing $u = 0$, $P(s)y = 0$ implies that $C(s)P(s)y = r(s)y = 0$, and hence y is not free since every component is a solution of the same delay-differential equation. □

Remark 3.9. Theorem 3.8 holds more generally for a behavior $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ with $R(s) \in \mathcal{A}^{p\times q}$, under the additional assumption that at least one of its rank minors belongs to $\mathcal{H}_m$.

The following theorem shows in which case the existence of an image representation of $\ker_{\mathcal{E}} R(s)$ implies the existence of a generalized inverse of $R(s)$ over $\mathcal{A}$.

THEOREM 3.10. *Let $R(s) \in \mathcal{H}_m^{p\times q}$ and suppose that $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ is a regular behavior with a single input. Then, if $\mathcal{B}$ admits an image representation, $R(s)$ admits a generalized inverse.*

Proof. By Theorem 2.12 and (38), there is a column $M(s) \in \mathcal{H}_m^{q\times 1}$ that is constructed with the q rank minors of $R(s)$, such that

$$\ker_{\mathcal{O}} R(s)\circ = \operatorname{im}_{\mathcal{O}} M(s)\circ.$$

If we show that $1 \in \operatorname{im}_{\mathcal{A}} \circ M(s)$, we have proved that the rank minors of $R(s)$ satisfy a Bézout equation over $\mathcal{A}$ and, by Theorem 2.9, $R(s)$ admits a generalized inverse.

Since the behavior admits an image representation, say $\mathcal{B} = \operatorname{im}_{\mathcal{E}} N(s)$ with $N(s) \in \mathcal{A}^{q\times d}$, then $R(s)N(s) = 0$. Thus each row of $N(s)$ is in $\ker_{\mathcal{O}} R(s)\circ$, and so we have that

$$\exists Y(s) \in \mathcal{O}^{1\times d} : \ M(s)Y(s) = N(s) \ \Rightarrow \ m_i(s)y_j(s) = n_{ij}(s).$$

By Proposition 2.3-3 $n_{ij}(s)/m_i(s) = y_j(s) \in \mathcal{O}$ is a Paley–Wiener function and so $Y(s) \in \mathcal{A}^{1\times d}$. Now, since $M(s)Y(s) = N(s)$, where every matrix is an operator, $\operatorname{im}_{\mathcal{E}} N(s) \subseteq \operatorname{im}_{\mathcal{E}} M(s)$. Therefore,

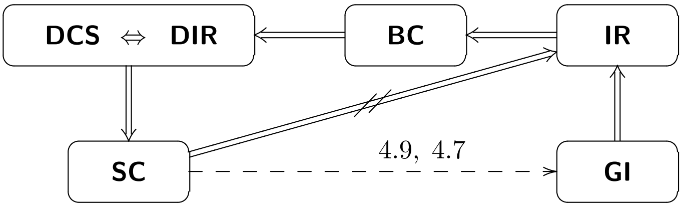
$$\ker_{\mathcal{E}} R(s) = \operatorname{im}_{\mathcal{E}} N(s) \subseteq \operatorname{im}_{\mathcal{E}} M(s) \subseteq \ker_{\mathcal{E}} R(s),$$

so $\operatorname{im}_{\mathcal{E}} M(s)$ is closed and, by Propositions 2.1 and 2.2, $\operatorname{im}_{\mathcal{A}} \circ M(s)$ is also closed.

We know that the rank minors of $R(s)$ that are the elements of $M(s)$ have no common zeros, and thus by Theorem 2.7 $\operatorname{im}_{\mathcal{A}} \circ M(s) = \mathcal{A}$. □

4. Controllability of multidelay differential systems. In this section we will show that the equivalence between all the conditions introduced in Definition 3.1 does hold if we restrict our attention to a particular class of delay-differential systems that contains delay systems in state space form.

The following scheme describes the relations already established and two other results that give sufficient conditions for spectral controllability to imply the existence of a generalized inverse, yielding in this way the equivalence of all the conditions listed in Definition 3.1.



In order to illustrate the assumptions that the class of systems which we will consider must satisfy, we need preliminary definitions and results that concern exponential polynomials in $\mathcal{R}_m[s]$ and elements in $\mathcal{H}_m$.

We remind the reader that $\mathcal{R}_m$ is isomorphic to the ring of Laurent polynomials in m variables and that $\mathcal{R}_m[s]$ is isomorphic to the ring of polynomials in $m + 1$

variables to which we refer for factorization properties. We extend now the definition of monic polynomials to $\mathcal{H}_m$ following [10, Def. 2.2].

DEFINITION 4.1. *An element $r(s) \in \mathcal{R}_m[s]$ is* monic *if the corresponding polynomial in $m+1$ variables is monic with respect to s, i.e., the highest degree power of s is contained in one single monomial of the form $ks^n e^{s\tau}$, $k, \tau \in \mathbb{R}$, $k \neq 0$. An element $h(s) \in \mathcal{H}_m$ is* monic *if it can be written as $h(s) = n(s)/d(s)$ with $n(s) \in \mathcal{R}_m[s]$ monic and $d(s) \in \mathbb{R}[s]$.*

The following proposition concerns the solvability of Bézout equations over $\mathcal{A}$ for delay-differential elements.

PROPOSITION 4.2. *Suppose that the elements $r_1(s), \dots, r_l(s) \in \mathcal{H}_m$ have no common zeros and let $\mathcal{I} \triangleq (r_1(s), \dots, r_l(s))_{\mathcal{A}}$ be the ideal over $\mathcal{A}$ that they generate. Then the following conditions are equivalent.*

1. *$\mathcal{I} = \mathcal{A}$.*
2. *There exists $p(s) \in \mathbb{R}[s]$ such that $p(s) \in \mathcal{I}$.*
3. *There exists $h(s) \in \mathcal{H}_m$ monic such that $h(s) \in \mathcal{I}$.*

In order to prove this proposition we need several technical results.

This first lemma shows that the ideal generated over $\mathcal{H}_m$ by elements in the same ring that have no common zeros contains an exponential polynomial with a particular property.

LEMMA 4.3. *If the elements $r_i(s) \in \mathcal{H}_m$, $i = 1 \dots, l$, do not have common zeros, then there exist $x_i \in \mathcal{R}_m[s]$, $d(s) \in \mathbb{R}[s]$, and $n(s) \in \mathcal{R}_m$ such that*

$$\sum_{i=1}^{l} x_i(s) r_i(s) = d(s)n(s).$$

Proof. By Proposition 2.3-1, $r_i(s) = n_i(s)/d_i(s)$, with $n_i(s) \in \mathcal{R}_m[s]$ and $d_i(s) \in \mathbb{R}[s]$. Let $g(s)$ be the common factor of $n_i(s)$ (viewed as polynomials), and suppose that $g(s_0) = 0$. Then for every i, $n_i(s_0) = r_i(s_0)d_i(s_0) = 0$. Observe that $r_1(s), \dots, r_l(s)$ have no common zeros and thus there is at least a j such that $d_j(s_0) = 0$. This shows that every zero of $g(s)$ is a zero of $d_1(s) \cdots d_l(s) \in \mathbb{R}[s]$, and therefore $g(s)$ has only a finite number of zeros.

Since by Lemma 2.5 the only elements without zeros in $\mathcal{R}_m[s]$ are exponentials, there is a $d(s) \in \mathbb{R}[s]$ and a $\tau \in \mathbb{R}$ such that $g(s) = e^{s\tau} d(s)$, and we can factor $n_i(s) = \bar{n}_i(s)d(s)$ with $\bar{n}_i(s) \in \mathcal{R}_m[s]$ without common (polynomial) factors.

It is known [28, Thm. 2, part 2] that there always exists a linear combination of factor coprime polynomials that is independent of one variable, i.e., there are exponential polynomials $\bar{x}_i(s) \in \mathcal{R}_m[s]$ such that

$$\sum_{i=1}^{l} \bar{x}_i(s)\bar{n}_i(s) = n(s) \in \mathcal{R}_m.$$

So, if we put $x_i(s) = \bar{x}_i(s)d_i(s)$, we get

$$\sum_{i=1}^{l} x_i(s)r_i(s) = \sum_{i=1}^{l} \bar{x}_i(s)d_i(s)r_i(s) = \sum_{i=1}^{l} \bar{x}_i(s)n_i(s) = d(s)\sum_{i=1}^{l} \bar{x}_i(s)\bar{n}_i(s) = d(s)n(s). \qquad \square$$

We will be concerned with properties of holomorphic functions regarding the placement of their zeros. The following notation will be very useful.

Definition 4.4. *For any holomorphic function $a(s) \in \mathcal{O}$, we let $\mathcal{Z}(a)$ be the set of zeros of $a(s)$.*

We remark that $\mathcal{Z}(a)$ has no limit points in $\mathbb{C}$. This is a fundamental property of holomorphic functions. Moreover, we note that the definition of $\mathcal{Z}(a)$ does not take multiplicities into account; they are, in this particular context, an unnecessary complication.

In the next lemma we give an estimate of the position of zeros of two classes of exponential polynomials.

Lemma 4.5.

1. *If $a(s) \in \mathcal{R}_m$, there exists $C > 0$ such that $|\operatorname{Re} s| \leq C$ for every $s \in \mathcal{Z}(a)$.*
2. *If $a(s) \in \mathcal{R}_m[s]$ is monic, there exist two constants $A, B > 0$ such that $|\operatorname{Im} s| \leq Ae^{B|\operatorname{Re} s|}$ for every $s \in \mathcal{Z}(a)$.*

Proof. We prove that if $a(s) \in \mathcal{R}_m$, then there exists a constant $C \geq 0$ such that if $a(s_0) = 0$, then $\operatorname{Re} s_0 \leq C$, the opposite being analogous. We can suppose that $a(s) = 1 + \sum a_i e^{-b_i s}$ with $b_i > 0$; otherwise we could collect a suitable exponential factor ke^{bs}. Now, assume that $a(s_0)$ is zero and $\operatorname{Re} s_0 \geq 0$. Then $1 = |\sum a_i e^{-b_i s_0}|$. If we let $A = \sum |a_i|$ and $B = \min\{b_i\}$, then we obtain $1 = |\sum a_i e^{-b_i s_0}| \leq \sum |a_i|\, e^{-b_i \operatorname{Re} s_0} \leq Ae^{-B \operatorname{Re} s_0}$, which implies that $\operatorname{Re} s_0 \leq \frac{1}{B} \log A$.

Regarding the second inequality, it is not restrictive to suppose that $a(s) = s^n + \sum_{i=0}^{n-1} a_i(s) s^i$ with $a_i(s) \in \mathcal{R}_m$. Consider first only complex numbers $s_0 \in \mathbb{C}$ such that $|s_0| \geq 1$. Then, whenever s_0 is a zero of $a(s)$, we obtain

$$s_0^n = -\sum_{i=0}^{n-1} a_i(s_0) s_0^i \;\Rightarrow\; s_0 = -\sum_{i=0}^{n-1} a_i(s_0) s_0^{i-n+1},$$

and, employing elementary inequalities, we obtain

$$|\operatorname{Im} s_0| \leq |s_0| \leq \sum_{i=0}^{n-1} |a_i(s_0)|\, |s_0^{i-n+1}| \leq \sum_{i=0}^{n-1} |a_i(s_0)|.$$

Note that every polynomial $a_i(s) \in \mathcal{R}_m$ can be written as $a_i(s) = \sum_{j=1}^{\nu_i} a_{ij} e^{b_{ij} s}$, and so

$$|a_i(s_0)| \leq \sum_{j=1}^{\nu_i} |a_{ij}|\, |e^{b_{ij} s_0}| \leq \nu_i \max_j\{|a_{ij}|\} e^{\max_j\{|b_{ij}|\} |\operatorname{Re} s_0|}.$$

Therefore, for some $\bar{A} > 0, B > 0$, we have $|\operatorname{Im} s_0| \leq \bar{A} e^{B|\operatorname{Re} s_0|}$ for $|s_0| \geq 1$. Taking $A = \max\{1, \bar{A}\}$, we get the claimed result. □

The following proposition extends to $\mathcal{H}_m$ an estimate from above that is known for exponential polynomials in $\mathcal{R}_m[s]$.

Proposition 4.6. *For every $h(s) \in \mathcal{H}_m$ there exist real constants $K, M, N, E > 0$ such that*

$$|h(s)| \geq \frac{K \operatorname{dist}(s, \mathcal{Z}(h))^M e^{-E|\operatorname{Re} s|}}{(1 + |s|)^N}. \tag{30}$$

Proof. Write $h(s)$ as $n(s)/d(s)$, where, by Proposition 2.3-1, $n(s) \in \mathcal{R}_m[s]$ and $d(s) \in \mathbb{R}[s]$. Notice that if $d(s)$ has degree ν, there is a $D > 0$ such that $|d(s)| \leq D(1+|s|)^\nu$. Indeed, if $d(s) = \sum_0^\nu d_i s^i$, let $D \triangleq \max\{|d_i|/\binom{\nu}{i}\}$ and in this way we have

$|d(s)| \leq \sum |d_i| \, |s|^i \leq D \sum \binom{\nu}{i} |s|^i = D(1+|s|)^\nu$. As stated in [3, Prop. 1], since $n(s)$ is an exponential polynomial, there are real constants $\tilde{K}, \tilde{N}, M, E > 0$ such that

$$|n(s)| \geq \frac{\tilde{K} \operatorname{dist}(s, \mathcal{Z}(n))^M e^{-E|\operatorname{Re} s|}}{(1+|s|)^{\tilde{N}}}.$$

Therefore, we obtain

$$|h(s)| = \frac{|n(s)|}{|d(s)|} \geq \frac{\tilde{K} \operatorname{dist}(s, \mathcal{Z}(n))^M e^{-E|\operatorname{Re} s|}}{D(1+|s|)^{\tilde{N}+\nu}}. \tag{31}$$

Now let $\mathcal{Z}_e \triangleq \mathcal{Z}(n) \setminus \mathcal{Z}(h)$ be that subset of zeros of $d(s)$ that eliminate, by division, some of the zeros of $n(s)$. We employ now Lemma B.1 by letting $\mathcal{V} = \mathcal{Z}(h)$ and $\mathcal{W} = \mathcal{Z}_e$, which is finite and thus compact. We obtain that fixing any positive $L < 1$, there is an $R > 0$ such that

$$\operatorname{dist}(s, \mathcal{Z}(n)) \geq L \operatorname{dist}(s, \mathcal{Z}(h)) \text{ as } |s| > R.$$

Therefore, if we let $N \triangleq \tilde{N} + \nu$, from (31) we get

$$|h(s)| \geq \frac{\tilde{K} L^M}{D} f(s) \text{ for every } s \text{ such that } |s| > R, \tag{32}$$

where $f(s) \triangleq \operatorname{dist}(s, \mathcal{Z}(h))^M e^{-E|\operatorname{Re} s|}(1+|s|)^{-N}$.

Note that, since $L < 1$, this relation still holds inside the closed disk, i.e., for $|s| \leq R$, wherever we have $\operatorname{dist}(s, \mathcal{Z}(n)) = \operatorname{dist}(s, \mathcal{Z}(h))$, that corresponds to the condition $\operatorname{dist}(s, \mathcal{Z}(h)) \leq \operatorname{dist}(s, \mathcal{Z}_e)$.

Now we have only to consider the compact subset $\mathcal{U}$ of the closed disk constituted by $s \in \mathbb{C}$ such that $|s| \leq R$ and $\operatorname{dist}(s, \mathcal{Z}(h)) \geq \operatorname{dist}(s, \mathcal{Z}_e)$. The function $h(s)$ has no zeros in $\mathcal{U}$, and therefore there exists a constant $H > 0$ such that $|h(s)| \geq H$ for all $s \in \mathcal{U}$. Moreover, $f(s)$ has maximum F on the compact $\mathcal{U}$. So, if we let $K \triangleq \min\{\frac{H}{F}, \frac{\tilde{K}L^M}{D}\}$, then

$$|h(s)| \geq H \geq KF \geq Kf(s) \ \ \forall s \in \mathcal{U},$$

and, by (32), on the remaining points of $\mathbb{C}$ we have

$$|h(s)| \geq \frac{\tilde{K} L^M}{D} f(s) \geq K f(s) \text{ for every } s \in \mathbb{C} \setminus \mathcal{U},$$

thus completing the proof of the Proposition. □

We are now in a position to prove Proposition 4.2.

Proof. The first condition implies obviously the last two. We have to prove the converse.

($2 \Rightarrow 1$) Let $R(s) = [r_i(s)] \in \mathcal{H}_m^{1 \times l}$ be the row vector containing the given polynomials. By hypothesis we know that there is a column vector $A(s) \in \mathcal{A}^{l \times 1}$ such that we have

$$p(s) \triangleq R(s)A(s) = \sum_i r_i(s) a_i(s) \in \mathbb{R}[s].$$

If s_0 is a zero of $p(s)$, then $R(s_0)A(s_0) = 0$. Since $r_i(s)$ have no common zeros, $R(s_0) \neq 0$ and hence $A(s_0) \in \ker_{\mathbb{C}} R(s_0)^\circ$. By Theorem 2.12 there is a matrix $M(s) \in$

$\mathcal{H}_m^{t\times d}$ with constant $\operatorname{rank}_{\mathbb{C}} M(\lambda)$ for every $\lambda \in \mathbb{C}$, such that $\ker_{\mathcal{O}} R(s)\circ = \operatorname{im}_{\mathcal{O}} M(s)\circ$. Therefore, $\ker_{\mathbb{C}} R(s_0)\circ = \operatorname{im}_{\mathbb{C}} M(s_0)\circ$, and so there exists a column $c \in \mathbb{C}^{d\times 1}$ such that $A(s_0) = M(s_0)c$. Since $R(s)M(s)c = 0$, we can write

$$p(s) = R(s)(A(s)-M(s)c) \;\Rightarrow\; \tilde{p}(s) \triangleq \frac{p(s)}{s-s_0} = R(s)\frac{A(s)-M(s)c}{s-s_0} = R(s)\tilde{A}(s),$$

where $\tilde{p}(s)$ is a polynomial with lower degree than $p(s)$ and $\tilde{A}(s)$ is a vector with entries in $\mathcal{A}$ by Proposition 2.3-3. Iterating this procedure, we get a Bézout equation for $r_i(s)$ with coefficients in $\mathcal{A}$.

($3 \Rightarrow 2$) The element $h(s) \in \mathcal{I}$ may be written as a quotient in $\mathcal{R}_m[s]$ whose numerator $a(s) \in \mathcal{R}_m[s]$ is monic and belongs to $\mathcal{I}$. By Lemma 4.5-2, we have that $|\operatorname{Im} s| \le Ae^{B|\operatorname{Re} s|}$ for every $s \in \mathcal{Z}(a)$. Moreover, by Lemma 4.3, $r_i(s)$ generate an exponential polynomial $b(s) = d(s)n(s)$ such that $d(s) \in \mathbb{R}[s]$ and $n(s) \in \mathcal{R}_m$. Since $d(s)$ has a finite number of zeros and the real part of the zeros of $n(s)$ is bounded by Lemma 4.5-1, there is a constant $C > 0$ such that $|\operatorname{Re} s| \le C$ for every $s \in \mathcal{Z}(b)$.

So, if we consider the closed disk $\mathcal{C}$ centered at the origin with radius R such that $R^2 > A^2e^{2BC} + C^2$, it contains (strictly) the region containing the common zeros of $a(s)$ and $b(s)$. We can build two polynomials $p(s)$ and $q(s)$ having those zeros of $a(s)$ and $b(s)$, respectively, that lie in $\mathcal{C}$. In this way we can define the elements in $\mathcal{H}_m$

$$h_a(s) \triangleq \frac{a(s)}{p(s)} \in \mathcal{H}_m \text{ and } h_b(s) \triangleq \frac{b(s)}{q(s)} \in \mathcal{H}_m$$

that have no zeros inside $\mathcal{C}$ and thus, in particular, have no common zeros. Moreover, by construction there exists a constant $D > 0$ such that $\operatorname{dist}(\mathcal{Z}(h_a), \mathcal{Z}(h_b)) \ge 2D$. Therefore, we can define the sets

$$\mathcal{C}_a \triangleq \{s \in \mathbb{C} \text{ such that } \operatorname{dist}(s, \mathcal{Z}(h_a)) \le \operatorname{dist}(s, \mathcal{Z}(h_b))\},$$
$$\mathcal{C}_b \triangleq \{s \in \mathbb{C} \text{ such that } \operatorname{dist}(s, \mathcal{Z}(h_a)) \ge \operatorname{dist}(s, \mathcal{Z}(h_b))\},$$

and so $\operatorname{dist}(s, \mathcal{Z}(h_a)) \ge D$ for every $s \in \mathcal{C}_b$ and $\operatorname{dist}(s, \mathcal{Z}(h_b)) \ge D$ for every $s \in \mathcal{C}_a$. As stated in Proposition 4.6, there are suitable positive constants such that

$$|h_a(s)| \ge \frac{K_a \operatorname{dist}(s, \mathcal{Z}(h_a))^{M_a} e^{-E_a|\operatorname{Re} s|}}{(1+|s|)^{N_a}} \ge \frac{K_a D^{M_a} e^{-E_a|\operatorname{Re} s|}}{(1+|s|)^{N_a}} \;\forall s \in \mathcal{C}_b,$$
$$|h_b(s)| \ge \frac{K_b \operatorname{dist}(s, \mathcal{Z}(h_b))^{M_b} e^{-E_b|\operatorname{Re} s|}}{(1+|s|)^{N_b}} \ge \frac{K_b D^{M_b} e^{-E_b|\operatorname{Re} s|}}{(1+|s|)^{N_b}} \;\forall s \in \mathcal{C}_a.$$

So, since $\mathbb{C} = \mathcal{C}_a \cup \mathcal{C}_b$, if we let $K \triangleq \min\{K_a D^{M_a}, K_b D^{M_b}\}$, $E \triangleq \max\{E_a, E_b\}$, and $N \triangleq \max\{N_a, N_b\}$, we obtain

$$|h_a(s)| + |h_b(s)| \ge \frac{Ke^{-E|\operatorname{Re} s|}}{(1+|s|)^N} \;\; \forall s \in \mathbb{C}.$$

We know [14] that this condition ensures the existence of two Paley–Wiener functions $x_a(s), x_b(s) \in \mathcal{A}$ such that $h_a(s)x_a(s) + h_b(s)x_b(s) = 1$, i.e.,

$$a(s)q(s)x_a(s) + b(s)p(s)x_b(s) = p(s)q(s),$$

and therefore the ideal $\mathcal{I}$ contains the polynomial $p(s)q(s)$. $\square$

An immediate consequence of the previous proposition is the following result for delay-differential systems.

THEOREM 4.7. *Let $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ with $R(s) \in \mathcal{H}_m^{p\times q}$. If the ideal generated by the rank minors of $R(s)$ over $\mathcal{A}$ contains a monic element $a(s) \in \mathcal{H}_m$, then the conditions in Definition* 3.1 *are all equivalent.*

Proof. We have only to prove that, under this hypothesis, **SC** $\Rightarrow$ **GI**, or, equivalently, by Theorem 2.9, that the rank minors satisfy a Bézout equation in $\mathcal{A}$. This is true by Proposition 4.2. □

Remark 4.8. Note that for systems with only one delay, spectral controllability always implies that the ideal generated by the rank minors contains a polynomial only in s, and therefore the hypothesis of Theorem 4.7 always holds. We have in this way that all the conditions of Definition 3.1 are always equivalent for systems with one delay, as already proved in [9, 23].

Remark 4.9. The previous theorem does not hold in general for a behavior $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ with $R(s) \in \mathcal{A}^{p\times q}$, unless its rank minors belong to $\mathcal{H}_m$. However, note that the proof ($2 \Rightarrow 1$) of Proposition 4.2 still holds if $r_i(s) \in \mathcal{A}$, and therefore we have the following result.

Let $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ with $R(s) \in \mathcal{A}^{p\times q}$. If the ideal generated by the rank minors of $R(s)$ over $\mathcal{A}$ contains a polynomial $p(s) \in \mathbb{R}[s]$, then the conditions in Definition 3.1 are all equivalent.

Theorem 4.7 constitutes a powerful tool even though the check of the fact that there exists a monic element in the ideal over $\mathcal{A}$ generated by the rank minors is difficult to perform in practice. On the other hand, if $R(s) \in \mathcal{H}_m^{p\times q}$, by Gröbner bases techniques it can be checked whether there exists a monic polynomial in the ideal over $\mathcal{H}_m$ generated by the rank minors, and this is actually still a sufficient condition for the equivalence of the conditions of Definition 3.1. Indeed, by Proposition 2.3-1 we can write the rank minors of $R(s)$ as $r_i(s) = p_i(s)/\rho(s)$, where $p_i(s) \in \mathcal{R}_m[s]$ and $\rho(s) \in \mathbb{R}[s]$. It is not difficult to prove that there exists a monic $h(s) \in \mathcal{H}_m$ in the ideal over $\mathcal{H}_m$ generated by the rank minors if and only if there exists a monic $p(s)$ in the ideal over $\mathcal{R}_m[s]$ generated by the Laurent exponential polynomials $p_i(s)$.

The nicest situation in which the previous theorem can be applied is in relation with the class of systems in state space form

$$\dot{x} = A(\sigma_{\tau_1}, \dots, \sigma_{\tau_m})x + B(\sigma_{\tau_1}, \dots, \sigma_{\tau_m})u. \tag{33}$$

Actually, the first rank minor of the matrix

$$R(s) = [sI - A(e^{-s\tau_1}, \dots, e^{s\tau_m})\ B(e^{-s\tau_1}, \dots, e^{s\tau_m})],$$

which provides a kernel representation of the system (33), is monic.

5. Conclusions.

In this paper we have shown some remarkable properties related to controllability of systems described by delay-differential or convolutional equations. We have shown, however, that the equivalence between spectral controllability, behavior controllability, and the existence of an image representation does not hold in this general setup, essentially due to the fact that images may not be closed in general.

There are two ways to overcome this difficulty. One is to weaken the notion of image representation. The second is to find subclasses of systems for which the equivalence holds. We showed that for the class of systems in state space form such equivalence holds.

Several problems remain unsolved. We conclude this paper by giving a brief list of open problems, which will be the object of our future investigation.

1. It would be interesting to understand if the equivalence of conditions **SC**, **DCS**, and **DIR** can be extended to nonregular behaviors, i.e., systems admitting a kernel representation $\mathcal{B} = \ker_{\mathcal{E}} R(s)$ in which the matrix $R(s)$ is not full row rank.

2. Another open problem is connected with the possibility of extending Theorem 3.10, showing that **IR** and **GI** are equivalent for regular systems with a single input to a more general situation. As pointed out by H. Glüsing-Lüerßen in a personal communication, it is not possible to extend such a result to nonregular systems, as proved by the following example:

$$\mathcal{B} = \ker_{\mathcal{E}} \begin{bmatrix} r_1(s) & 0 \\ r_2(s) & 0 \end{bmatrix} = \mathrm{im}_{\mathcal{E}} \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

where $r_1(s), r_2(s)$ are defined in (5) within the proof of Proposition 2.6. The matrix providing the kernel representation of $\mathcal{B}$ has no generalized inverse over $\mathcal{A}$.

3. An interesting direction of future research concerns a very detailed investigation of the class of systems for which all the conditions of Definition 3.1 are equivalent. We conjecture that the lack of such equivalence is the typical property of systems whose spectral controllability is not robust with respect to variations of the delays.

Appendix A. Proof of Theorem 2.12. The proof needs a rather involved notation that two simple examples will help to explain.

Let $R(s) = [a\ b\ c\ d] \in \mathcal{A}^{1\times 4}$. (We omit the dependence of the elements on s.) By Remark 2.10, the hypotheses of Theorem 2.12 are satisfied with $r = p = 1$, $q = 4$ as soon as $R(\lambda) \neq 0$ for all $\lambda \in \mathbb{C}$.

We show how we can construct a matrix $M(s) \in \mathcal{A}^{4\times 6}$ of rank $q - r = q - p = 3$ such that $R(s)M(s) = 0$. Consider the matrix

$$\tilde{R}(s) = \begin{bmatrix} w & x & y & z \\ a & b & c & d \end{bmatrix},$$

where the vector $[w\ x\ y\ z]$ belongs to the image of $\circ R(s)$; i.e., it is a scalar multiple of $R(s)$. In this case $\tilde{R}(s)$ has still rank $p = 1$ and therefore its minors of order $p+1$ are zero. Note that the minors of $\tilde{R}(s)$ are linear functions of w, x, y, or z. We will show that we can write them as particular row-column products.

We have $6 = \binom{4}{2} = \binom{q}{p+1}$ different (ordered) sets of $2 = p + 1$ columns of $R(s)$,

$$\rho_1 = \{1,4\}, \ \rho_2 = \{2,4\}, \ \rho_3 = \{3,4\}, \ \rho_4 = \{1,3\}, \ \rho_5 = \{2,3\}, \ \rho_6 = \{1,2\},$$

that correspond to every minor of order $p+1$ of $\tilde{R}(s)$. The minor given by ρ_1 is $wd - za = 0$, which can be written (up to multiplication by -1) as

$$\begin{bmatrix} w & x & y & z \end{bmatrix} \begin{bmatrix} -d \\ 0 \\ 0 \\ a \end{bmatrix},$$

i.e., the first element in the column is $-d$, the minor of $R(s)$ corresponding to the set of columns $\{4\} = \rho_1 \setminus \{1\}$. (We consider it with the opposite sign if the row we are considering, 1, occupies an odd position in ρ.) The second and third elements are zero ($2, 3 \notin \rho_1$), the fourth element is a, and the minor of column $\{1\} = \rho_1 \setminus \{4\}$ of $R(s)$.

From every ρ_i we can construct such a row and obtain the matrix

$$M(s) = \begin{bmatrix} -d & 0 & 0 & -c & 0 & -b \\ 0 & -d & 0 & 0 & -c & a \\ 0 & 0 & -d & a & b & 0 \\ a & b & c & 0 & 0 & 0 \end{bmatrix},$$

which satisfies the equation $R(s)M(s) = 0$.

Note that $-d^3$ is a minor of order $3 = q - p$ of $M(s)$. The third power of every other minor of $R(s)$ is a minor of order 3 of $M(s)$. Some of the rank minors of $R(s)$ are not zero, so $M(s)$ has at least rank 3. It cannot have rank 4 because otherwise the equation $R(s)M(s) = 0$ would imply that $R(s) = 0$.

Without being so detailed, we show what happens with a full rank matrix having dimensions $p = 2$ and $q = 4$:

$$R(s) = \begin{bmatrix} a & b & c & d \\ \alpha & \beta & \gamma & \delta \end{bmatrix}.$$

Matrix $\tilde{R}(s)$ is now

$$\tilde{R}(s) = \begin{bmatrix} w & x & y & z \\ a & b & c & d \\ \alpha & \beta & \gamma & \delta \end{bmatrix},$$

where $[w\ x\ y\ z]$ is any linear combination of the rows of $R(s)$. Its minors of order $p + 1 = 3$ correspond to columns in the $d = \binom{q}{p+1} = \binom{4}{3} = 4$ sets

$$\rho_1 = \{1, 3, 4\},\ \rho_2 = \{2, 3, 4\},\ \rho_3 = \{1, 2, 4\},\ \rho_4 = \{1, 2, 3\}$$

and permit us to construct $M(s)$ in the following way:

$$M(s) = \begin{bmatrix} d\gamma - c\delta & 0 & d\beta - b\delta & c\beta - b\gamma \\ 0 & d\gamma - c\delta & a\delta - d\alpha & a\gamma - c\alpha \\ a\delta - d\alpha & b\delta - d\beta & 0 & b\alpha - a\beta \\ c\alpha - a\gamma & c\beta - b\gamma & b\alpha - a\beta & 0 \end{bmatrix}.$$

$M(s)$ has obviously rank $2 = q - p$, since among its minors of order 2 there are the squares of the minors of order $p = 2$ of $R(s)$.

Proof of Theorem 2.12. Since this proof is quite long, we divide it into steps.

First step. We prove the existence of $M(s)$ such that $R(s)M(s) = 0$. If $r = q$, then $R(s)$ is left invertible over $\mathcal{O}$. There exists $G(s) \in \mathcal{O}^{q \times p}$ such that $G(s)R(s) = I$. This implies that $R(s)\circ$ is injective and $\circ R(s)$ surjective over $\mathcal{O}$. Therefore, $M(s) = 0 \in \mathcal{S}^{q \times 1}$ satisfies the given conditions.

Let us suppose that $r < q$. If $r(s)$ is any element in $\mathcal{O}^p R(s)$ and $\bar{R}(s) \in \mathcal{O}^{r \times q}$ is one of the $\binom{p}{r}$ submatrices of $R(s)$ with r rows, we can build the matrix

$$\tilde{R}(s) = \begin{bmatrix} r(s) \\ \bar{R}(s) \end{bmatrix}. \tag{34}$$

$\tilde{R}(s)$ has rank at most r, so every minor of order $r+1$ is zero and is a linear combination of $r+1$ elements of $r(s)$, the coefficients being $r+1$ minors of order r of $\bar{R}(s)$ and thus rank minors of $R(s)$.

More precisely, let $\rho \subseteq \mathbb{N}$ be a subset of $r+1$ elements of the set $\{1, 2, \ldots, q\}$. We suppose that ρ is ordered and write $\rho(i)$ to indicate the ith element of ρ. Furthermore, we denote by

$$\bar{\rho}(i) \triangleq \rho \setminus \{i\} \tag{35}$$

the ordered set with r elements which has the elements of ρ except of i, and we let

$$n_\rho(i) \triangleq \begin{cases} 0 & \text{if } i \notin \rho, \\ (-1)^k & \text{if } \rho(k) = i. \end{cases} \tag{36}$$

That is to say if $i \in \rho$, then $n_\rho(i)$ is equal to 1 when i occupies an even "position" in ρ.

We know by basic combinatorics that there are exactly $\bar{d} = \binom{q}{r+1}$ different sets ρ_j, so we can construct a matrix $\bar{M}(s) \in \mathcal{S}^{q \times \bar{d}}$ with elements $\bar{m}_{ij}(s)$ defined as

$$\bar{m}_{ij}(s) \triangleq n_{\rho_j}(i) \bar{R}(s)_{\bar{\rho}_j(i)}, \tag{37}$$

where $\bar{R}(s)_\rho$ is the determinant of the matrix formed by those columns of $\bar{R}(s)$ which are indexed by elements in the set ρ.

We see that if $r(s)$ in (34) has elements $r_i(s)$ and $\bar{M}_j(s)$ is the jth column of $\bar{M}(s)$, then, remembering that by definition (36) $n_{\rho_j}(i) = 0$ when $i \notin \rho_j$, we have

$$0 = \tilde{R}(s)_{\rho_j} = \sum_{i \in \rho_j} r_i(s) n_{\rho_j}(i) \bar{R}(s)_{\bar{\rho}_j(i)} = \sum_{i=1}^{q} r_i(s) n_{\rho_j}(i) \bar{R}(s)_{\bar{\rho}_j(i)} = r(s) \bar{M}_j(s).$$

Since $r(s)$ may be any row in $\mathcal{O}^p R(s)$ and thus any row of $R(s)$, this proves that $R(s)\bar{M}(s) = 0$. Then, if we let

$$d \triangleq \bar{d} \binom{p}{r} = \binom{q}{r+1} \binom{p}{r} \tag{38}$$

and $M(s) \in \mathcal{S}^{q \times d}$ be the block column matrix containing the $\binom{p}{r}$ matrices $\bar{M}(s)$ obtained by varying the set of rows defining $\bar{R}(s)$, we still have

$$R(s) M(s) = 0. \tag{39}$$

Second step. $M(s)$ has rank $q - r$ and admits a generalized inverse over $\mathcal{O}$. Let $\bar{R}(s) \in \mathcal{O}^{r \times q}$ be a submatrix of $R(s)$ with r rows, as in the previous step, and consider $\bar{M}(s)$ defined by (37). Let $\rho \triangleq \{q-r+1, q-r+2, \ldots, q-1, q\}$ be the set that indicates the last r columns of $\bar{R}(s)$, and assume that the sets ρ_j, $j = 1, \ldots, \bar{d}$ are such that

$$\rho_j = \{j\} \cup \rho \quad \forall j \in \{1, 2, \ldots, q-r\}.$$

By definition (35) $\bar{\rho}_j(j) = \rho$, and thus by (37)

$$|\bar{m}_{jj}(s)| = |\bar{R}(s)_\rho| \quad \forall j \in \{1, 2, \ldots, q-r\}.$$

Again the definition (37) of $\bar{m}_{ij}(s)$ and (36) imply that

$$\forall i,j \in \{1,2,\dots,q-r\}\,,\; i \neq j \;\Rightarrow\; n_{\rho_j}(i) = 0, \text{and therefore } \bar{m}_{ij}(s) = 0,$$

so the submatrix containing the first $q-r$ rows and columns of $\bar{M}(s)$ has nonzero entries only on the main diagonal and these are equal, up to the sign, to the minor $\bar{R}(s)_\rho$ of $R(s)$.

It is obvious, by the symmetric structure of the problem, that the set of minors of order $q-r$ of $M(s)$ contains every $(q-r)$th power of the rank minors of $R(s)$. This shows that $M(s)$ has rank $m \geq q-r$.

If we denote by $\mathcal{M}$ the field of fractions of $\mathcal{O}$, then the dimension of $\ker_{\mathcal{M}} R(s)\circ$ is $q-r$ and the dimension of $\operatorname{im}_{\mathcal{M}} M(s)\circ$ is m. Therefore, since $R(s)M(s) = 0$, the first vector space includes the second; hence $m \leq q-r$ and therefore $m = q-r$.

Since the maximal minors of $R(s)$ have no common zeros and their mth power is contained in the set of rank minors of $M(s)$, even these minors cannot have common zeros. Therefore, by Theorem 2.9, $M(s)$ also has a generalized inverse.

Third step. We prove that $\operatorname{im}_{\mathcal{O}} \circ R(s) = \ker_{\mathcal{O}} \circ M(s)$ and $\ker_{\mathcal{O}} R(s)\circ = \operatorname{im}_{\mathcal{O}} M(s)\circ$. Equation (39) implies that $\operatorname{im}_{\mathcal{O}} \circ R(s) \subseteq \ker_{\mathcal{O}} \circ M(s)$. In order to prove the converse inclusion, let us consider again kernels and images over the field of fractions of $\mathcal{O}$. In this case $\operatorname{im}_{\mathcal{M}} \circ R(s)$ is a subspace of $\ker_{\mathcal{M}} \circ M(s)$. However, $\operatorname{im}_{\mathcal{M}} \circ R(s)$ has dimension r and $\ker_{\mathcal{M}} \circ M(s)$ has dimension $q-m = r$, and therefore they coincide. So, for every $x(s) \in \ker_{\mathcal{O}} \circ M(s)$ there is an $\bar{y}(s) \in \mathcal{M}^p$ such that $x(s) = \bar{y}(s)R(s)$. Multiplying by a suitable $d(s) \in \mathcal{O}$, we obtain $d(s)x(s) = y(s)R(s)$ with $y(s) \in \mathcal{O}^p$. By Lemma 2.11, $d(s)x(s) \in \operatorname{im}_{\mathcal{O}} \circ R(s) = \ker_{\mathcal{O}} \circ (I - G(s)R(s))$. Since $\mathcal{O}$ is a domain, also $x(s) \in \ker_{\mathcal{O}} \circ (I - G(s)R(s)) = \operatorname{im}_{\mathcal{O}} \circ R(s)$, and this proves that $\ker_{\mathcal{O}} \circ M(s) \subseteq \operatorname{im}_{\mathcal{O}} \circ R(s)$.

The equation $\ker_{\mathcal{O}} R(s)\circ = \operatorname{im}_{\mathcal{O}} M(s)\circ$ follows analogously, since $\ker_{\mathcal{M}} R(s)\circ$ and $\operatorname{im}_{\mathcal{M}} M(s)\circ$ are the same vector space with dimension $q-r$. □

Appendix B. A technical lemma.

Lemma B.1. *Let* $\emptyset \neq \mathcal{V}, \mathcal{W} \subset \mathbb{C}$ *with* $\mathcal{W}$ *compact, and let* $0 < L < 1$. *Then there exists an* $R > 0$ *such that for every* $s \in \mathbb{C}$ *such that* $|s| > R$*,*

$$\operatorname{dist}(s, \mathcal{V} \cup \mathcal{W}) \geq L \operatorname{dist}(s, \mathcal{V}). \tag{40}$$

Proof. Since the function $\operatorname{dist}(w, \mathcal{V})$ of w is continuous [5, Thm. IX.4.3], it has a maximum over the compact $\mathcal{W}$ that we denote by $D \geq 0$. We want to show that the following inequality holds true:

$$\operatorname{dist}(s, \mathcal{V}) \leq \operatorname{dist}(s, \mathcal{V} \cup \mathcal{W}) + D \quad \forall s \in \mathbb{C}. \tag{41}$$

Define the set

$$\mathcal{U} \triangleq \{s \in \mathbb{C} \text{ such that } \operatorname{dist}(s, \mathcal{V}) > \operatorname{dist}(s, \mathcal{W})\}\,. \tag{42}$$

Observe that $\operatorname{dist}(s, \mathcal{V} \cup \mathcal{W}) = \min\{\operatorname{dist}(s, \mathcal{V}), \operatorname{dist}(s, \mathcal{W})\}$. Thus, inequality (41) is trivial if $s \notin \mathcal{U}$, because in this case $\operatorname{dist}(s, \mathcal{V} \cup \mathcal{W}) = \operatorname{dist}(s, \mathcal{V})$. So, assume that $\operatorname{dist}(s, \mathcal{V} \cup \mathcal{W}) = \operatorname{dist}(s, \mathcal{W})$. By definition of distance, the relation

$$\operatorname{dist}(s, \mathcal{V}) = \inf_{v \in \mathcal{V}} |s - v| \leq |s - w| + \inf_{v \in \mathcal{V}} |w - v| = |s - w| + \operatorname{dist}(w, \mathcal{V}) \leq |s - w| + D$$

holds for every $w \in \mathcal{W}$. Therefore, it is true for $|s - w| = \operatorname{dist}(s, \mathcal{W})$, so proving (41).

Fix now $0 < L < 1$. Note that, if $s \notin \mathcal{U}$, condition (40) is always satisfied. Therefore, if $\mathcal{U}$ is bounded, the claim is proved choosing an $R > 0$ such that $|s| > R$ implies that $s \notin \mathcal{U}$.

Suppose, conversely, that $\mathcal{U}$ is not bounded and that s belongs to $\mathcal{U}$. As $|s| \to \infty$, even $\operatorname{dist}(s, \mathcal{W}) \to \infty$, being that $\mathcal{W}$ is compact, and so does $\operatorname{dist}(s, \mathcal{V})$, by definition (42). Therefore, there is an $R > 0$ such that $\operatorname{dist}(s, \mathcal{V}) \geq D/(1-L)$ for every s such that $|s| > R$. This is equivalent to

$$1 - \frac{D}{\operatorname{dist}(s, \mathcal{V})} \geq L.$$

Thus, using this inequality and formula (41), we obtain that

$$\operatorname{dist}(s, \mathcal{V} \cup \mathcal{W}) \geq \operatorname{dist}(s, \mathcal{V}) - D = \operatorname{dist}(s, \mathcal{V}) \left(1 - \frac{D}{\operatorname{dist}(s, \mathcal{V})}\right) \geq L \operatorname{dist}(s, \mathcal{V}),$$

which proves condition (40) for every s such that $|s| > R$. □

REFERENCES

[1] C. A. Berenstein and M. A. Dostal, *The Ritt theorem in several variables*, Ark. Mat., 12 (1974), pp. 267–280.

[2] C. A. Berenstein and D. C. Struppa, *Complex analysis and convolution equations*, in Several Complex Variables, Encyclopaedia of Mathematical Sciences, Springer–Verlag, Berlin, 1993, pp. 1–108.

[3] C. A. Berenstein and A. Yger, *On Lojasiewicz-type inequalities for exponential polynomials*, J. Math. Anal. Appl., 129 (1988), pp. 166–195.

[4] K. P. S. Bhaskara Rao, *On generalized inverses of matrices over integral domains*, Linear Algebra Appl., 49 (1983), pp. 179–189.

[5] J. Dugundji, *Topology*, Allyn and Bacon, Inc., Boston, 1966.

[6] L. Ehrenpreis, *Solution of some problem of division. Part* II. *Division by a punctual distribution*, Amer. J. Math., 77 (1955), pp. 286–292.

[7] M. Fliess and H. Mounier, *Interpretation and comparison of various types of delay system controllabilities*, in Proceedings of the International Federation of Automatic Control Conference on System Structure and Control, Nantes, France, 1995, pp. 330–335.

[8] M. Fliess and H. Mounier, *Tracking control and π-freeness of infinite dimensional linear systems*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 301–314; also available online from http://www.emath/Maths/Cocv.

[9] H. Glüsing-Lüerssen, *A behavioral approach to delay-differential systems*, SIAM J. Control Optim., 35 (1997), pp. 480–499.

[10] H. Glüsing-Lüerssen, *First-order representations of delay-differential systems in a behavioral setting*, European J. Control, 3 (1997), pp. 137–149.

[11] L. C. G. J. M. Habets, *System equivalence for AR-systems over rings—with an application to delay-differential systems*, Math. Control Signals Systems, 12 (1999), pp. 219–244.

[12] O. Helmer, *Divisibility properties of integral functions*, Duke Math. J., 6 (1940), pp. 345–356.

[13] O. Helmer, *The elementary divisor theorem for certain rings without chain condition*, Bull. Amer. Math. Soc. (N.S.), 49 (1943), pp. 225–236.

[14] L. Hörmander, *Generators for some rings of analytic functions*, Bull. Amer. Math. Soc. (N.S.), 73 (1967), pp. 943–949.

[15] E. W. Kamen, *On an algebraic theory of systems defined by convolution operators*, Math. Systems Theory, 9 (1975), pp. 57–74.

[16] J. L. Kelley and I. Namioka, *Linear Topological Spaces*, Van Nostrand, New York, 1963.

[17] B. J. Levin, *Distribution of Zeros of Entire Functions*, AMS, Providence, Rhode Island, 1980.

[18] B. Malgrange, *Existence et approximation des solutions des équations aux dérivées partielles et des équations de convolution*, Ann. Inst. Fourier (Grenoble), 6 (1955–1956), pp. 271–355.

[19] G. H. Meisters, *Periodic distributions and non-Liouville numbers*, J. Funct. Anal., 26 (1977), pp. 68–88.

[20] H. Mounier, *Algebraic interpretations of the spectral controllability of a linear delay system*, Forum Math., 10 (1998), pp. 39–58.

[21] A. W. Olbrot and L. Pandolfi, *Null controllability of a class of functional-differential systems*, Internat. J. Control, 47 (1988), pp. 193–208.

[22] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Springer-Verlag, Berlin, 1997.

[23] P. Rocha and J. C. Willems, *Behavioral controllability of delay-differential systems*, SIAM J. Control Optim., 35 (1997), pp. 254–264.

[24] F. Treves, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, New York, 1967.

[25] P. Vettori, *Delay-Differential Systems in the Behavioral Approach*, Ph.D. thesis, Department of Electrical Engineering and Computing Science, University of Padova, Italy, 1999.

[26] J. C. Willems, *Models for dynamics*, in Dynamics Reported, John Wiley & Sons, Chichester, UK, 1989, pp. 171–269.

[27] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 259–294.

[28] D. C. Youla and G. Gnavi, *Notes on n-dimensional system theory*, IEEE Trans. Circuits and Systems, 26 (1979), pp. 105–111.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 757–775

ON THE DUALITY BETWEEN FILTERING AND NEVANLINNA–PICK INTERPOLATION[*]

CHRISTOPHER I. BYRNES[†] AND ANDERS LINDQUIST[‡]

Abstract. Positive real rational functions play a central role in both deterministic and stochastic linear systems theory, as well as in circuit synthesis, spectral analysis, and speech processing. For this reason, results about positive real transfer functions and their realizations typically have many applications and manifestations.

In this paper, we study certain manifolds and submanifolds of positive real transfer functions, describing a fundamental geometric duality between filtering and Nevanlinna–Pick interpolation. Not surprisingly, then, this duality, while interesting in its own right, has several corollaries which provide solutions and insight into some very interesting and intensely researched problems. One of these is the problem of parameterizing all rational solutions *of bounded degree* of the Nevanlinna–Pick interpolation problem, which plays a central role in robust control, and for which the duality theorem yields a complete solution. In this paper, we shall describe the duality theorem, which we motivate in terms of both the interpolation problem and a fast algorithm for Kalman filtering, viewed as a nonlinear dynamical system on the space of positive real transfer functions.

We also outline a new proof of the recent solution to the rational Nevanlinna–Pick interpolation problem, using an algebraic topological generalization of Hadamard's global inverse function theorem.

Key words. Nevanlinna–Pick interpolation, filtering, positive real functions, foliations, degree constraint

AMS subject classifications. 30E05, 47N70, 58C99, 93B29, 93E11

PII. S0363012999351115

1. Introduction.

Modulo a conformal equivalence, the classical Nevanlinna–Pick problem amounts to determining a function which is *positive real*, i.e., is analytic and has nonnegative real part in $\mathbb{D}^c := \{z \in \mathbb{C} \mid |z| > 1\}$, and which satisfies the interpolation condition

$$f(z_k) = w_k \quad \text{for } k = 0, 1, \dots, n, \tag{1.1}$$

where $z_0, z_1, \dots, z_n \in \mathbb{D}^c$ and $w_0, w_1, \dots, w_n \in \mathbb{C}$. This problem has a solution if and only if the associated Pick matrix P is positive semidefinite. It is unique if P is singular, and there are infinitely many solutions if $P > 0$ (see [35, 33]). We are interested in a particular subset of these solutions, namely those which are rational of degree at most n, and we shall refer to the problem of determining these as the *Nevanlinna–Pick problem with degree constraints* [12].

For simplicity, in this paper we shall consider the special case that the interpolation points are all distinct and fixed and with $z_0 = \infty$. Then the Pick matrix becomes

$$P = \left[\frac{w_k + \bar{w}_\ell}{1 - z_k^{-1} \bar{z}_\ell^{-1}} \right]_{k,\ell=0}^{n}.$$

[*] Received by the editors February 4, 1999; accepted for publication (in revised form) April 4, 2000; published electronically October 4, 2000. This research was supported in part by grants from AFOSR, TFR, the Göran Gustafsson Foundation, and Southwestern Bell.

http://www.siam.org/journals/sicon/39-3/35111.html

[†] Department of Systems Science and Mathematics, Washington University, St. Louis, MO 63130 (chrisbyrnes@seas.wustl.edu).

[‡] Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, 100 44 Stockholm, Sweden (alq@math.kth.se).

Moreover, we assume that the sets $z_0, z_1, \dots, z_n$ and $w_0, w_1, \dots, w_n$ are self-conjugate so that only real interpolants f need to be considered. We also normalize the problem by setting

$$w_0 = 1$$

so that $f(\infty) = 1$. Finally, we assume that the interpolant is *strictly positive real* in the sense that

$$f(e^{i\theta}) + f(e^{-i\theta}) > 0 \quad \text{for all } \theta \in [-\pi, \pi].$$

Any such function can, in a unique fashion, be written as

$$f(z) + f(z^{-1}) = v(z)v(z^{-1}), \tag{1.2}$$

where v is a *minimum-phase spectral factor* having all zeros in the open unit disc. These zeros will be called *the spectral zeros* of f. As we have remarked above, there are several conformal equivalents of this problem, including Nevanlinna–Pick interpolation for bounded-real, or Schur, functions. Indeed, even for positive real functions there are two conventions, one dealing with interpolation problems inside the unit disc and one outside the disc, as considered here. Our convention is motivated by the desire to have spectral factors which are stable and minimum-phase and therefore may be realized, in control engineering terms, by a stable discrete-time linear system.

We shall show that the space of all strictly positive real, rational functions of at most degree n, $\mathcal{P}_n$, admits two foliations: an *interpolation foliation* with one leaf for each choice of interpolation values $w_1, w_2, \dots, w_n$ satisfying the Pick condition, and a *filtering foliation* with one leaf for each choice of spectral zeros. These foliations are complementary, each pair of leaves with one from each foliation intersecting in one point under nonzero angle. This result is analogous to that obtained in [6] for the case that $z_0 = z_1 = \cdots = z_n = \infty$, the rational covariance extension problem. We note that the corresponding decompositions for the space of functions which are positive real, rather than strictly positive real, are not necessarily disjoint, nor are the equivalence classes necessarily smooth manifolds. For these reasons, we shall work with strictly positive real functions.

More generally, in section 6 we also prove that $\mathcal{P}_n$ is diffeomorphic to $\mathcal{W}_n^+ \times \mathcal{S}_n$, where $\mathcal{W}_n^+$ is the space of all $w_1, w_2, \dots, w_n$ satisfying the Pick condition, and $\mathcal{S}_n$ is the space of (real) *Schur polynomials* of degree n, i.e., real monic polynomials of degree n with all zeros in the open unit disc. Since, in addition, it can be shown that both $\mathcal{W}_n^+$ and $\mathcal{S}_n$ are diffeomorphic to $\mathbb{R}^n$, this implies that $\mathcal{P}_n$ is Euclidean of dimension $2n$.

2. Preliminaries. Let H_2 be the Hardy space of all real functions which are analytic in the exterior of the unit disc, $\mathbb{D}^c := \{z \in \mathbb{C} \mid |z| < 1\}$, and have square-integrable radial limits

$$\lim_{r \to +1} \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(re^{i\theta})|^2 d\theta < \infty$$

on the boundary. Denoting by L_2 the space of all real functions which are square-integrable on the unit circle, we may identify H_2 with the subspace of L_2 consisting of those functions with vanishing positively indexed Fourier coefficients. More precisely, for $f \in H_2$,

$$f(z) = f_0 + f_1 z^{-1} + f_2 z^{-2} + \cdots .$$

Similarly, let $\bar{H}_2$ be the conjugate Hardy space of L_2-functions which are analytic in the open unit disc and thus have vanishing negatively indexed Fourier coefficients so that

$$f(z) = f_0 + f_1 z + f_2 z^2 + \cdots$$

for $f \in \bar{H}_2$. Hence, if $f^*(z) := f(z^{-1})$, $f \in H_2$ if and only if $f^* \in \bar{H}_2$.

The space L_2 is a Hilbert space with inner product

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\theta}) g^*(e^{i\theta}) d\theta.$$

Next, given the interpolation points $z_1, z_2, \dots, z_n$, define the Blaschke product

$$B(z) := \prod_{k=1}^{n} \frac{1 - z_k^{-1} z}{z - \bar{z}_k^{-1}}.$$

As is well known, the subspace BH_2 is invariant under the shift z^{-1}. In order to set notation, we remark that BH_2 is the kernel of the evaluation operator $E : H_2 \to \mathbb{C}^n$ defined

$$E(f) = \begin{bmatrix} f(z_1) \\ \vdots \\ f(z_n) \end{bmatrix},$$

and, if $z_0 = \infty$, $z^{-1}BH_2$ is the kernel of

$$\hat{E}(f) = \begin{bmatrix} f(z_0) \\ f(z_1) \\ \vdots \\ f(z_n) \end{bmatrix}.$$

In this paper, the coinvariant subspaces $H(B) := H_2 \ominus BH_2$,

$$\mathcal{K} := H(z^{-1}B) = H_2 \ominus z^{-1}BH_2, \quad \text{and} \quad \mathcal{L} := z^{-1}H(B) \tag{2.1}$$

will play an important part. They are all finite-dimensional. In fact, given the polynomial

$$\tau(z) = \prod_{k=1}^{n} (z - \bar{z}_k^{-1}), \tag{2.2}$$

$\mathcal{K}$ consists of all rational functions

$$r(z) = \frac{\pi(z)}{\tau(z)}$$

for which the polynomial π is of degree at most n, and hence $\mathcal{K}$ is $(n+1)$-dimensional. The spaces $H(B)$ and $\mathcal{L}$ are n-dimensional subspaces of $\mathcal{K}$. In particular, $\mathcal{L}$ consists of those rational functions $r \in \mathcal{K}$ for which $r(\infty) = 0$. We shall also need the subset $\mathcal{R}$ of functions in $r \in \mathcal{K}$ with the property that $r - 1 \in \mathcal{L}$ and r is *minimum-phase* in the sense that the numerator polynomial π has all its zeros in the open unit disc.

In fact, to say that $r \in \mathcal{R}$ is to say that $\pi \in \mathcal{S}_n$, the n-dimensional space of (monic) Schur polynomials defined in section 1.

Finally, we shall need the subspace

$$\mathcal{Q} := \mathcal{K} + \mathcal{K}^*, \tag{2.3}$$

in terms of which we have the orthogonal decomposition

$$L_2 = zB^*\bar{H}^2 \oplus \mathcal{Q} \oplus z^{-1}BH^2 \tag{2.4}$$

and the subspace $\mathcal{D} \subset \mathcal{Q}$ defined as

$$\mathcal{D} := \{Q = q + q^* \mid q \in \mathcal{K}\}. \tag{2.5}$$

An important convex $(n+1)$-dimensional subset $\mathcal{D}_n^+$ of $\mathcal{D}$ consists of those $D \in \mathcal{D}$ which are positive real, i.e., satisfy the condition that $D(e^{i\theta}) > 0$ for all $\theta \in [-\pi, \pi]$. Also define the n-dimensional subset $\mathcal{Z}_n^+$ of $\mathcal{D}_n^+$ of all $D \in \mathcal{D}_n^+$ which are normalized so that $D(1) = 1$. It is immediately seen that $\mathcal{Z}_n^+$ is also convex.

The following lemma is a trivial modification of the unit circle version of Orlando's formula [15] (also see [5, Lemma 5.5]).

LEMMA 2.1. *Let $a \in \mathcal{R}$, and define $S(a) : \mathcal{K} \to \mathcal{D}$ to be the linear mapping defined by*

$$S(a)v = av^* + a^*v.$$

Then $\ker S(a) = 0$.

3. The interpolation foliation.

3. The interpolation foliation. Any rational function f of degree at most n has a representation

$$f(z) = \frac{b(z)}{a(z)}, \quad a, b \in \mathcal{K}. \tag{3.1}$$

If, in addition, f is strictly positive real, the zeros of the rational functions a and b in (3.1) must be located in the open unit disc. Therefore, if we also assume that $f(\infty) = 1$, it is no restriction to choose $a, b \in \mathcal{R}$. Consequently, we define $\mathcal{P}_n$ to be the space of all pairs (a, b) with $a, b \in \mathcal{R}$ such that f is strictly positive real. The following result was established in [6]. We note that $\mathcal{R}$ is diffeomorphic to Euclidean space $\mathbb{R}^n$ because $\mathcal{S}_n \simeq \mathbb{R}^n$ [4].

PROPOSITION 3.1. *The space $\mathcal{P}_n$ is a smooth, connected, real manifold of dimension $2n$.*

Next, denote by $\mathcal{W}_n^+$ the space of all $w \in \mathbb{C}^n$ with components $w_1, w_2, \dots, w_n \in \mathbb{C}$ satisfying the Pick condition $P > 0$ and forming a self-conjugate set.

PROPOSITION 3.2. *$\mathcal{W}_n^+$ is a smooth, connected, real manifold of dimension n.*

Proof. It is clear that $\mathcal{W}_n^+$ is a smooth manifold having real dimension n. From the form of the Pick matrix, one can also see that $\mathcal{W}_n^+$ is convex and hence connected. □

Let $\eta : \mathcal{P}_n \to \mathcal{W}_n^+$ be the restriction of the evaluation operator E to $\mathcal{P}_n$. Then, for each $w \in \mathcal{W}_n^+$,

$$\mathcal{P}_n(w) = \eta^{-1}(w) \tag{3.2}$$

is the space of all $f \in \mathcal{P}_n$ satisfying the interpolation condition (1.1) corresponding to w.

THEOREM 3.3. *The connected components of the sets $\{\mathcal{P}_n(w) \mid w \in \mathcal{W}_+\}$ form the leaves of an n-dimensional foliation of $\mathcal{P}_n$.*

Remark 3.4. Below, we shall prove that the submanifold $\mathcal{P}_n(w)$ is actually connected. This fact is a nontrivial consequence of the transversality lemma we shall prove in section 5.

To prove this theorem, we need to show that η is a submersion [26], i.e., that the Jacobian $\mathrm{Jac}(\eta)|_{(a,b)}$ is everywhere surjective. To this end, for any $u, v \in \mathcal{L}$, first form the directional derivative of f in the direction (u, v), i.e.,

$$D_{(u,v)}f = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[\frac{b + \epsilon v}{a + \epsilon u} - \frac{b}{a} \right] = \frac{av - bu}{a^2}.$$

Then, the directional derivative of η in the direction (u, v) is

$$D_{(u,v)}\eta = \begin{bmatrix} D_{(u,v)}f(z_1) \\ D_{(u,v)}f(z_2) \\ \vdots \\ D_{(u,v)}f(z_n) \end{bmatrix},$$

which is zero if and only if

$$av - bu = rB, \quad \text{where } r \in \mathcal{L}.$$

Consequently,

$$\ker \mathrm{Jac}(\eta)|_{(a,b)} = \{(u, v) \in \mathcal{L} \times \mathcal{L} \mid av - bu \in B\mathcal{L}\}.$$

LEMMA 3.5. *The tangent space of $\mathcal{P}_n(w)$ at (a, b) has dimension n and is given by*

$$T_{(a,b)}\mathcal{P}_n(w) = \{(u, v) \in \mathcal{L} \times \mathcal{L} \mid av - bu \in B\mathcal{L}\}.$$

Proof. The tangent vectors of $\mathcal{P}_n(w)$, as defined by (3.2), are precisely the vectors in the nullspace of the Jacobian of η at (a, b). For simplicity of notation, denote this space by V. To prove that $\dim V = n$, let $M_{(a,b)} : V \to B\mathcal{L}$ be the mapping $M_{(a,b)}(u, v) = av - bu$. Let n_0 be the number of common zeros of a and b. Then there are three proper rational functions, each taking the value 1 at infinity, namely θ of degree n_0 and $\tilde{a}$ and $\tilde{b}$ of degree $n - n_0$, such that $a = \theta\tilde{a}$ and $b = \theta\tilde{b}$ and $\tilde{a}$ and $\tilde{b}$ have no nontrivial common factors. Now, if $(u, v) \in \ker M_{(a,b)}$, we have $av - bu = 0$, and hence

$$\frac{v}{u} = \frac{b}{a} = \frac{\tilde{b}}{\tilde{a}},$$

so there must be a rational function ϑ of degree n_0 vanishing at infinity such that $u = \vartheta\tilde{a}$ and $v = \vartheta\tilde{b}$. Consequently, since ϑ is completely arbitrary,

$$\dim \ker M_{(a,b)} = n_0.$$

Moreover, for $(u, v) \in V$,

$$av - bu = \theta(\tilde{a}v - \tilde{b}u) = Br \quad \text{for some } r \in \mathcal{L}.$$

Therefore, since $\dim \mathcal{L} = n$ and θ is fixed of degree n_0,

$$\dim M_{(a,b)}(V) = n - n_0.$$

Therefore, by complementarity between rank and nullity,

$$\dim V = \dim M_{(a,b)}(V) + \dim \ker M_{(a,b)} = n,$$

as claimed. □

Proof of Theorem 3.3. Since the Jacobian $\mathrm{Jac}(\eta)$ is a linear map from the $2n$-dimensional tangent space of $\mathcal{P}_n$ to the n-dimensional tangent space of $\mathcal{W}_+$, complementarity of rank and nullity for $\ker \mathrm{Jac}(\eta)$ and the fact that $\dim \ker \mathrm{Jac}(\eta) = n$ (Lemma 3.5) imply that the range of $\mathrm{Jac}(\eta)$ has dimension n. Hence η is a submersion, proving the statement of the theorem [26, p. 2]. □

4. The filtering foliation. The following lemma is a trivial reformulation of results presented in [28, 29] concerning a fast filtering algorithm for Kalman filtering [27] (see also [5]).

LEMMA 4.1. *Given any $(a, b) \in \mathcal{P}_n$, consider the dynamical system*

$$\begin{aligned} a_{t+1}(z) &= \frac{1}{2(1+\gamma_t)}[(1+z)a_t(z) + (1-z)b_t(z)], \quad a_0(z) = a(z), \\ b_{t+1}(z) &= \frac{1}{2(1-\gamma_t)}[(1-z)a_t(z) + (1+z)b_t(z)], \quad b_0(z) = b(z), \end{aligned} \tag{4.1}$$

where

$$\gamma_t = \left(z\frac{b_t(z) - a_t(z)}{2} \right) |_{z=\infty}. \tag{4.2}$$

Then, for $t = 0, 1, 2, \dots,$

$$(a_t, b_t) \in \mathcal{P}_n \tag{4.3}$$

and

$$r_t S(a_t) b_t = S(a)b, \quad \textit{where} \quad r_t = \prod_{k=0}^{t-1}(1 - \gamma_k^2). \tag{4.4}$$

Moreover, as $t \to \infty$, $\gamma_t \to 0$, $r_t \to r_\infty$, and

$$(a_t, b_t) \to (\sigma, \sigma), \quad \textit{where } \sigma \in \mathcal{R}. \tag{4.5}$$

The parameters (4.2) are the Schur parameters (reflection coefficients) corresponding to the function f, and, consequently, $|\gamma_t| < 1, t = 0, 1, 2, \dots,$ whenever f is strictly positive real. The connection to the Schur algorithm and Kalman filtering is explained in the appendix, where, for convenience, an independent proof of Lemma 4.1 is given. For initial conditions $(a, b) \notin \mathcal{P}_n$, the fast filtering algorithm exhibits much more complicated (and interesting) dynamical behavior, which is investigated in detail in [5]. Here, however, we are only interested in its behavior on the set $\mathcal{P}_n$.

In view of (3.1), we have

$$f(z) + f(z^{-1}) = \frac{S(a)b}{aa^*}, \tag{4.6}$$

and hence Lemma 4.1 implies that

$$f(z) + f(z^{-1}) = r_\infty \frac{\sigma(z)\sigma(z^{-1})}{a(z)a(z^{-1})}, \tag{4.7}$$

showing that the spectral factor in (1.2) is

$$v(z) = \sqrt{r_\infty}\frac{\sigma(z)}{a(z)}.$$

We note that $\mathcal{P}_n$ is invariant under the dynamical system (4.1); i.e., whenever the initial condition $(a, b) \in \mathcal{P}_n$, the iterates $(a_t, b_t) \in \mathcal{P}_n$. Moreover, the dynamical system (4.1) converges to the limit point (σ, σ) along the invariant manifold (4.4) [5]. Hence, the equilibrium set is

$$\mathcal{P}_n(\hat{w}), \quad \text{where } \hat{w} := \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{4.8}$$

Furthermore, (4.8) defines the center manifold for the dynamical system (4.1) evolving on $\mathcal{P}_n$, and no equilibrium in the center fold has a nontrivial unstable manifold. The invariant set (4.4) may also be written as

$$\rho_t S(a_t)b_t = S(\sigma)\sigma, \quad \text{where } \rho_t = \Pi_{k=t}^{\infty}(1-\gamma_k^2)^{-1}.$$

Then, for each $\sigma \in \mathcal{R}$,

$$\mathcal{W}^s(\sigma) = \{(a, b) \in \mathcal{P}_n \mid \rho S(a)b = S(\sigma)\sigma \quad \text{for some } \rho \in \mathbb{R}_+\} \tag{4.9}$$

is the stable manifold in $\mathcal{P}_n$ through (σ, σ). In view of (4.6), $S(a)b$ is positive on the unit circle for all $(a, b) \in \mathcal{P}_n$, and hence we can eliminate the variable ρ in $\rho S(a)b = S(\sigma)\sigma$ by dividing by $[S(a)b](1)$. Therefore, we define the mapping $h : \mathcal{P}_n \to \mathcal{Z}_n^+$ as

$$h(a, b) = \frac{S(a)b}{[S(a)b](1)}, \tag{4.10}$$

where $\mathcal{Z}_n^+$ is the n-dimensional convex space defined in section 2. Then, the manifold (4.9) may also be written as

$$\mathcal{W}^s(\sigma) = h^{-1}(\kappa(\sigma)), \tag{4.11}$$

where

$$\kappa(\sigma) := \frac{S(\sigma)\sigma}{[S(\sigma)\sigma](1)} \in \mathcal{Z}_n^+.$$

THEOREM 4.2. *The connected components of the sets* $\{\mathcal{W}^s(\sigma) \mid \sigma \in \mathcal{R}\}$ *form the leaves of an n-dimensional foliation of* $\mathcal{P}_n$.

Remark 4.3. The stable manifolds $\mathcal{W}^s(\sigma)$ are in fact connected. In this paper we shall sketch a proof of this fact based on the transversality lemma.

For the proof we need to show that the Jacobian $\mathrm{Jac}(h)|_{(a,b)}$ has full rank. To this end, we compute the directional derivative of h in the direction (u,v) for arbitrary $u,v \in \mathcal{L}$ as

$$D_{(u,v)}h = \lim_{\epsilon \to 0} \frac{h(a+\epsilon u, b+\epsilon v) - h(a,b)}{\epsilon} = \frac{S(a)q + S(b)p}{[S(a)b](1)}, \tag{4.12}$$

where

$$p = u - \mu b, \quad q = v - \mu a, \quad \mu = \frac{1}{2}\left[\frac{S(a)v + S(b)u}{S(a)b}\right](1). \tag{4.13}$$

In this computation, we have also used the fact that $S(a)b = S(b)a$.

LEMMA 4.4. *The tangent space of* $\mathcal{W}^s(\sigma)$ *at* (a,b) *has dimension* n *and is given by*

$$T_{(a,b)}\mathcal{W}^s(\sigma) = \{(u,v) \in \mathcal{L} \times \mathcal{L} \mid S(a)q + S(b)p = 0\},$$

where $p, q \in \mathcal{K}$ *depend on* (u,v) *as in* (4.13).

Proof. The tangent space $T_{(a,b)}\mathcal{W}(\sigma)$ is precisely the kernel of the Jacobian $\mathrm{Jac}(h)|_{(a,b)}$ of $h^{-1}(\kappa(\sigma))$, i.e., the space of (u,v) for which the directional derivative (4.12) is zero. This yields the expression of the lemma. Since the n algebraic equations contained in

$$h(a,b) = \kappa(\sigma)$$

are obtained by eliminating the variable ρ from the $n+1$ algebraic equations contained in

$$\rho S(a)b = S(\sigma)\sigma,$$

$T_{(a,b)}\mathcal{W}(\sigma)$ has the same dimension as $\ker \mathrm{Jac}(F)|_{(\rho,a,b)}$, where $F : \mathbb{R}_+ \times \mathcal{P}_n \to \mathcal{D}_+$ is defined as

$$F(\rho, a, b) = \rho S(a)b.$$

Now, the directional derivative of F in the direction $(\lambda, u, v) \in \mathbb{R} \times \mathcal{L} \times \mathcal{L}$ is given by

$$D_{(\lambda,u,v)}F(\rho,a,b) = S(a)[\rho v + \lambda b] + S(b)u,$$

so $T_{(a,b)}\mathcal{W}(\sigma)$ has the same dimension as

$$W := \{(r,u) \in \mathcal{K} \times \mathcal{L} \mid S(a)r + S(b)u = 0\}.$$

Then, exactly the same proof as in [5, Lemma 5.11] shows that $\dim W = n$. □

We note in passing that Lemma 5.11 in [5] also shows that if we extend $\mathcal{W}(\sigma)$ outside the positive region $\mathcal{P}_n$ we encounter singularities where the rank of the Jacobian is deficient precisely at the points where a and b have common reciprocal zeros.

Proof of Theorem 4.2. Given Lemma 4.4 and hence that $\ker \mathrm{Jac}(h) = n$, the rest of the proof is completely analogous to that of Theorem 3.3. □

5. The transversality lemma and the geometry of positive real functions. The following result is modeled after the corresponding result in [6, Lemma 4.5].

THEOREM 5.1 (transversality lemma). *Let $\mathcal{K}$ and $\mathcal{L}$ be the spaces defined in* (2.1)*. Then there are no nonzero p and q in $\mathcal{K}$ such that*

$$aq - bp \in B\mathcal{L} \tag{5.1}$$

and

$$S(a)q + S(b)p = 0. \tag{5.2}$$

Proof. We want to prove that if $p \in \mathcal{K}$ and $q \in \mathcal{K}$ satisfy (5.1) and (5.2), then $p = q = 0$. To this end, first note that (5.2) may be written as

$$h(z) + h(z^{-1}) = 0, \tag{5.3}$$

where

$$h(z) := a(z^{-1})q(z) + b(z)p(z^{-1}),$$

and that $h \in \mathcal{Q}$, where $\mathcal{Q}$ is defined by (2.3). Moreover, in view of (5.1),

$$g(z) := \frac{q(z)}{b(z)} - \frac{p(z)}{a(z)} = B(z)\frac{r(z)}{a(z)b(z)}, \quad \text{where } r \in \mathcal{L}.$$

Since $a(\infty) = b(\infty) = 1$ and $r(\infty) = 0$, the rational function $\frac{r}{ab}$ has a Laurent expansion

$$\frac{r(z)}{a(z)b(z)} = c_1 z^{-1} + c_2 z^{-2} + c_3 z^{-3} + \cdots$$

about infinity which holds on and outside the unit circle, and hence $g \in z^{-1}BH^2$. Therefore, $g^* \in zB^*\bar{H}^2$, and, consequently, by (2.4), both g and g^* are orthogonal to $\mathcal{Q}$ and hence to h. In particular,

$$\langle h, g - g^* \rangle = 0. \tag{5.4}$$

Now, a simple calculation shows that

$$g - g^* = \frac{h}{ba^*} - \frac{h^*}{ab^*} = \frac{S(a)b}{aa^*bb^*}h,$$

where (5.3) has been used to obtain the second equality, and therefore (5.4) yields

$$\left\langle h, \frac{S(a)b}{aa^*bb^*}h \right\rangle = 0.$$

However, since $(a, b) \in \mathcal{P}_n$ is positive real, $S(a)b$ is positive on the unit circle, and so is aa^*bb^*. Hence h must be zero, implying that $g = g^*$, i.e., g is constant and thus contained in $\mathcal{Q}$. But g is orthogonal to $\mathcal{Q}$, so g must be zero also. Then

$$q(z) = \frac{b(z)}{a(z)}p(z),$$

which, substituted into (5.2), yields

$$\left[\frac{b}{a}+\frac{b^*}{a^*}\right][ap^*+a^*p]=0.$$

Since $(a,b)\in\mathcal{P}_n$, the first factor is positive on the unit circle, and so

$$a(e^{i\theta})p(e^{-i\theta})+a(e^{-i\theta})p(e^{i\theta})=0$$

for all θ, and therefore, by the identity theorem,

$$S(a)p=0.$$

However, by Lemma 2.1, $S(a)$ has full rank, so p, and hence q, are zero. □

The transversality lemma has the following important consequence.

LEMMA 5.2. *Suppose that the point $(a,b)\in\mathcal{P}_n$ lies on the submanifolds $\mathcal{P}_n(w)$ and $\mathcal{W}^s(\sigma)$. Then*

$$T_{(a,b)}\mathcal{P}_n(w)\cap T_{(a,b)}\mathcal{W}^s(\sigma)=0.$$

Proof. Taking $(u,v)\in T_{(a,b)}\mathcal{P}_n(w)\cap T_{(a,b)}\mathcal{W}(\sigma)$, we see from Lemma 4.4 that (5.2) holds with p and q defined by (4.13). Moreover, since

$$aq-bp=av-\mu ab-bu+\mu ab=av-bu$$

for this choice of p and q, (5.1) also holds by Lemma 3.5. Hence, by Theorem 5.1, we must have $p=q=0$. But then evaluating at ∞, we obtain from (4.13) that $\mu=p(\infty)=q(\infty)$, which a fortiori must be zero, hence implying that $(u,v)=0$. □

It remains to show that the submanifolds $\mathcal{W}^s(\sigma)$ and $\mathcal{P}_n(w)$ are connected and thus constitute the leaves of the filtering foliation and the interpolation foliation, respectively.

COROLLARY 5.3. *The stable manifolds $\{\mathcal{W}^s(\sigma)\mid\sigma\in\mathcal{R}\}$ are diffeomorphic to $\mathcal{W}_n^+$ and thus connected. In particular, the stable manifolds of the fast filtering algorithm* (4.1) *decompose the space $\mathcal{P}_n$ into the leaves of a foliation.*

Proof. Consider again the mapping

$$\eta:\,\mathcal{P}_n\to\mathcal{W}_n^+$$

with $\eta^{-1}(w)=\mathcal{P}_n$. The restriction η_σ of η to $\mathcal{W}^s(\sigma)$ is a map of n-manifolds

$$\eta_\sigma:\,\mathcal{W}^s(\sigma)\to\mathcal{W}_n^+.$$

We claim that

$$\det\operatorname{Jac}(\eta_\sigma)|_{(a,b)}\neq 0$$

for all $(a,b)\in\mathcal{W}^s(\sigma)$. To prove this, we need to show that the directional derivative

$$D_{(u,v)}\eta_\sigma=\operatorname{Jac}(\eta_\sigma)\begin{bmatrix}u\\v\end{bmatrix}$$

is zero for any $(u,v)\in T_{(a,b)}\mathcal{W}^s(\sigma)$ only if $(u,v)=0$. However,

$$\ker\operatorname{Jac}(\eta_\sigma)\subset\ker\operatorname{Jac}(\eta)=T_{(a,b)}\mathcal{P}_n(w)$$

(Lemma 3.5), and hence this follows from Lemma 5.2. To proceed, we also need to show that η_σ is *proper*, i.e., that the inverse image $\eta_\sigma^{-1}(K)$ is compact for each compact set in the range space.

LEMMA 5.4. *The mapping η_σ is proper.*

Proof. Suppose $w_k \to w$ in $\mathcal{W}_n^+$ with $w_k = \eta_\sigma(a_k, b_k)$. Since $\mathcal{P}_n$ and hence $\overline{\mathcal{W}^s(\sigma)}$ are relatively compact, the sequence (a_k, b_k) has a cluster point (a, b) in $\overline{\mathcal{W}^s(\sigma)} \subset \overline{\mathcal{P}_n}$, where a and b have all their zeros in the closed unit disc. We need to show that $(a, b) \in \mathcal{W}^s(\sigma)$. Now, suppose instead that $(a, b) \in \partial\mathcal{W}^s(\sigma)$, the boundary of $\mathcal{W}^s(\sigma)$. Then $(a, b) \in \partial\mathcal{P}_n$. In fact, if $(a, b) \in \mathcal{P}_n$, then, by Theorem 4.2, $(a, b) \in \mathcal{W}^s(\hat{\sigma})$ for some $\hat{\sigma} \in \mathcal{R}$ such that $\hat{\sigma} \neq \sigma$. But then

$$\frac{S(a)b}{[S(a)b](1)} = \kappa(\hat{\sigma}) \neq \kappa(\sigma),$$

which is impossible by continuity. Now, the boundary $\partial\mathcal{P}_n$ consists of those (a, b) for which either $S(a)b$ has a zero on the unit circle or $S(a)b$ is identically zero. Since the zeros of $S(a_k)b_k$ are fixed and therefore independent of k and lie inside the unit disc, $S(a)b$ cannot have zeros on the unit circle without being identically zero. Therefore, the function $f = b/a$ has the property $f + f^* = 0$, and hence f must have all poles and zeros on the unit circle. Then, it is well known [23] and easy to check that f takes the form

$$f(z) = \prod_{k=1}^{m} \frac{z - \mu_k}{z + \mu_k}, \quad |\mu_k| = 1, \quad m \leq n,$$

and, consequently,

$$F(z) = \frac{f(z) - 1}{f(z) + 1}$$

is a Blaschke product of degree m. Thus, modulo a trivial conformal equivalence, Corollary 2.3 in [16, p. 9] states that the rank of the corresponding Pick matrix equals m. Therefore, since $m < n + 1$, the Pick matrix is singular, and the corresponding value vector w must lie in the boundary of $\mathcal{W}_n^+$, contrary to assumption. Consequently, $(a, b) \notin \partial\mathcal{W}^s(\sigma)$, and thus $(a, b) \in \mathcal{W}^s(\sigma)$ as claimed. □

Since η_σ is proper and has a nowhere vanishing Jacobian, $\eta_\sigma^{-1}(w)$ is a finite set with cardinality δ, which is independent of w [30]. Therefore, $\eta_\sigma : \mathcal{W}^s(\sigma) \to \mathcal{W}_n^+$ is a δ-fold covering $\mathcal{W}_n^+$ [30]. Consider the point $\hat{w} \in \mathcal{W}_n^+$ defined by (4.8). For $(a, b) \in \mathcal{W}^s(\sigma)$, to say that $\eta_\sigma(a, b) = \hat{w}$ is to say that $a = b$. Since (a, a) is an equilibrium for the fast filtering algorithm of Lemma 4.1 and lies on the stable manifold of the equilibrium (σ, σ), we must have $(a, a) = (\sigma, \sigma)$, or $(a, b) = (\sigma, \sigma)$. Therefore, $\delta = 1$ and the map $\eta_\sigma : \mathcal{W}^s(\sigma) \to \mathcal{W}_n^+$ is a diffeomorphism. □

COROLLARY 5.5. *The submanifolds $\{\mathcal{P}_n(w) \mid w \in \mathcal{W}_+\}$ are connected. In particular, Nevanlinna–Pick interpolation defines a foliation of the space $\mathcal{P}_n$.*

Proof. Suppose $(a^{(1)}, b^{(1)})$ and $(a^{(2)}, b^{(2)})$ lie in $\mathcal{P}_n(w)$. Since $\mathcal{P}_n$ is connected, there is a continuous path $\gamma : [0, 1] \to \mathcal{P}_n$ with $\gamma(0) = (a^{(1)}, b^{(1)})$ and $\gamma(1) = (a^{(2)}, b^{(2)})$. Composing γ with η, we obtain a closed curve

$$\tilde{\gamma} = \eta \circ \gamma : \quad [0, 1] \to \mathcal{W}_+$$

with initial (and final) point w, i.e., $w = \eta(a^{(i)}, b^{(i)})$, $i = 1, 2$. Since $\mathcal{W}_n^+$ is convex, it is simply connected and therefore $\tilde{\gamma}$ can be contracted to the "constant curve" w

through a homotopy $\tilde{H}$ [22]; i.e.,

$$\tilde{H} : [0,1] \times [0,1] \to \mathcal{W}_+$$

jointly continuous and satisfying

$$\begin{aligned} \tilde{H}(r,0) &= \tilde{\gamma}(r), \\ \tilde{H}(r,1) &= w, \\ \tilde{H}(0,t) &= w, \\ \tilde{H}(1,t) &= w. \end{aligned}$$

We now construct a lifting of the homotopy $\tilde{H}$ to a homotopy H, with values in $\mathcal{P}_n$, covering $\tilde{H}$; i.e., $\eta \circ H = \tilde{H}$. Returning first to the curve γ, each point $\gamma(r)$ lies in a unique stable manifold, which we denote by $\mathcal{W}^s(\sigma(r))$. Since η_σ is a diffeomorphism for each σ, for each r fixed we can lift the curve $\tilde{H}_r$, defined as $\tilde{H}_r(t) = \tilde{H}(r,t)$ for $t \in [0,1]$, to a curve in $\mathcal{W}^s(\sigma(r))$ covering $\tilde{H}_r$ by defining $H_r(t) = \eta^{-1}_{\sigma(r)}(\tilde{H}_r(t))$. Note that H_r is a curve lying in $\mathcal{P}_n$ with initial point $\gamma(r)$. Now define $H : [0,1]\times[0,1] \to \mathcal{P}_n$ via

$$H(r,t) = H_r(t) = \eta^{-1}_{\sigma(r)}(\tilde{H}(r,t)).$$

We claim that H is jointly continuous. To see this, suppose $(r_k, t_k) \to (r,t)$ and set

$$(a_k, b_k) = H(r_k, t_k), \quad (a,b) = H(r,t).$$

We next note that $w_k := \tilde{H}(r_k, t_k) \to \tilde{H}(r,t) =: \tilde{w}$, $\gamma(r_k) \to \sigma(r)$, and consequently that $\sigma(r_k) \to \sigma(r)$, as $k \to \infty$. To prove that H is jointly continuous, it suffices to prove that every neighborhood of (a,b) contains the points (a_k, b_k) for all k sufficiently large. Now, $(a,b) \in \mathcal{P}_n(\tilde{w})$, and, using the implicit function theorem, we can choose neighborhoods $N(a,b)$ which are rectangular in the sense that a neighborhood of (a,b) in $\mathcal{P}_n(\tilde{w})$ serves as the vertical axis, while the horizontal axes consist of unique "slices" consisting of n-manifolds to which the restriction of η will be a diffeomorphism.

That is, the horizontal slices will be open subsets of $\mathcal{W}^s(\sigma)$. Since $\sigma(r_k) \to \sigma(r)$, and since the foliation defined by the stable manifolds of the fast filtering algorithm is itself defined by a submersion, such a neighborhood $N(a,b)$ will intersect $\mathcal{W}^s(\sigma(r_k))$ for all k sufficiently large. Now, (a_k, b_k) is the endpoint of the unique curve $H_{r_k}(t)$ for $t \in [0, t_k]$ in $\mathcal{W}^s(\sigma(r_k))$ covering $\tilde{H}_{r_k}$. Similarly, for any $\bar{t}$ satisfying $0 \le \bar{t} < t_k$, (a_k, b_k) is the endpoint of the unique curve in $\mathcal{W}^s(\sigma(r_k))$ covering $\tilde{H}_{r_k}$ on $[\bar{t}, t_k]$. Since $\eta(N(a,b))$ is open, there exists a $\bar{t}$, $0 \le \bar{t} \le t_k$, for all k sufficiently large so that

$$\tilde{H}_{r_k}[\bar{t}, t_k] \subset \eta(N(a,b)).$$

In particular, since $\eta_{\sigma(r_k)}$ is a (global) diffeomorphism, there exist unique lifts γ_k of these curves in $\mathcal{P}_n$ which lie in $\mathcal{W}^s(\sigma(r_k)) \cap N(a,b)$ and cover $\tilde{H}_{r_k}$ on $[\bar{t}, t_k]$ and have initial points $\eta^{-1}_{\sigma(r_k)}(\tilde{H}(r_k, \bar{t}))$. Since such liftings are unique, it follows that γ_k and H_{r_k} coincide on the subinterval $[\bar{t}, t_k]$, and therefore

$$H_{r_k}(t) \subset N(a,b) \quad \text{for } t \in [\bar{t}, t_k].$$

Consequently,

$$(a_k, b_k) = H_{r_k}(t_k) \in N(a,b)$$

for all k sufficiently large.

We have established that H is jointly continuous. The mapping H also satisfies

$$\begin{aligned} &H(r,0)=\gamma(r),\\ &H(r,1)\subset \mathcal{P}_n(w) \quad \text{for } 0\le r\le 1,\\ &H(0,t)=\gamma(0)=(a^{(1)},b^{(1)}),\\ &H(1,t)=\gamma(1)=(a^{(2)},b^{(2)}). \end{aligned}$$

In particular, $H(\cdot,1)$ is a continuous path in $\mathcal{P}_n(w)$ joining $(a^{(1)},b^{(1)})$ and $(a^{(2)},b^{(2)})$. Since these points are arbitrary in $\mathcal{P}_n(w)$, this manifold is path connected and hence connected. □

Remark 5.6. The foliation by stable manifolds does, of course, define an integrable connection on the distribution tangent to the interpolation foliation, and it is tempting to believe that we can deduce a path-lifting result from the existence of this connection. At this point in the proofs we do not, however, know whether $\eta : \mathcal{P}_n \to \mathcal{W}_n^+$ is a fiber bundle or even a fibration. Moreover, η is definitely not proper, so one could at best expect a path lifting on a sufficiently small subinterval. For this reason, we directly established the homotopy lifting property for curves. We remark that it is possible to go further, showing that $\eta : \mathcal{P}_n \to \mathcal{W}_n^+$ is a fibration. In this case, one could then deduce path connectedness of the fiber from the fact that $\mathcal{W}_n^+$ is simply connected, using the long exact homotopy sequence of the fibration. Since we only needed the sequence for curves and connected components, we instead used a constructive approach to defining the boundary operator in the sequence.

6. Main results.

Another consequence of the transversality lemma is that the leaves of the interpolation foliation intersect the leaves of the filtering foliation transversely; i.e., the two foliations are complementary. Actually, a much deeper relationship exists between these foliations, having several interesting corollaries.

THEOREM 6.1. *The filtering foliation and the interpolation foliation are complementary. Moreover, each leaf $\mathcal{P}_n(w)$ intersects each leaf $\mathcal{W}^s(\sigma)$ of the filtering foliation in one, and only one, point in $\mathcal{P}_n$.*

The first assertion follows immediately from Lemma 5.2 after it has been established that $\mathcal{P}_n(w)$ and $\mathcal{W}^s(\sigma)$ are connected and so are the leaves of respective foliation (Corollary 5.5 and Corollary 5.3). Consequently, there are two complementary foliations of $\mathcal{P}_n$, namely,

$$\mathcal{F}_1 : \quad \mathcal{P}_n = \bigcup_{w\in\mathcal{W}_+} \mathcal{P}_n(w), \tag{6.1}$$

indexed by the interpolation values $w \in \mathcal{W}_+$, and

$$\mathcal{F}_2 : \quad \mathcal{P}_n = \bigcup_{\sigma\in\mathcal{R}} \mathcal{W}(\sigma), \tag{6.2}$$

indexed by the equilibrium points (4.8) of the dynamical system (4.1), or, equivalently, by the spectral zeros in the form of a point in $\mathcal{S}_n$. This suggests that, given a set of admissible interpolation values and a set of stable spectral zeros, there is a unique solution of the Nevanlinna–Pick problem represented by the intersection between the corresponding leaves of the foliations $\mathcal{F}_1$ and $\mathcal{F}_2$. This is precisely the second assertion of the theorem and is a consequence of Proposition 6.3 to be proven below.

To this end, first note that the fact that the filtering foliation and the interpolation foliation are complementary says that this uniqueness does occur to first order, in the following sense.

LEMMA 6.2. *Let $h_w : \mathcal{P}_n(w) \to \mathcal{Z}_n^+$ be the restriction of h, defined by* (4.10), *to $\mathcal{P}_n(w)$. Then, for each $(a,b) \in \mathcal{P}_n(w)$, the Jacobian matrix Jac(h_w) of h_w is nonsingular.*

Proof. To prove this, we need to show that the directional derivative

$$D_{(u,v)}h = \mathrm{Jac}(h_w)\begin{bmatrix} u \\ v \end{bmatrix}$$

is zero for any $(u,v) \in T_{(a,b)}\mathcal{P}_n(w)$ only if $(u,v) = 0$. But this follows from Lemma 5.2 precisely as in the proof of Corollary 5.3. $\square$

It is interesting to note that the duality between interpolation and filtering is reflected in a symmetry between the restricted mappings

$$\eta_\sigma : \mathcal{W}^s(\sigma) \to \mathcal{W}_n^+$$

and

$$h_w : \mathcal{P}_n(w) \to \mathcal{Z}_n^+.$$

Recall that η_σ is the restriction of $\eta : \mathcal{P}_n \to \mathcal{W}_n^+$ to $\mathcal{W}^s(\sigma) = h^{-1}(\kappa(\sigma))$, and h_w is the restriction of $h : \mathcal{P}_n \to \mathcal{Z}_n^+$ to $\mathcal{P}_n(w) = \eta^{-1}(w)$. Moreover, we have the following result.

PROPOSITION 6.3. *The mappings η_σ and h_w are diffeomorphisms. In particular, each choice of σ and w determines and is determined by precisely one element $(a,b) \in \mathcal{P}_n$.*

Proof. We have already shown in the proof of Corollary 5.3 that η_σ is a diffeomorphism. Concerning h_w, we first establish properness.

LEMMA 6.4. *The mapping h_w is proper.*

Proof. To show this, consider a sequence (κ_k) in $\mathcal{Z}_n^+$ with $\kappa_k = h_w(a_k, b_k)$ which converges to $\kappa \in \mathcal{Z}_n^+$ as $k \to \infty$, and prove that any cluster point (a,b) of (a_k,b_k) lies in $\mathcal{P}_n(w)$. Since $\mathcal{P}_n$ is relatively compact, $(a,b) \in \overline{\mathcal{P}_n(w)} \in \overline{\mathcal{P}_n}$. Now, suppose (a,b) is not in $\mathcal{P}_n(w)$ but in the boundary $\partial\mathcal{P}_n(w)$. Then $(a,b) \in \partial\mathcal{P}_n$ because if $(a,b) \in \mathcal{P}_n$, then, by Theorem 3.3, $(a,b) \in \mathcal{P}_n(\hat{w})$ for some $\hat{w} \neq w$, which contradicts continuity of $\eta(a,b)$. But if $(a,b) \in \partial\mathcal{P}_n$, then $S(a)b$ either has a zero on the unit circle or is identically zero, while $a, b \in \mathcal{R}$ of course remain nonzero. Therefore, if there is no zero at $z = 1$, $\kappa_k \to \partial\mathcal{Z}_n^+$ and if $[S(a)b](1) = 0$, then $\kappa_k \to \infty$, contradicting the assumption that $\kappa \in \mathcal{Z}_n^+$ in both cases. $\square$

Since h_w is a proper map with nonvanishing Jacobian (Lemma 6.2), $h_w : \mathcal{P}_n(w) \to \mathcal{Z}_n^+$ is a δ-fold covering. Since $\mathcal{P}_n(w)$ is connected and $\mathcal{Z}_n^+$ is convex, and hence simply connected, the number, δ, of sheets must be one [30]. Therefore, h_w is a diffeomorphism. $\square$

This concludes the proof of Theorem 6.1.

These geometric implications of the transversality lemma allow us to give an alternative geometric proof and amplification of the following result in [20], where, however, spectral zeros on the unit circle are also allowed, and in [12], using convex analysis to the minima of a functional defined using generalized entropy gains.

COROLLARY 6.5 (spectral zero assignability theorem for Nevanlinna–Pick interpolation). *Suppose w determines a positive definite Pick matrix. The positive real* *interpolants* (a, b) *in* $\mathcal{P}_n(w)$ *can be uniquely determined by a choice of stable spectral zeros.*

This corollary shows that the spectral zeros are design parameters which can be used, for example, in designing robust bounded real closed loop systems. This result also holds in the case that all the interpolation points $z_0 = z_1 = \cdots = z_n = \infty$, a situation of great interest in signal processing, spectral analysis, and stochastic systems [6, 7, 8, 9, 10, 11] (see also [17, 18], where the first proofs of existence were presented). In this case, the design parameter is intuitively very appealing, since it represents a choice of zeros for shaping filters which can shape white noise into a process matching a finite window of covariance data.

The following theorem, finally, is also a consequence of the transversality lemma. Here $\simeq$ denotes "diffeomorphic," and $\mathcal{S}_n$ is the space of Schur polynomials of degree n introduced in section 1.

THEOREM 6.6. *The space* $\mathcal{P}_n$ *is Euclidean of dimension* $2n$. *More specifically,*

$$\mathcal{P}_n \simeq \mathcal{W}_n^+ \times \mathcal{Z}_n^+ \simeq \mathcal{W}_n^+ \times \mathcal{S}_n,$$

where $\mathcal{W}_n^+$, $\mathcal{Z}_n^+$, *and* $\mathcal{S}_n$ *are all diffeomorphic to* $\mathbb{R}^n$.

For the proof we need the following "folk theorem," for which we have been unable to find a direct reference.

LEMMA 6.7. *An open, convex set* $D \in \mathbb{R}^n$ *is diffeomorphic to* $\mathbb{R}^n$.

It is well known and easy to see that an open convex set $D \in \mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$ [3, p. 2]. Except for $n = 4$, this implies that D is also diffeomorphic to $\mathbb{R}^n$, so the problem is only for $n = 4$ [31, p. 5]. However, convexity gives us much more, and it is simpler to give a direct proof. The following is an outline of a proof provided by O. Viro.[1]

Convexity allows us to construct a C^∞-function $\varphi : \mathbb{R}^n \to [0, 1]$, such that $\varphi(0) = 1$, $\varphi(x) > 0$ in D, and $\varphi(x) = 0$ outside D, which is monotonely nondecreasing along any ray $\{\lambda y \mid \|y\| = 1, \lambda \geq 0\}$. (We place the origin inside D.) In fact, for each supporting hyperplane H_k to D, one can construct a function φ_k which is zero in the half-space not containing D and which is monotonely nonincreasing along the normal direction from the origin, with the value one on a parallel hyperplane $\hat{H}_k$ and in the whole half-space beyond it. If D is a polytope, there are finitely many supporting hyperplanes H_k, and we may take $\varphi(x) = \prod_k \varphi_k(x)$. In general, we choose the hyperplanes H_k on a dense set of the boundary and let the distances d_k between each pair $\hat{H}_k$ and H_k be a sequence which tends to zero. Then, only a finite number of φ_k are different from one at any point in D, and hence the construction still works. The function $\psi : D \to \mathbb{R}^n$, with $\psi(0) = 1$ and $\psi(x) = x/\varphi(x)$ otherwise, is then a diffeomorphism. In fact, the monotonicity implies that the Jacobian does not vanish in D.

Proof of Theorem 6.6. Since $\mathcal{W}_n^+$ and $\mathcal{Z}_n^+$ are open and convex sets, they are diffeomorphic to $\mathbb{R}^n$ by Lemma 6.7. In the case of $\mathcal{Z}_n^+$, this can also be seen from the facts that $\mathcal{Z}_n^+ \simeq \mathcal{S}_n$ [6, p. 1849] and $\mathcal{S}_n \simeq \mathbb{R}^n$ [4]. Then, the rest follows from Proposition 6.3. □

Appendix. Fast Kalman filtering and the Schur algorithm.

Modulo a trivial reformulation, Lemma 4.1 is proven in [27, 28, 29] in the context of Kalman filtering, using the Szegö polynomials orthogonal on the unit circle and the Levinson recursion. Obviously, the recursion (4.1) is related to the Schur algorithm

[1] Similar ideas of a proof have also been suggested to us by H. Shapiro and M. Benedicks.

[1], as was established in, for example, [13]. In this appendix, we give a simple proof in this context.

Given any $(a,b) \in \mathcal{P}_n$ and the corresponding strictly positive real function $f = b/a$, define

$$\varphi(z) := \frac{f(z)-1}{f(z)+1} = \frac{b(z)-a(z)}{b(z)+a(z)} =: \frac{P(z)}{Q(z)},$$

which is a *Schur function* in the sense that it maps the exterior of the unit disc, $\mathbb{D}^c$, into the open unit disc $\mathbb{D}$. The Schur algorithm

$$\varphi_{t+1}(z) = z\frac{\varphi_t(z)-\varphi_t(\infty)}{1-\varphi_t(\infty)\varphi_t(z)}, \quad \varphi_0(z) = \varphi(z), \tag{A.1}$$

defines a sequence $\varphi_t(z)$, $t = 0,1,2,\dots$, of Schur functions, and the Schur parameters

$$\gamma_t = \varphi_{t+1}(\infty), \quad t = 0,1,2,\dots, \tag{A.2}$$

are less than one in modulus [1].

Proposition A.1. *For* $t = 0,1,2,\dots$,

$$\varphi_{t+1}(z) = \frac{zP_t(z)}{Q_t(z)}, \tag{A.3}$$

where P_t *and* Q_t *are polynomials satisfying the recursions*

$$\begin{cases} Q_{t+1}(z) = Q_t(z) - \gamma_t z P_t(z), & Q_0(z) = Q(z), \\ P_{t+1}(z) = zP_t(z) - \gamma_t Q_t(z), & P_0(z) = P(z). \end{cases} \tag{A.4}$$

Here Q_t *is of degree* n *having leading coefficient*

$$r_t = \prod_{k=0}^{t-1}(1-\gamma_k^2). \tag{A.5}$$

Proof. Clearly,

$$\varphi_1(z) = z\varphi_0(z) = \frac{zP(z)}{Q(z)},$$

so (A.3) holds for $t = 0$. Now let $t \geq 1$, and suppose that

$$\varphi_t(z) = \frac{zP_{t-1}(z)}{Q_{t-1}(z)}.$$

Then, the Schur algorithm (A.1) together with (A.2) yields

$$\varphi_{t+1}(z) = z\frac{zP_{t-1}(z) - \gamma_{t-1}Q_{t-1}(z)}{Q_{t-1}(z) - \gamma_{t-1}zP_{t-1}(z)} = \frac{zP_t(z)}{Q_t(z)},$$

and hence (A.3) holds for $t = 1,2,\dots$ by induction. Moreover, (A.3) and (A.4) yield

$$\frac{Q_{t+1}(z)}{Q_t(z)} = 1 - \gamma_t\varphi_{t+1}(z),$$

which, evaluated at $z = \infty$, becomes $1 - \gamma_t^2$ by (A.2). But $|\gamma_t| < 1$, and hence $\deg Q_{t+1} = n$ whenever $\deg Q_t = n$. Since $\deg Q_0 = n$, it thus follows by induction that $\deg Q_t = n$ for $t = 0, 1, 2, \ldots$. More precisely, r_t, given by (A.5), is the leading coefficient of $Q_t(z)$. $\square$

The recursion (A.4) is precisely the fast algorithm for Kalman filtering [27] in the formulation of [28]. In fact, suppose $\{y_0, y_1, y_2, \ldots\}$ is a stationary stochastic process with spectral density

$$\Phi(z) = \frac{1}{2}[f(z) + f(z^{-1})],$$

where $f = b/a$ has a minimal realization

$$f(z) = 1 + 2h(zI - F)^{-1}g.$$

Then the linear least squares estimate $\hat{y}_t$ of y_t given $y_0, y_1, \ldots, y_{t-1}$ is generated by the Kalman filter

$$\begin{cases} \hat{x}_{t+1} = F\hat{x}_t + k_t(y_t - \hat{y}_t), \\ \hat{y}_t = h\hat{x}_t, \end{cases}$$

where k_t is determined from $Q_t(z)$ in the following way: If (F, g, h) is chosen so that $h = (1, 0, \ldots, 0)$, F has characteristic polynomial $\chi_F(z) = z^n + \alpha_1 z^{n-1} + \cdots + \alpha_n$, and

$$Q_t(z) = r_t[z^n + q_1(t)z^{n-1} + \cdots + q_n(t)],$$

then the gain k_t is given by

$$k_t = q(t) - \alpha.$$

Moreover, Q_t/r_t is the characteristic polynomial of the feedback matrix

$$F - k_t h, \tag{A.6}$$

which hence is stable.

In the same way as in [28], a direct calculation using (A.4) yields

$$\begin{aligned} &Q_{t+1}(z)Q_{t+1}(z^{-1}) - P_{t+1}(z)P_{t+1}(z^{-1}) \\ &\qquad = (1 - \gamma_t^2)[Q_t(z)Q_t(z^{-1}) - P_t(z)P_t(z^{-1})] \end{aligned} \tag{A.7}$$

for $t = 0, 1, 2, \ldots$.

Now set

$$a_t(z) := \frac{Q_t(z) - P_t(z)}{2r_t\tau(z)}, \quad b_t(z) := \frac{Q_t(z) + P_t(z)}{2r_t\tau(z)}. \tag{A.8}$$

We first note that $a_0 = a$ and $b_0 = b$. Moreover, since $|z^{-1}| < 1$ in $\mathbb{D}^c$, $z^{-1}\varphi_{t+1}(z)$ is a Schur function, and, consequently,

$$f_t(z) := \frac{b_t(z)}{a_t(z)} = \frac{1 + z^{-1}\varphi_{t+1}(z)}{1 - z^{-1}\varphi_{t+1}(z)} \tag{A.9}$$

is strictly positive real for $t = 0, 1, 2, \ldots$ so that $(a_t, b_t) \in \mathcal{P}_n$. This verifies (4.3).

From (A.4) we readily obtain the recursion (4.1), Moreover, in view of (A.9),

$$\varphi_{t+1}(z) = z\frac{b_t(z) - a_t(z)}{b_t(z) + a_t(z)},$$

and hence (4.2) follows from (A.2). We also note that (4.4) is equivalent to (A.7). It now only remains to verify (4.5). To this end, we recall that, for rational positive real functions, the Schur parameters form an ℓ_2 sequence [18, p. 447], and hence $\gamma_t \to 0$ as $t \to \infty$. Consequently, r_t tends to some limit r_∞ as $t \to \infty$, and it follows from (A.4) that $Q_t(z)$ tends to a constant polynomial Q_∞, which is the characteristic polynomial of the steady-state feedback matrix (A.6) defined by the steady-state Kalman gain. Hence $r_\infty^{-1} Q_\infty \in \mathcal{S}_n$. It also follows that $P_t(z)$ tends to zero. Therefore, by (A.8), a_t and b_t tend to σ as $t \to \infty$, where

$$\sigma(z) = \frac{Q_\infty(z)}{r_\infty \tau(z)}.$$

Clearly, $\sigma \in \mathfrak{R}$, as claimed.

This proves the claims made in Lemma 4.1. In [5], a much more refined analysis of the global phase portrait of the fast filtering algorithm is given, with the explicit derivation of the global stable manifolds which we employ in section 6. This analysis has many other consequences. For example, it can be shown [6] that for a rational strictly positive real function the sequence of Schur parameters decays to zero geometrically, generalizing previous results in the literature on conditional and absolute summability of the corresponding series of Schur parameters [21, 2, 18].

Acknowledgments. We would like to thank Professors H. Shapiro, M. Benedicks, and O. Viro for technical advice.

REFERENCES

[1] M. Bakonyi and T. Constantinescu, *Schur's Algorithm and Several Applications*, Pitman Res. Notes Math. Ser. 261, Longman, Harlow, John Wiley and Sons, New York, 1992.

[2] A. Bultheel, *Convergence of Schur parameters and transmission zeros of a meromorphic spectrum*, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 281–296.

[3] H. Busemann, *Convex Surfaces*, Interscience Publishers, New York, 1958.

[4] C.I. Byrnes and A. Lindquist, *On the geometry of the Kimura-Georgiou parameterization of modelling filters*, Internat. J. Control, 50 (1989), pp. 2301–2312.

[5] C. I. Byrnes, A. Lindquist, and Y. Zhou, *On the nonlinear dynamics of fast filtering algorithms*, SIAM J. Control Optim., 32 (1994), pp. 744–789.

[6] C. I. Byrnes, A. Lindquist, S.V. Gusev, and A. S. Matveev, *A complete parametrization of all positive rational extensions of a covariance sequence*, IEEE Trans. Automat. Control, AC-40 (1995), pp. 1841–1857.

[7] C. I. Byrnes and A. Lindquist, *On the partial stochastic realization problem*, IEEE Trans. Automat. Control, AC-42 (1997), pp. 1049–1070.

[8] C. I. Byrnes, H. J. Landau, and A. Lindquist, *On the well-posedness of the rational covariance extension problem*, in Current and Future Directions in Applied Mathematics, M. Alber, B. Hu, and J. Rosenthal, eds., Birkhäuser Boston, Boston, 1997, pp. 83–108.

[9] C. I. Byrnes and A. Lindquist, *On duality between filtering and control*, in Systems and Control in the Twenty-First Century, C. I. Byrnes, B. N. Datta, D. S. Gilliam, and C. F. Martin, eds., Birkhäuser Boston, Boston, 1997, pp. 101–136.

[10] C. I. Byrnes, S.V. Gusev, and A. Lindquist, *A convex optimization approach to the rational covariance extension problem,* SIAM J. Control Optim., 37 (1998), pp. 211–229.

[11] C. I. Byrnes, P. Enqvist, and A. Lindquist, *Cepstral coefficients, covariance lags and pole-zero models for finite data strings,* IEEE Trans. Signal Process., submitted.

[12] C.I. Byrnes, T.T. Georgiou, and A. Lindquist, *A generalized entropy criterion for Nevanlinna–Pick interpolation: A convex optimization approach to certain problems in systems and control*, IEEE Trans. Automat. Control, to appear.

[13] C. Chang and T. T. Georgiou, *The Schur algorithm: Connections with the difference Riccati equation, the Chandrasekhar algorithm, and Square-root filtering*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods, M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, eds., Birkhäuser Boston, Boston, 1990, pp. 207–213.

[14] Ph. Delsarte, Y. Genin, and Y. Kamp, *On the role of the Nevanlinna–Pick problem in circuits and system theory*, Internat. J. Circuit Theory Appl., 9 (1981), pp. 177–187.

[15] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[16] J. B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[17] T.T. Georgiou, *Partial Realization of Covariance Sequences*, Ph.D. thesis, Center for Mathematical Systems Theory, University of Florida, Gainesville, FL, 1983.

[18] T.T. Georgiou, *Realization of power spectra from partial covariance sequences*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-35 (1987), pp. 438–449.

[19] T. T. Georgiou, *A topological approach to Nevanlinna–Pick interpolation*, SIAM J. Math. Anal., 18 (1987), pp. 1248–1260.

[20] T. T. Georgiou, *The interpolation problem with a degree constraint*, IEEE Trans. Automat. Control, 44 (1999), pp. 631–635.

[21] L. Ya. Geronimus, *Orthogonal Polynomials: Estimates, Asymptotic Formulas, and Series of Polynomials Orthogonal on the Unit Circle and on an Interval*, Consultants Bureau, New York, 1961.

[22] M. Greenberg, *Lectures on Algebraic Topology*, Benjamin, New York, 1967.

[23] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, Los Angeles, CA, 1958.

[24] S. Haykin, *Nonlinear Methods of Spectral Analysis*, Springer-Verlag, Berlin, New York, 1983.

[25] H. Helson, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.

[26] H. Blaine Lawson, Jr., *The Quantitative Theory of Foliations*, CBMS Reg. Conf. Ser. Math. 27, Amer. Math. Soc., Providence, RI, 1977.

[27] A. Lindquist, *A new algorithm for optimal filtering of discrete-time stationary processes*, SIAM J. Control, 12 (1974), pp. 736–746.

[28] A. Lindquist, *Some reduced-order non-Riccati equations for linear least-squares estimation: The stationary, single-output case*, Internat. J. Control, 24 (1976), pp. 821–842.

[29] A. Lindquist, *Linear least-squares prediction based on covariance data from stationary processes with finite-dimensional realizations*, in Proceedings of the Second European Congress on Operations Research, Stockholm, Sweden, North-Holland, Amsterdam, 1976, pp. 281–286.

[30] J. W. Milnor, *Topology from the Differentiable Viewpoint*, University Press of Virginia, Charlottesville, VA, 1965.

[31] T. Petrie and J. Randall, *Connections, Definite Forms, and Four-Manifolds*, Oxford Science Publications, Clarendon Press, Oxford University Press, New York, 1990.

[32] B. Porat, *Digital Processing of Random Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[33] D. Sarason, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.

[34] I. Schur, *On power series which are bounded in the interior of the unit circle* I *and* II, J. Reine Angew. Math., 148 (1918), pp. 122–145.

[35] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, Amer. Math. Soc. Colloq. Publ. 20, Amer. Math. Soc., Providence, RI, 1956.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 776–797

OPTIMALITY OF ENERGY ESTIMATES FOR THE WAVE EQUATION WITH NONLINEAR BOUNDARY VELOCITY FEEDBACKS*

JUDITH VANCOSTENOBLE† AND PATRICK MARTINEZ†

Abstract. We consider the wave equation damped by a nonlinear boundary velocity feedback $q(u_t)$.

First we consider the case where q has a linear growth at infinity. We prove that the usual decay rate estimates proved by Nakao [*Differential Integral Equations,* 8 (1995), pp. 681–688], Haraux and Zuazua [*Arch. Rational Mech. Anal.,* 100 (1988), pp. 191–206], Conrad, Leblond, and Marmorat [in *Proceedings of the Fifth International Federation of Automatic Control Symposium on Control of Distributed Parameter Systems*, Perpignan, 1989, pp. 101–116], Zuazua [*SIAM J. Control Optim.,* 28 (1990), pp. 466–477], and Komornik [in *Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena, Internat. Ser. Numer. Math.* 118, Birkhäuser, Basel, 1994, pp. 253–266] when q has a polynomial behavior at zero and by the second author [*ESAIM Control Optim. Calc. Var.,* 4 (1999), pp. 419–444] in the general case are in fact optimal in one space dimension. More generally, we prove that the energy decays exactly like the solution of an explicit and simple ordinary differential equation.

Next we study the problem when q is bounded at infinity. We prove that strong solutions decay exponentially to zero, and we exhibit a sequence of weak solutions for which the associated energy decays to zero at infinity as slowly as the iterated logarithms go to infinity at infinity.

Key words. damped wave equation, asymptotic behavior, boundary feedback, optimal energy estimates

AMS subject classifications. 35L05, 35B35, 35B40

PII. S0363012999354211

1. Introduction.

We consider the wave equation in one space dimension, damped by a boundary velocity feedback $q(u_t)$, where $q : \mathbb{R} \longrightarrow \mathbb{R}$ is a nonlinear function:

$$(1.1)\qquad \begin{cases} u_{tt} - u_{xx} = 0, & x \in (0,1),\ t \geq 0, \\ u(0,t) = 0, & t \geq 0, \\ u_x(1,t) = -q(u_t(1,t)), & t \geq 0, \\ (u(x,0), u_t(x,0)) = (u^0(x), v^0(x)), & x \in (0,1), \end{cases}$$

where $(u^0(x), v^0(x))$ is given in $V \times L^2(0,1)$ (with $V = \{v \in H^1(0,1) \mid v(0) = 0\}$). We refer to [22] for the physical motivations of this model.

We define the energy of u by

$$\forall t \geq 0, \quad E_u(t) = \frac{1}{2}\int_0^1 (u_x^2(x,t) + u_t^2(x,t))\,dx.$$

First we study the problem of asymptotic decay of $E_u(t)$ as $t \to +\infty$. When $q : \mathbb{R} \longrightarrow \mathbb{R}$ is a continuous nondecreasing function, has a polynomial behavior at zero, and a linear growth at infinity, it means that when there exist four positive

*Received by the editors March 29, 1999; accepted for publication (in revised form) April 27, 2000; published electronically October 4, 2000. This work was done while the authors were working at I.R.M.A.R., Université Rennes I and École Normale Supérieure de Cachan, Antenne de Bretagne.
http://www.siam.org/journals/sicon/39-3/35421.html

†M.I.P., Université Paul Sabatier Toulouse III, 118 route de Narbonne, 31 062 Toulouse Cedex 4, France (vancoste@mip.ups-tlse.fr, martinez@mip.ups-tlse.fr).

constants C_1, C_2, C_3, and C_4 such that

$$\begin{cases} \forall |s| \le 1,\ C_1|s|^p \le |q(s)| \le C_2|s|^{1/p} \quad \text{with } p>1, \\ \forall |s| \ge 1,\ C_3|s| \le |q(s)| \le C_4|s|, \end{cases} \tag{1.2}$$

then $E_u(t)$ satisfies the estimate

$$\forall t \ge 0, \quad E_u(t) \le \frac{C(E_u(0))}{(1+t)^{2/(p-1)}}, \tag{1.3}$$

where $C(E_u(0)) > 0$ is a constant which depends on $E_u(0)$. (See [23, 10, 5] for (1.1) and see also [17, 6, 9, 4, 2, 11, 19, 12, 20] for the similar problem of the wave equation damped by a distributed nonlinear feedback (2.2). More generally, these authors studied the case of higher space dimensions.)

More recently, the previous results were completed in [15] with explicit estimates under weaker assumptions than (1.2) on the function q (in particular, when q is weaker than any polynomial in zero; see also Lasiecka and Tataru [13] and Liu and Zuazua [14]). For example, we proved that, if $q(s) = \operatorname{sgn}(s)\, e^{-1/|s|}$ in a neighborhood of zero (with linear growth at infinity), then $E_u(t)$ satisfies the estimate

$$\forall t \ge 2, \quad E_u(t) \le \frac{C}{(\ln t)^2}. \tag{1.4}$$

The purpose of this article is to establish the optimality of these estimates in case of (1.1), i.e., in the case of a boundary feedback in one space dimension. Note that we use here the term of "optimality of estimates" in the sense of "two-sides estimates." (For example, in (1.3), we wonder if $2/(p-1)$ is the best choice of the exponent. In some cases, our method will also lead to the best choice of the factor $C(E_u(0))$ but it is not the main object of this paper.) First, we treat the polynomial case with a method announced in [21]: for particular solutions (we choose particular initial conditions which simplify the problem), this method produces an explicit equivalent of the energy and proves the optimality of (1.3). We exhibit some solutions u of (1.1) that satisfy

$$E_u(t) \underset{t\to+\infty}{\sim} \frac{C_p}{t^{2/(p-1)}}$$

(see Theorem 2.1). Then we extend the preceding result in the case of general feedbacks. In general, we prove that the solutions decay exactly like the solutions of a very simple ordinary differential equation. Moreover, under some additional assumptions, we give lower bounds of the energy, which prove, in particular, the optimality of estimates like (1.3) and (1.4) (see Theorem 3.1, and Propositions 3.1, 3.2, and 3.3). Our method also gives similar results of optimality for other examples: we can treat the one-dimensional wave equation with boundary feedbacks at both extremities and also the wave equation with a boundary feedback *in dimension* 3.

Next we turn to another optimality problem: in a previous paper [16], we studied the decay rate of *strong* solutions of the wave equation damped with an internal nonlinear weak feedback (that means $\frac{q(s)}{s} \longrightarrow 0$ when $|s| \longrightarrow +\infty$) in space dimension 2. We proved that, if $q'(0) \neq 0$, *strong* solutions decay exponentially to zero, but with a decay rate depending on the norm of the initial conditions in $H^2(\Omega) \times H_0^1(\Omega)$. Haraux and Conrad asked us if the decay rate really depends on that quantity, and what can be said about *weak* solutions.

Here we study this question in the case of the one-dimensional wave equation with boundary damping (1.1): with a carefully chosen function q, which satisfies

$$q'(0) \neq 0 \quad \text{and} \quad q(s) \longrightarrow \pm 1 \quad \text{when} \quad s \longrightarrow \pm\infty,$$

we prove that *strong* solutions decay exponentially but *not uniformly* to zero: more precisely, we prove that given (u^0, v^0) in $W^{1,\infty}(0,1) \times L^\infty(0,1)$ (and $u^0(0) = 0$), there exists some (explicit) constant $C^0 = C^0(\|(u^0, v^0)\|_{W^{1,\infty} \times L^\infty})$ such that

$$\forall t \geq 0, \; E_u(t) \leq C^0 \, 3^{-t};$$

and on the other hand we construct a sequence $(u_m)_{m \geq 1}$ of solutions of (1.1), corresponding to initial conditions $({u_m}^0, {v_m}^0)$ in $W^{1,\infty}(0,1) \times L^\infty(0,1)$ that satisfy

$$E_{u_m}(0) = 1 \quad \text{and} \quad \|({u_m}^0, {v_m}^0)\|_{W^{1,\infty} \times L^\infty} = 2m,$$

and

$$\forall t \geq m, \; E_{u_m}(t) = \frac{9^m}{9m^2} \, 3^{-t};$$

hence the exponential decrease of the energy to zero cannot be uniform.

Concerning *weak* solutions, we consider, motivated by a recent work of Cannarsa, Komornik, and Loreti [3], the sequence of iterated logarithms

$$\begin{cases} \ln_1(t) = \ln(t), \\ \ln_{p+1}(t) = \ln(\ln_p(t)); \end{cases}$$

given $p \geq 1$, we exhibit a weak solution u_p of (1.1) such that the energy of u_p decays very slowly to zero. Indeed, for t large enough, we have (see Theorem 4.1)

$$E_{u_p}(t) \geq \frac{1}{\ln_p(t)}.$$

The paper is organized as follows: in section 2, we treat the polynomial case; in section 3, we extend the previous result in the case of general feedbacks; in section 4, we study the case of a special weak dissipation; the last section is devoted to the proofs of the results.

2. Polynomial feedbacks.

2.1. Optimality of the energy estimates for (1.1). We assume that $q : \mathbb{R} \longrightarrow \mathbb{R}$ is a continuous nondecreasing function given by

$$\forall s \in [-s_0, s_0], \quad q(s) = s\,|s|^{p-1}, \quad \text{with} \quad p > 1, \tag{2.1}$$

in a neighborhood $[-s_0, s_0]$ of zero, and that q has a linear growth at infinity. For all initial data $(u^0(x), v^0(x)) \in V \times L^2(0,1)$, there exists a unique solution u of (1.1) with the regularity $u \in \mathcal{C}(\mathbb{R}_+, V) \cap \mathcal{C}^1(\mathbb{R}_+, L^2(0,1))$. This solution u verifies the estimate (1.3).

There are very few results of optimality. Although this has not been proved until now, it was reasonable to think that the estimates (1.3) were optimal (at least for some particular solutions). Haraux [8] obtained partial results on the problem of the wave equation damped by a distributed feedback:

$$\begin{array}{ll} u_{tt} - u_{xx} = -q(u_t), & x \in (0,1), \; t \geq 0, \\ u(0,t) = u(1,t) = 0, & t \geq 0, \\ (u(x,0), u_t(x,0)) = (u^0(x), v^0(x)), & x \in (0,1), \end{array} \tag{2.2}$$

where $(u^0(x), v^0(x))$ is given in $H^1_0(0,1) \times L^2(0,1)$, where $H^1_0(0,1) := \{u \in H^1(0,1) \mid u(0) = 0 = u(1)\}$. For smooth solutions, he proved that

$$\limsup_{t \to +\infty} \; E_u(t)(1+t)^{3/(p-1)} > 0.$$

However, the problem of optimality is unsolved, since we would like to obtain (at least for some solutions)

$$\limsup_{t \to +\infty} \; E_u(t)(1+t)^{2/(p-1)} > 0. \tag{2.3}$$

Haraux [8] noted also that, when we change the Dirichlet conditions in (2.2) by the homogeneous Neumann conditions $u_x(0,t) = u_x(1,t) = 0$ and when $q(s) = s|s|^{p-1}$, then the nontrivial spatially homogeneous solutions decay precisely like $t^{-1/(p-1)}$ as $t \to +\infty$. (See also Haraux [7, p. 46] for another result of optimality for an ordinary differential equation.)

Concerning (1.1), no result of optimality was known before (even in the case of dimension 1). Note that Aassila [1] obtained estimates like (1.3) and proved their optimality for a nonlinear wave equation (with a nonlinear term in u' and a nonlinear term in ∇u), but his proof cannot be extended to (1.1) or (2.2).

We prove that (1.3) is optimal for particular initial conditions.

Theorem 2.1. *Assume* (2.1). *For* $(u^0(x), v^0(x)) = (2A_0 x, 0)$, *with* $A_0 \in \mathbb{R}$, $A_0 \neq 0$, *the solution* u *of* (1.1) *satisfies*

$$E_u(t) \underset{t \to +\infty}{\sim} \frac{C_p}{t^{2/(p-1)}}, \qquad \textit{with} \;\; C_p = \frac{1}{2\,(p-1)^{2/(p-1)}}. \tag{2.4}$$

Remarks. 1. Theorem 2.1 is still true if we replace (2.1) by

$$q(0) = 0 \;\; \text{and} \quad \forall s \in [-s_0, s_0],\; s \neq 0, \quad q(s) = \operatorname{sgn}(s)\,|s|^{1/p}, \text{ with } p > 1. \tag{2.5}$$

(For all $s \neq 0$, we denote $\operatorname{sgn}(s)$ the sign of s.) Note that the two inverse functions defined by (2.1) and (2.5) lead exactly to the same equivalent for the energy.

2. In fact, (2.3) is true for more general initial conditions (see Proposition 3.2 and the following remarks).

3. The proof of (2.4) does not yield any information on the time T_1 for which we can say that if $t \geq T_1$, $E_u(t)$ is close to $C_p t^{-2/(p-1)}$. From a practical point of view, this could also be interesting. In Propositions 3.1 and 3.2, we will give explicit lower bounds for the energy, which gives us explicit constants $c_0 > 0$ and $T_0 > 0$ such that

$$\forall t \geq T_0,\; E_u(t) \geq \frac{c_0}{t^{2/(p-1)}}.$$

2.2. Other examples. We can also study the damped wave equation with feedbacks at both extremities: let $q_1,\ q_2\ : \mathbb{R} \longrightarrow \mathbb{R}$ be two continuous nondecreasing functions and consider the following system:

$$\begin{cases} u_{tt} - u_{xx} = 0, & x \in (0,1),\; t \geq 0, \\ u_x(0,t) = q_1(u_t(0,t)), & t \geq 0, \\ u_x(1,t) = -q_2(u_t(1,t)), & t \geq 0, \\ (u(x,0), u_t(x,0)) = (u^0(x), v^0(x)), & x \in (0,1), \end{cases} \tag{2.6}$$

where $(u^0(x), v^0(x))$ is given in $H^1(0,1) \times L^2(0,1)$.

We define the energy of the solution u by

$$\forall t \geq 0, \quad E_u(t) = \frac{1}{2} \int_0^1 (u_x^2(x,t) + u_t^2(x,t))\, dx.$$

Assume that

$$\forall s \in [-s_0, s_0], \quad q_1(s) = s\, |s|^{p_1 - 1}, \quad \text{with} \quad p_1 > 1, \tag{2.7}$$

and

$$\forall s \in [-s_0, s_0], \quad q_2(s) = s\, |s|^{p_2 - 1}, \quad \text{with} \quad p_2 > 1. \tag{2.8}$$

Then we have the following.

Proposition 2.1. *Assume* (2.7) *and* (2.8). *For* $(u^0(x), v^0(x)) = (2A_0 x, 0)$, *with* $A_0 \in \mathbb{R}$, $A_0 \neq 0$, *the solution* u *of* (2.6) *satisfies the following:*

(i) *if* $p_1 = p_2 = p$,

$$E_u(t) \underset{t \to +\infty}{\sim} \frac{C'_p}{t^{2/(p-1)}}, \qquad \textit{with } C'_p = \frac{2}{(2^p\,(p-1))^{2/(p-1)}}; \tag{2.9}$$

(ii) *if* $p_1 \neq p_2 = p$,

$$E_u(t) \underset{t \to +\infty}{\sim} \frac{C_{p_0}}{t^{2/(p_0-1)}}, \qquad \textit{with } C_{p_0} = \frac{1}{2\,(p_0 - 1)^{2/(p_0-1)}}, \tag{2.10}$$

where $p_0 = \min(p_1, p_2)$.

Remark. The result of Proposition 2.1 holds true if we replace the assumptions (2.7) and (2.8) by the assumptions (2.11) and (2.8), or by (2.7) and (2.12), or by (2.11) and (2.12) with

$$q_1(0) = 0 \ \text{ and } \ \forall s \in [-s_0, s_0],\ s \neq 0,\ \ q_1(s) = \operatorname{sgn}(s)\, |s|^{1/p_1}, \tag{2.11}$$

and

$$q_2(0) = 0 \ \text{ and } \ \forall s \in [-s_0, s_0],\ s \neq 0,\ \ q_2(s) = \operatorname{sgn}(s)\, |s|^{1/p_2} \tag{2.12}$$

(always with $p_1 > 1$ and $p_2 > 1$).

2.3. Example in dimension 3. With the same method, we also obtain similar results for the wave equation with boundary feedback *in dimension* 3, which show that the previous results of optimality are not specific to dimension 1.

In space dimension 3, set $B_1 = B(0,1)$ and $B_2 = B(0,2)$ and let Ω be the open subset $B_2 \setminus B_1$ of $\mathbb{R}^3$ with boundary $\Gamma = \partial B_1 \cup \partial B_2$. Then we consider the following system:

$$\begin{cases} u_{tt} - \Delta u = 0, & x \in \Omega,\ t \geq 0, \\ u = 0, & x \in \partial B_1,\ t \geq 0, \\ \partial_\nu u + \frac{1}{2} u = -q(u_t), & x \in \partial B_2,\ t \geq 0, \\ (u(x,0), u_t(x,0)) = (u^0(x), v^0(x)), & x \in \Omega, \end{cases} \tag{2.13}$$

where $(u^0(x), v^0(x))$ is given in $\tilde{V} \times L^2(\Omega)$ (with $\tilde{V} = \{v \in H^1(\Omega) \mid v_{|\Gamma_1} = 0\}$).

Under (1.2), the energy E_u of the solution satisfies the estimate (1.3) (see, e.g., Zuazua [23]). Here we prove that (1.3) is optimal.

Proposition 2.2. *Assume that q is odd and satisfies* (2.1). *Then there exist* $(u^0, v^0) \in \tilde{V} \times L^2(\Omega)$ *and* $\tilde{C}_p > 0$ *such that*

$$E_u(t) \underset{t\to+\infty}{\sim} \frac{\tilde{C}_p}{t^{2/(p-1)}}. \tag{2.14}$$

(Note that we construct *radial* solutions satisfying (2.14).)

3. General feedbacks.

3.1. Known estimates.

When the feedback q has no polynomial behavior in zero, like (1.2), and has a linear growth at infinity, Lasiecka and Tataru [13] prove that the energy decays as fast as the solution of some associated differential equation. In [14], Liu and Zuazua give another proof of this result, which provides a simpler dissipative ordinary differential equation describing the decay rate.

In subsection 3.2, we will prove that the solutions of (1.1) decay exactly like the solution of an ordinary differential equation (which is very simple in this case).

On the other hand, in [15], we obtained explicit estimates of the decay rate of the energy. Let us recall these estimates: assume that $q : \mathbb{R} \longrightarrow \mathbb{R}$ is a strictly continuous increasing function (with linear growth at infinity) such that

$$|g(s)| \le |q(s)| \le |g^{-1}(s)|$$

in a neighborhood of zero, where $g : \mathbb{R} \longrightarrow \mathbb{R}$ is a strictly increasing and odd function of class $\mathcal{C}^1$ (and where g^{-1} denotes the inverse function of g). Set $H(s) = g(s)/s$ and assume that $H(0) = 0$ and that H is increasing on $[0, \eta]$ for some $\eta > 0$. Then, for (1.1), $E_u(t)$ satisfies the estimate

$$\forall t \ge 1, \quad E_u(t) \le C(E_u(0)) \left[g^{-1}\left(\frac{1}{t}\right)\right]^2. \tag{3.1}$$

For example, if g is defined on some $[0, \eta]$ by

$$g(s) = e^{-1/y^p} \tag{3.2}$$

for some $p > 0$, then

$$\forall t \ge 2, \quad E_u(t) \le \frac{C}{(\ln t)^{2/p}}. \tag{3.3}$$

And if

$$g(s) = e^{-e^{1/y}}, \tag{3.4}$$

then

$$\forall t \ge 3, \quad E_u(t) \le \frac{C}{(\ln(\ln t))^2}. \tag{3.5}$$

(Note that (3.1) does not directly give the optimal estimate (1.3) when g has a polynomial behavior in zero, but still allows us to get it after some additional computations.) In subsection 3.3, we prove that (3.3) and (3.5) are optimal. And more generally, in subsection 3.4, we prove that (3.1) is optimal for a class of nonpolynomial feedbacks (including (3.2) and (3.4)).

3.2. Behavior of the energy in the general case.

THEOREM 3.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing and odd function of class $\mathcal{C}^1$ such that $g(0) = 0$ and $g'(0) = 0$. Assume $q : \mathbb{R} \to \mathbb{R}$ is a continuous nondecreasing function such that*

$$q = g \quad \text{or} \quad q = g^{-1}$$

in a neighborhood of zero.

For all $(u^0(x), v^0(x)) = (2A_0x, 0)$ with $A_0 \in \mathbb{R}$, $A_0 \neq 0$, the solution u of (1.1) *satisfies*

$$\exists n_0 \in \mathbb{N},\ \forall n \geq n_0, \quad E_u(2n) = \frac{1}{2}V(2t_n)^2, \tag{3.6}$$

where $(t_n)_{n \geq n_0}$ is a real positive increasing sequence such that $t_n \sim n$ as $n \to +\infty$, and where $V : \mathbb{R}_+ \to \mathbb{R}_+$ is the solution of the ordinary differential equation

$$V'(s) = -g(V(s)), \quad s \geq 0, \tag{3.7}$$

with $V(0) = 2\sqrt{E_u(2n_0)/2}$.

Remarks. 1. Note that $q = g$ and $q = g^{-1}$ lead exactly to the same estimate.

2. By solving the differential equation (3.7), we find that the estimates (1.3), (3.3), and (3.5) are optimal. However, this needs some computations, which is why we give two explicit lower bound of the energy in the following.

3.3. Some explicit lower bounds of the energy. It is interesting to study if the result of Theorem 2.1 can be extended to more general initial conditions (which means if (2.3) holds true for more general initial conditions) and to have an explicit lower bound of the energy when q has no polynomial growth in zero. In the following, we provide two formulae that allow us to prove the optimality of the estimates of (1.3), (3.3), and (3.5) even when we consider more general initial conditions than the special ones used in Theorem 2.1.

PROPOSITION 3.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing, odd, and convex function of class $\mathcal{C}^1$ such that $g(0) = 0$ and $g'(0) = 0$. Assume $q : \mathbb{R} \to \mathbb{R}$ is a continuous nondecreasing function such that*

$$|q(s)| \leq |g(s)| \quad \text{or} \quad |q(s)| \geq |g^{-1}(s)|$$

$\forall s$ in a neighborhood of 0. *Moreover we denote $h = \frac{1}{2}g^{-1}$ and we assume that*

$$s \mapsto s(h'(s) - 1) \quad \text{is increasing in a neighborhood of zero.} \tag{3.8}$$

Then $\forall (u^0(x), v^0(x)) = (2A_0x, 0)$ with $A_0 \in \mathbb{R}$, $A_0 \neq 0$, the solution u of (1.1) *satisfies*

$$\exists n_0, n_1 \in \mathbb{N},\ \forall n \geq n_0, \quad E_u(2n) \geq \frac{1}{2}\left[(g')^{-1}\left(\frac{1}{2(n+n_1)}\right)\right]^2. \tag{3.9}$$

Remark. It is easy to check that (3.8) is satisfied when

$$g(s) = s^p \quad \text{or} \quad g(s) = e^{-1/s^p} \quad \text{or} \quad g(s) = e^{-e^{1/s}}$$

on $[0, \eta]$ for some $\eta > 0$; then (3.9) gives the optimality of the corresponding estimates (1.3), (3.3), and (3.5) if q has a linear growth at infinity.

Since it may be not easy to compute $(g')^{-1}$, we give a simpler lower bound of the energy.

PROPOSITION 3.2. *Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing and convex function of class $\mathcal{C}^1$ such that $g(0) = 0$ and $g'(0) = 0$. Assume $q : \mathbb{R} \to \mathbb{R}$ is a continuous nondecreasing function such that*

$$|q(s)| \leq |g(s)| \quad \text{or} \quad |q(s)| \geq |g^{-1}(s)|$$

$\forall s$ in a neighborhood of 0. Moreover we denote $h = \frac{1}{2}g^{-1}$ and we assume that there exist $\beta > 0$ and $k \geq 1$ such that

$$\forall n \geq 1,\ h'\left(\frac{\beta}{n^k}\right) \geq n + 1. \tag{3.10}$$

For all $(u^0(x), v^0(x)) = (2A_0 x, 0)$ with $A_0 \in \mathbb{R},\ A_0 \neq 0$, the solution u of (1.1) *satisfies*

$$\exists \gamma > 0,\ \exists n_0 \in \mathbb{N},\ \forall n \geq n_0, \quad E_u(2n) \geq \frac{1}{2}\left[g^{-1}\left(\frac{\gamma}{n^k}\right)\right]^2. \tag{3.11}$$

As an application of Proposition 3.2, we can easily check the optimality of (1.3) and also of (3.3), and (3.5): it is sufficient to choose $k = \frac{p}{p-1}$ when $g(s) = s^p$ and $k = 2$ in the two other examples.

Remark. With the same proof, we can also obtain similar results for more general initial conditions $(u^0, v^0) \in (W^{1,\infty}(0,1) \cap V) \times L^\infty(0,1)$ provided that $\|(u^0, v^0)\|_{W^{1,\infty}\times L^\infty}$ is small enough.

3.4. Optimality of estimate (3.1) for a class of nonpolynomial feedbacks. At last, we prove that the estimate (3.1) is optimal for a class of nonpolynomial feedbacks (including (3.2) and (3.4)): we have the following.

PROPOSITION 3.3. *Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing and convex function of class $\mathcal{C}^1$ such that $g(0) = 0$ and $g'(0) = 0$. Assume $q : \mathbb{R} \to \mathbb{R}$ is a continuous nondecreasing function such that*

$$|q(s)| \leq |g(s)| \quad \text{or} \quad |q(s)| \geq |g^{-1}(s)|$$

$\forall s$ in a neighborhood of 0. Moreover, we assume that there exist $\alpha > 0$ and $M > 0$ such that

$$2\alpha \frac{g(2s)g'(\alpha s)}{g(\alpha s)^2} \leq M. \tag{3.12}$$

For all $(u^0(x), v^0(x)) = (2A_0 x, 0)$ with $A_0 \in \mathbb{R},\ A_0 \neq 0$, the solution u of (1.1) *satisfies*

$$\exists C > 0,\ \exists n_0 \in \mathbb{N},\ \forall n \geq n_0, \quad E_u(2n) \geq \frac{2}{\alpha^2}\left[g^{-1}\left(\frac{1}{Mn + C}\right)\right]^2. \tag{3.13}$$

3.5. Remark concerning (2.2). Even in the polynomial case, we have no result of optimality of energy estimates for (2.2). However, if we change the Dirichlet conditions by the homogeneous Neumann conditions $u_x(0,t) = u_x(1,t) = 0$, then we can prove similar results of optimality as for (1.1).

We consider the nontrivial homogeneous solutions of

$$\left\{ \begin{array}{ll} u_{tt} - u_{xx} = -q(u_t), & x \in (0,1),\ t \geq 0, \\ u_x(0,t) = u_x(1,t) = 0, & t \geq 0. \end{array} \right. \tag{3.14}$$

We set $v = u_t$. Then $v(t)$ is solution of the ordinary differential equation

$$v'(t) = -q(v(t)), \quad t \geq 0, \tag{3.15}$$

which is the same as (3.7) when $q = g$. In particular, under the hypotheses of Theorem 3.1 and Propositions 3.1 and 3.2, we can easily obtain similar lower bounds of the energy. (Note that when $q = g^{-1}$, the asymptotic behavior is different: in this case v converges to 0 in finite time.)

4. Decay rate estimates when the feedback is weak at infinity. In a previous paper (see [16]), we studied the behavior of the energy of the solutions of the following problem:

$$\left\{ \begin{array}{ll} u_{tt} - \Delta u = -q(u_t), & x \in \Omega,\ t \geq 0, \\ u(y,t) = 0, & y \in \partial\Omega,\ t \geq 0, \\ (u(x,0), u_t(x,0)) = (u^0(x), v^0(x)), & x \in \Omega, \end{array} \right. \tag{4.1}$$

where Ω is a bounded domain of class $\mathcal{C}^2$ of $\mathbb{R}^2$ and q is a nondecreasing function that satisfies

$$\frac{q(s)}{s} \longrightarrow 0 \text{ when } |s| \longrightarrow +\infty$$

(in this case, the feedback is called *weak*). When $q'(0) \neq 0$, we proved that the energy of strong solutions decays exponentially to zero, but with a decay rate depending on the norm of the initial conditions in $H^2(\Omega) \times H_0^1(\Omega)$: given $(u^0, v^0) \in H^2(\Omega) \times H_0^1(\Omega)$, there exists $\omega = \omega(\|(u^0, v^0)\|_{H^2(\Omega) \times H_0^1(\Omega)}) > 0$ such that

$$\forall t \geq 0,\ E(t) \leq E(0) e^{1-\omega t}.$$

This result improved previous results of Komornik [11] and of Nakao [18] (who also treated more general cases). Haraux and Conrad asked us if the decay rate really depends on the quantity $\|(u^0, v^0)\|_{H^2(\Omega) \times H_0^1(\Omega)}$. We are able to give some answers in the case of (1.1): consider the special function q defined by

$$\left\{ \begin{array}{l} \forall |s| \leq 2,\ q(s) = \frac{s}{2}, \\ \forall |s| \geq 2,\ q(s) = \text{sgn }(s) 1. \end{array} \right. \tag{4.2}$$

We also consider the sequence of iterated logarithms

$$\left\{ \begin{array}{l} \forall t > 1,\ \ln_1(t) = \ln(t), \\ \forall t > T_{p+1},\ \ln_{p+1}(t) = \ln(\ln_p(t)), \end{array} \right. \tag{4.3}$$

where $(T_p)_p$ is defined by

$$\left\{ \begin{array}{l} T_1 = 1, \\ T_{p+1} = e^{T_p}. \end{array} \right.$$

The functions $\ln_p$ are well defined on $[T_p, +\infty[$ and go slowly to infinity at infinity.

We have the following.

THEOREM 4.1. 1. *Let q be the function defined by* (4.2). *If $(u^0, v^0) \in W^{1,\infty}(0,1) \times L^\infty(0,1)$, then the energy of the strong solution of* (1.1) *decays exponentially to zero, but not uniformly with respect to (u^0, v^0).*

2. *Given $p \geq 1$, there exist $(u^0, v^0) \in V \times L^2(0,1)$, $T'_p > T_p$, such that the energy of the associated solution u of* (1.1) *satisfies*

$$\forall t \geq T'_p,\ E_u(t) \geq \frac{1}{\ln_p(t)}. \tag{4.4}$$

Remarks. 1. This result allows us to measure the gap between the decrease of the energy of strong solutions and the one of weak solutions. Therefore, when the feedback is weak at infinity, the behavior of q at infinity has a great importance on the decrease of the energy.

2. We conjecture that the following stronger result is also true: given $f : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ a decreasing function that goes to zero at infinity, there exist $(u^0, v^0) \in H^1_0(0,1) \times L^2(0,1)$ and $\varepsilon > 0$ such that the energy of the associated solution u of (1.1) satisfies

$$\forall t \geq 0,\ E_u(t) \geq \varepsilon f(t).$$

3. More precisely about the nonuniform exponential decay of the energy of strong solutions, we prove that given (u^0, v^0) in $W^{1,\infty}(0,1) \times L^\infty(0,1)$ (and $u^0(0) = 0$), there exists some (explicit) constant $C^0 = C^0(\|(u^0, v^0)\|_{W^{1,\infty} \times L^\infty})$ such that

$$\forall t \geq 0,\ E_u(t) \leq C^0\ 3^{-t};$$

and on the other hand we construct a sequence $(u_m)_{m \geq 1}$ of solutions of (1.1), corresponding to initial conditions $({u_m}^0, {v_m}^0)$ in $W^{1,\infty}(0,1) \times L^\infty(0,1)$ that satisfy

$$E_{u_m}(0) = 1 \text{ and } \|({u_m}^0, {v_m}^0)\|_{W^{1,\infty} \times L^\infty} = 2m,$$

and

$$\forall t \geq m,\ E_{u_m}(t) = \frac{9^m}{9m^2}\ 3^{-t};$$

hence the exponential decrease of the energy to zero cannot be uniform.

4. The proof of (4.4) is based on the construction of explicit special initial conditions.

5. Proofs.

5.1. Polynomial feedbacks. In order to prove Theorem 2.1 and Proposition 2.1, we first prove the following.

LEMMA 5.1. *We denote by $F : \mathbb{R} \longrightarrow \mathbb{R}$ the continuous strictly increasing function defined by $F = \mathrm{Id} + q$ where q satisfies* (2.1) *or* (2.5). *Then we have the following:*

(i) *if q satisfies* (2.1),

$$F^{-1}(r) = r - r|r|^{p-1} + o\left(|r|^p\right) \quad \text{as } r \to 0,$$

(ii) *if q satisfies* (2.5),

$$F^{-1}(r) = r|r|^{p-1} + o\left(|r|^p\right) \quad \text{as } r \to 0.$$

Remark. Note that in both cases, this implies that

$$(r+F^{-1}(-2r))^2 = r^2\Big(1-2^p|r|^{p-1}+\mathrm{o}\left(|r|^{p-1}\right)\Big)^2$$
$$= r^2\Big(1-2^{p+1}|r|^{p-1}+\mathrm{o}\left(|r|^{p-1}\right)\Big), \quad \text{as } r\to 0.$$

We will see in the proof that the behavior of the solution is closely related to the behavior at 0 of the function $r \longrightarrow (r+F^{-1}(-2r))^2$. In particular, since their behavior at 0 is the same in both cases, we will obtain exactly the same equivalent of the energy in both cases.

Proof of Lemma 5.1.

(i) Assume (2.1). Set $r = F(s) = s + s|s|^{p-1}$ $\forall s\in\mathbb{R}$. Note that $r\to 0$ as $s\to 0$ and

$$\frac{r}{F^{-1}(r)} = \frac{F(s)}{s} = 1+|s|^{p-1} \underset{s\to 0}{\to} 1.$$

Thus $F^{-1}(r) = r + o(|r|)$ as $r\to 0$. And since

$$r = s + \operatorname{sgn}(s)\,|s|^p = F^{-1}(r) + \operatorname{sgn}(r)\,|F^{-1}(r)|^p,$$

we obtain

$$F^{-1}(r) = r - \operatorname{sgn}(r)\,|r|^p + o(|r|^p), \quad \text{as } r\to 0.$$

(ii) Assume (2.5). Set $r = F(s) = s + \operatorname{sgn}(s)\,|s|^{1/p}$ $\forall s\in\mathbb{R}$. Then

$$\frac{F^{-1}(r)}{\operatorname{sgn}(r)\,|r|^p} = \frac{s}{\operatorname{sgn}(s)\,|F(s)|^p} = \left(\frac{|s|^{1/p}}{\left|s+\operatorname{sgn}(s)\,|s|^{1/p}\right|}\right)^p \underset{s\to 0}{\to} 1.$$

Thus

$$F^{-1}(r) = \operatorname{sgn}(r)\,|r|^p + o(|r|^p), \quad \text{as } r\to 0. \qquad \square$$

Proof of Theorem 2.1. Let $A_0\in\mathbb{R}$ be such that $A_0\neq 0$. First we introduce the real sequence $(A_n)_{n\in\mathbb{N}}$ defined by

$$\forall n\in\mathbb{N},\quad A_{n+1}+A_n = -q(A_{n+1}-A_n). \tag{5.1}$$

This sequence is well defined since (5.1) can also be written like

$$\forall n\in\mathbb{N},\ A_{n+1} = A_n + F^{-1}(-2A_n), \tag{5.2}$$

where $F:\mathbb{R}\longrightarrow\mathbb{R}$ is the continuous strictly increasing function defined by $F = \mathrm{Id}+q$.

Assume for a moment the following.

LEMMA 5.2.

$$A_n^2 \underset{n\to+\infty}{\sim} \frac{C_p'}{n^{2/(p-1)}}, \qquad \textit{with} \quad C_p' = \left(\frac{1}{2^p\,(p-1)}\right)^{2/(p-1)}.$$

Then we define an absolutely continuous function $f:(-1,+\infty)\longrightarrow\mathbb{R}$ such that

$$\forall n\in\mathbb{N},\quad \forall s\in(2n-1,2n+1),\quad f'(s) = A_n.$$

Note that (5.1) implies

$$f'(t+1)+f'(t-1)=-q(f'(t+1)-f'(t-1)) \tag{5.3}$$

almost everywhere (a.e.) $t\geq 0$.

Then, u defined by

$$u(x,t)=f(t+x)-f(t-x),\quad (x,t)\in(0,1)\times(0,\infty),$$

is the solution of problem (1.1). (Relation (5.3) gives $u_x(1,t)=-q(u_t(1,t))$.) And the energy of u is given by

$$\forall t\geq 0,\quad E_u(t)=\frac{1}{2}\int_0^1 (u_x^2(x,t)+u_t^2(x,t))\,dx=\int_{-1}^1 f'(t+s)^2\,ds.$$

In particular, Lemma 5.2 gives

$$E_u(2n)=\int_{2n-1}^{2n+1} f'(s)^2\,ds=2A_n^2 \underset{n\to+\infty}{\sim} \frac{2\,C_p'}{n^{2/(p-1)}}.$$

And since $t\mapsto E_u(t)$ is nonincreasing, we deduce

$$E_u(t) \underset{t\to+\infty}{\sim} \frac{C_p}{t^{2/(p-1)}}\quad \text{with } C_p=\frac{1}{2(p-1)^{2/(p-1)}}. \qquad \square$$

Proof of Lemma 5.2. First, let us prove that $A_n \underset{n\to+\infty}{\to} 0$. From (5.1), we have, $\forall n\in\mathbb{N}$,

$$A_{n+1}^2-A_n^2=-(A_{n+1}-A_n)\,q(A_{n+1}-A_n)\ \leq\ 0. \tag{5.4}$$

The sequence $(A_n^2)_{n\in\mathbb{N}}$ is nonincreasing and convergent. Thus $A_{n+1}^2-A_n^2 \underset{n\to+\infty}{\to} 0$. Relation (5.4) implies $(A_{n+1}-A_n)\,q(A_{n+1}-A_n) \underset{n\to+\infty}{\to} 0$ and consequently $A_{n+1}-A_n \underset{n\to+\infty}{\to} 0$. From (5.1), we also deduce $A_{n+1}+A_n \underset{n\to+\infty}{\to} 0$, and finally, $A_n \underset{n\to+\infty}{\to} 0$.

Then (5.2) and Lemma 5.1 give

$$A_{n+1}=-A_n+2^pA_n|A_n|^{p-1}+\mathrm{o}\,(|A_n|^p),\quad n\to+\infty,$$

when q is given by (2.1). (Note that in this case, for some $n_0\in\mathbb{N}$ large enough, the sign of the sequence $(A_n)_{n\geq n_0}$ is alternating.) On the other hand Lemma 5.1 gives

$$A_{n+1}=A_n-2^pA_n|A_n|^{p-1}+\mathrm{o}\,(|A_n|^p),\quad n\to+\infty,$$

when q is given by (2.5). (In this case, for some $n_0'\in\mathbb{N}$ large enough, the sign of the sequence $(A_n)_{n\geq n_0'}$ is constant.)

In both cases, we obtain

$$\begin{aligned} A_{n+1}^2 &= A_n^2\Big(1-2^p|A_n|^{p-1}+\mathrm{o}\,(|A_n|^{p-1})\Big)^2 \\ &= A_n^2\Big(1-2^{p+1}|A_n|^{p-1}+\mathrm{o}\,(|A_n|^{p-1})\Big). \end{aligned} \tag{5.5}$$

Note that (5.5) implies in particular that $|A_{n+1}| \underset{n\to+\infty}{\sim} |A_n|$. We deduce

$$\frac{1}{(A_{n+1})^{2\ (p-1)/2}} - \frac{1}{(A_n)^{2\ (p-1)/2}} = \frac{1 - \left(1 - 2^{p+1}|A_n|^{p-1} + \mathrm{o}\left(|A_n|^{p-1}\right)\right)^{(p-1)/2}}{|A_{n+1}|^{p-1}}$$

$$= \frac{\frac{p-1}{2}\, 2^{p+1}\, |A_n|^{p-1} + \mathrm{o}\left(|A_n|^{p-1}\right)}{|A_{n+1}|^{p-1}} \underset{n\to+\infty}{\to} 2^p\,(p-1).$$

Finally, Cesàro's theorem gives

$$\frac{1}{(A_n)^{2\ (p-1)/2}} \underset{n\to+\infty}{\sim} \sum_{k=0}^{n-1} \left(\frac{1}{(A_{k+1})^{2\ (p-1)/2}} - \frac{1}{(A_k)^{2\ (p-1)/2}} \right) \underset{n\to+\infty}{\sim} 2^p(p-1)\,n,$$

which proves Lemma 5.2. □

Idea of the proof of Proposition 2.1. Let $A_0 \in \mathbb{R}$ be such that $A_0 \neq 0$. We introduce the two real sequences $(A_n)_{n\geq 0}$ and $(B_n)_{n\geq -1}$ defined by $B_{-1} = A_0$ and

$$\forall n \geq 0, \quad A_n - B_n = q_1(A_n + B_n), \tag{5.6}$$

$$\forall n \geq -1, \quad A_{n+2} - B_n = -q_2(A_{n+2} + B_n). \tag{5.7}$$

We denote by F_1 and $F_2 : \mathbb{R} \to \mathbb{R}$ the two continuous strictly increasing functions defined by $F_1 = \mathrm{Id} + q_1$ and $F_2 = \mathrm{Id} + q_2$. Then (5.6) and (5.7) become

$$\forall n \geq 0,\ B_n = -A_n + F_1^{-1}(2A_n), \tag{5.8}$$

$$\forall n \geq -1,\ A_{n+2} = -B_n + F_2^{-1}(2B_n). \tag{5.9}$$

We define $f : (0,+\infty) \longrightarrow \mathbb{R}$ and $g : (-1,+\infty) \longrightarrow \mathbb{R}$ two absolutely continuous functions such that

$$\forall n \geq 0, \quad \forall s \in (n, n+1), \quad f'(s) = A_n,$$

and

$$\forall n \geq -1, \quad \forall s \in (n, n+1), \quad g'(s) = B_n.$$

Note that (5.6) and (5.7) imply

$$f'(t) - g'(t) = q_1(f'(t) + g'(t)) \quad \text{a.e. } t \geq 0, \tag{5.10}$$

and

$$f'(t+1) - g'(t-1) = -q_2(f'(t+1) + g'(t-1)) \quad \text{a.e. } t \geq 0. \tag{5.11}$$

Then, u defined by

$$u(x,t) = f(t+x) + g(t-x), \quad (x,t) \in (0,1) \times (0,\infty),$$

is the solution of problem (2.6). (Relation (5.10) gives $u_x(0,t) = q_1(u_t(0,t))$ and (5.11) gives $u_x(1,t) = -q_2(u_t(1,t))$.) The energy of u is given by

$$\forall t \geq 0, \quad E_u(t) = \frac{1}{2} \int_0^1 f'(t+x)^2 + g'(t-x)^2\, dx = \int_t^{t+1} f'(s)^2\, ds + \int_{t-1}^{t} g'(s)^2\, ds.$$

In particular,

$$\forall n \in \mathbb{N}, \quad E_u(n) = A_n^2 + B_{n-1}^2.$$

First applying the same reasoning as in Lemma 5.2, we easily see that

$$\forall n \geq -1, \quad A_{n+2}^2 \leq B_n^2 \leq A_n^2 \leq B_{n-2}^2$$

and that the sequences $(A_n)_n$ and $(B_n)_n$ go to zero as $n \to +\infty$.

Next assume (2.7) or (2.11). We use (5.6) and Lemma 5.1 to get that

$$B_n^2 = A_n^2\Big(1 - 2^{p_1+1}|A_n|^{p_1-1} + \mathrm{o}\left(|A_n|^{p_1-1}\right)\Big). \tag{5.12}$$

Similarly, under (2.8) or (2.12), (5.7) and Lemma 5.1 give that

$$A_{n+2}^2 = B_n^2\Big(1 - 2^{p_2+1}|B_n|^{p_2-1} + \mathrm{o}\left(|B_n|^{p_2-1}\right)\Big). \tag{5.13}$$

Note that $|A_n| \underset{n\to+\infty}{\sim} |B_n|$ and $|A_{n+2}| \underset{n\to+\infty}{\sim} |A_n|$. We distinguish the cases $p_1 = p_2$ and $p_1 < p_2$. The strategy we used in Lemma 5.2 allows one to prove (2.9) and (2.10).

Idea of the proof of Proposition 2.2. We will use the spherical coordinates $r \in [1,2]$, $\phi \in [0, 2\pi]$, $\theta \in [0,\pi]$ and we will construct *radial* solutions satisfying (2.14). Let $A_0 \in \mathbb{R}$ be such that $A_0 \neq 0$ and define $\tilde{q}$ by $\tilde{q}(s) = 2q(s/2)$ $\forall s \in \mathbb{R}$. We introduce the real sequence $(A_n)_{n\in\mathbb{N}}$ defined by (5.1) in the proof of Theorem 2.1 with $\tilde{q}$ instead of q. Thus Lemma 5.2 is still true with a constant $\tilde{C}_p'$ instead of C_p'.

We define an absolutely continuous function $\Phi : (-2, +\infty) \to \mathbb{R}$ such that

$$\forall n \in \mathbb{N}, \quad \forall s \in (2n-2, 2n), \quad \Phi'(s) = A_n.$$

Then, u defined by

$$u(r,t) = \frac{1}{r}(-\Phi(t+r-2) + \Phi(t-r))$$

is the solution of problem (2.13) since

$$\partial_\nu u(2,t) + \frac{1}{2}u(2,t) = u_r(2,t) + \frac{1}{2}u(2,t) = -q(u_t(2,t)).$$

It is not difficult to verify that the energy of this radial solution defined as usual by the formula

$$E_u(t) = \frac{1}{2}\int_\Omega {u_t}^2 + |\nabla u|^2\,dx + \frac{1}{2}\int_{\partial B_2} \frac{1}{2}u^2\,d\sigma$$

satisfies

$$\begin{aligned}\forall t \geq 0, \quad E_u(t) &= \frac{1}{2}\int_1^2\int_0^{2\pi}\int_0^\pi \left(u_t^2 + \left(\frac{1}{r}\frac{d}{dr}(ru)\right)^2\right) r^2 \sin\theta\,d\theta\,d\phi\,dr \\ &= 2\pi\int_1^2 \left(\Phi'(t+r-2)^2 + \Phi'(t-r)^2\right)dr.\end{aligned}$$

Consequently,

$$\forall n \in \mathbb{N}, \ E_u(2n) = 4\pi A_n^2$$

and Proposition 2.2 follows from Lemma 5.2. □

5.2. Behavior of the energy in the general case.

Proof of Theorem 3.1. Let $A_0 \in \mathbb{R}$ be such that $A_0 \neq 0$. As for Theorem 2.1, $E_u(2n) = 2A_n^2$ $\forall n \in \mathbb{N}$, where $(A_n)_{n\in\mathbb{N}}$ is the real sequence defined by (5.1). Denote $u_n := |A_n|$. We know that the sequence $(u_n)_{n\in\mathbb{N}}$ is nonincreasing and that $u_n \underset{n\to+\infty}{\to} 0$.

First we will prove that in both cases $q = g$ or $q = g^{-1}$, there exists some $n_0 \in \mathbb{N}$ such that the corresponding sequences $(u_n)_{n\geq n_0}$ will satisfy exactly the same relation:

$$\forall n \geq n_0, \quad u_{n+1} - u_n = -g(u_{n+1} + u_n). \tag{5.14}$$

We denote by $[-s_0, s_0]$ a neighborhood of 0 where q is given by $q = g$ or by $q = g^{-1}$. Since $g'(0) = 0$, we can assume that $0 \leq g'(s) < 1/2$ $\forall s \in [-s_0, s_0]$. Then we introduce $n_0 \in \mathbb{N}$ such that, $\forall n \geq n_0$, $A_n \in [-\frac{s_0}{2}, \frac{s_0}{2}]$. Note that this implies in particular $|g(A_n)| < |A_n|$ $\forall n \geq n_0$.

Assume $q = g$. Then the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies

$$\forall n \geq n_0, \quad A_{n+1} + A_n = -g(A_{n+1} - A_n). \tag{5.15}$$

Fix $n \geq n_0$ and assume, for example, $A_n > 0$. We introduce the strictly increasing function $\phi : s \in [-\frac{s_0}{2}, \frac{s_0}{2}] \mapsto s + g(s - A_n)$. Then

$$\phi(0) = g(-A_n) > -A_n = \phi(A_{n+1}).$$

Therefore $A_{n+1} < 0$. So the sign of the sequence $(A_n)_{n\geq n_0}$ is alternating and (5.15) becomes, since g is odd,

$$\forall n \geq n_0, \quad |A_{n+1}| - |A_n| = -g(|A_{n+1}| + |A_n|),$$

which proves (5.14).

Assume now $q = g^{-1}$. Then the sequence $(A_n)_{n\in\mathbb{N}}$ satisfies

$$\forall n \geq n_0, \quad A_{n+1} + A_n = -g^{-1}(A_{n+1} - A_n),$$

or

$$\forall n \geq n_0, \quad A_{n+1} - A_n = -g(A_{n+1} + A_n). \tag{5.16}$$

Fix $n \geq n_0$ and assume, for example, $A_n > 0$. We introduce the strictly increasing function $\psi : s \in [-\frac{s_0}{2}, \frac{s_0}{2}] \mapsto s + g(s + A_n)$. Then

$$\psi(0) = g(A_n) < A_n = \psi(A_{n+1}).$$

Therefore $A_{n+1} > 0$. So the sign of the sequence $(A_n)_{n\geq n_0}$ is constant and (5.16) becomes

$$\forall n \geq n_0, \quad |A_{n+1}| - |A_n| = -g(|A_{n+1}| + |A_n|),$$

which proves (5.14).

Finally, in both cases ($q = g$ and $q = g^{-1}$), since $E_u(2n) = 2A_n^2 = 2u_n^2$, the problem reduces to studying the positive nonincreasing sequence $(u_n)_{n\geq n_0}$ defined by (5.14).

Let us introduce two other sequences $(v_n)_{n\geq n_0}$ and $(w_n)_{n\geq n_0}$ defined by $v_{n_0} = w_{n_0} = u_{n_0}$ and by

$$\forall n \geq n_0, \quad v_{n+1} = v_n - g(2v_n) = G(v_n), \tag{5.17}$$

and

$$\forall n \geq n_0, \quad w_n = w_{n+1} + g(2w_{n+1}) = H(w_{n+1}) \quad \text{or} \quad w_{n+1} = H^{-1}(w_n), \tag{5.18}$$

where $G : [0, \frac{s_0}{2}] \to \mathbb{R}$ and $H : [0, \frac{s_0}{2}] \to \mathbb{R}$ are the two strictly increasing functions defined by $G(s) = s - g(2s)$ and $H(s) = s + g(2s)$ $\forall s \in [0, \frac{s_0}{2}]$. We verify that $(v_n)_{n\geq n_0}$ and $(w_n)_{n\geq n_0}$ are positive nonincreasing sequences such that $v_n \to 0$ and $w_n \to 0$ as $n \to +\infty$. And, with these notations, we prove that

$$\forall n \geq n_0, \quad 0 \leq v_n \leq u_n \leq w_n. \tag{5.19}$$

Indeed, $v_{n_0} = w_{n_0} = u_{n_0}$. Assume for some $p > n_0$, $v_p \leq u_p \leq w_p$. Then

$$u_{p+1} = u_p - g(u_{p+1} + u_p) \geq u_p - g(2u_p) = G(u_p) \geq G(v_p) = v_{p+1},$$

and

$$u_p = u_{p+1} + g(u_{p+1} + u_p) \geq u_{p+1} + g(2u_{p+1}) = H(u_{p+1}).$$

Thus

$$u_{p+1} \leq H^{-1}(u_p) \leq H^{-1}(w_p) = w_{p+1}.$$

Let U be the solution of the ordinary differential equation

$$U'(s) = -g(2U(s)), \qquad s \geq 0, \tag{5.20}$$

such that $U(0) = u_{n_0} > 0$. The function $s \mapsto U(s)$ is strictly positive and decreasing and $U(s) \to 0$ as $s \to +\infty$.

We denote by $(s_n)_{n\geq n_0}$, $(t_n)_{n\geq n_0}$, and $(r_n)_{n\geq n_0}$ the increasing sequences defined by $s_n = U^{-1}(v_n)$, $t_n = U^{-1}(u_n)$, and $r_n = U^{-1}(w_n)$ $\forall n \geq n_0$. Since U^{-1} is decreasing, we have $s_n \geq t_n \geq r_n$. Note also, since v_n, u_n, $w_n \underset{n\to+\infty}{\to} 0$, that $U(s_n)$, $U'(s_n), U(t_n)$, $U'(t_n), U(r_n)$, $U'(r_n) \underset{n\to+\infty}{\to} 0$.

First we prove that $s_n \underset{n\to+\infty}{\sim} n$: the relation (5.17) becomes

$$U(s_{n+1}) = U(s_n) + U'(s_n).$$

Thus

$$\begin{aligned} s_{n+1} &= U^{-1}(U(s_n) + U'(s_n)) \\ &= U^{-1}(U(s_n)) + U'(s_n)(U^{-1})'(U(s_n)) + \frac{1}{2}U'(s_n)^2(U^{-1})''(U(s_n')) \\ &= s_n + 1 + \frac{1}{2}U'(s_n)^2 \left(\frac{-U''(s_n')}{U'(s_n')^3} \right) \\ &= s_n + 1 + \frac{1}{2}U'(s_n)^2 \frac{2g'(2U(s_n'))}{U'(s_n')^2} \\ &= s_n + 1 + g'(2U(s_n')) \frac{g(2U(s_n))^2}{g(2U(s_n'))^2}, \end{aligned}$$

with $s_n \leq s'_n \leq s_{n+1}$. Since U and g are, respectively, decreasing and increasing, we have

$$1 \leq \frac{g(2U(s_n))}{g(2U(s'_n))} \leq \frac{g(2U(s_n))}{g(2U(s_{n+1}))} = \frac{g(2v_n)}{g(2v_{n+1})}.$$

Moreover,

$$g(2v_{n+1}) = g(2v_n) + 2(v_{n+1} - v_n)g'(2\alpha_n) = g(2v_n) - 2g(2v_n)g'(2\alpha_n),$$

with $v_{n+1} \leq \alpha_n \leq v_n$. Thus, since $g'(0) = 0$, $g(2v_{n+1})/g(2v_n) \to 1$ as $n \to +\infty$. Thus

$$g'(2U(s'_n))\frac{g(2U(s_n))^2}{g(2U(s'_n))^2} \underset{n\to+\infty}{\longrightarrow} 0.$$

And we conclude that

$$s_{n+1} = s_n + 1 + o(1).$$

Therefore $s_n \underset{n\to+\infty}{\sim} n$.

In the same way we prove that $r_n \underset{n\to+\infty}{\sim} n$. Since $s_n \geq t_n \geq r_n$, we have $t_n \underset{n\to+\infty}{\sim} n$. We still introduce $V : \mathbb{R}_+ \to \mathbb{R}_+$ defined by $V(s) = 2U(s/2)$ $\forall s \geq 0$. Then $E_u(2n) = 2U(t_n)^2 = \frac{1}{2}V(2t_n)^2$, where $t_n \underset{n\to+\infty}{\sim} n$. □

5.3. Some explicit lower bounds of the energy.

Proof of Proposition 3.1. Let $A_0 \in \mathbb{R}$ be such that $A_0 \neq 0$. We use the notations of the proof of Theorem 3.1. Then $E_u(2n) = 2A_n^2$ $\forall n \in \mathbb{N}$, where $(A_n)_{n\in\mathbb{N}}$ is the real sequence defined by (5.1). As for Theorem 3.1, we can prove that, in both cases $|q| \leq |g|$ or $|q| \geq |g^{-1}|$ in a neighborhood of zero, there exists some $n_0 \in \mathbb{N}$ such that the sequence $(u_n)_{n\geq n_0} = (|A_n|)_{n\geq n_0}$ satisfies

$$(5.21) \qquad \forall n \geq n_0, \quad u_{n+1} - u_n \geq -g(u_{n+1} + u_n) \geq -g(2u_n).$$

We introduce $h = \frac{1}{2}g^{-1}$. Set $n_1 \in \mathbb{N}$, and let $(\lambda_n)_{n\geq n_0}$ be the real decreasing sequence (convergent to 0) defined by $h'(\lambda_n) = n + n_1$. Note that $\forall n \geq n_0$,

$$h(\lambda_n) = \frac{1}{2}(g')^{-1}\left(\frac{1}{2(n+n_1)}\right).$$

Choose $n_1 \in \mathbb{N}$ such that $h(\lambda_{n_0}) \leq u_{n_0}$. We prove by induction that

$$\forall n \geq n_0,\ u_n \geq h(\lambda_n).$$

Assume $u_n \geq h(\lambda_n)$. Then

$$u_{n+1} \geq u_n - g(2u_n) = G(u_n) \geq G(h(\lambda_n)) \geq h(\lambda_n) - \lambda_n.$$

On the other hand,

$$h(\lambda_n) - h(\lambda_{n+1}) = (\lambda_n - \lambda_{n+1})h'(\mu_n) \geq (\lambda_n - \lambda_{n+1})h'(\lambda_n)$$

(where $\lambda_{n+1} \leq \mu_n \leq \lambda_n$). Since

$$\lambda_n(h'(\lambda_n) - 1) \geq \lambda_{n+1}(h'(\lambda_{n+1}) - 1) = \lambda_{n+1}h'(\lambda_n),$$

we see that

$$(\lambda_n - \lambda_{n+1})h'(\lambda_n) \geq \lambda_n,$$

and so $u_{n+1} \geq h(\lambda_{n+1})$. □

Proof of Proposition 3.2. We still use the notations of the proof of Proposition 3.1 and we study the sequence $(u_n)_{n \geq n_0} = (|A_n|)_{n \geq n_0}$ that satisfies (5.21). Let $0 < \gamma \leq \beta$ be such that $u_{n_0} \geq h(\frac{\gamma}{n_0^k})$. We prove by induction that

$$u_n \geq h\left(\frac{\gamma}{n^k}\right).$$

Assume that $u_n \geq h(\frac{\gamma}{n^k})$. Then

$$u_{n+1} \geq u_n - g(2u_n) = G(u_n) \geq G\left(h\left(\frac{\gamma}{n^k}\right)\right) \geq h\left(\frac{\gamma}{n^k}\right) - \frac{\gamma}{n^k}.$$

On the other hand,

$$\begin{aligned} h\left(\frac{\gamma}{n^k}\right) - h\left(\frac{\gamma}{(n+1)^k}\right) &\geq \left(\frac{\gamma}{n^k} - \frac{\gamma}{(n+1)^k}\right) h'\left(\frac{\gamma}{n^k}\right) \\ &\geq \gamma(n+1)\left(\frac{1}{n^k} - \frac{1}{(n+1)^k}\right) \geq \frac{\gamma}{n^k}. \end{aligned}$$

Thus

$$u_{n+1} \geq h\left(\frac{\gamma}{(n+1)^k}\right). \qquad \square$$

Lower bound for the energy of a more general class of initial conditions. The reasoning is exactly the same: set $(u^0, v^0) \in W^{1,\infty}(0,1) \times L^\infty(0,1)$ such that $\|(u^0, v^0)\|_{W^{1,\infty} \times L^\infty}$ is small enough. The solution of (1.1) can be written as

$$u(x,t) = f(t+x) - f(t-x),$$

and the boundary condition implies that the sequence $(A_n(s) = f'(s+2n))_{n \in \mathbb{N}}$ for $s \in (-1,1)$ satisfies

$$A_{n+1}(s) + A_n(s) + q(A_{n+1}(s) - A_n(s)) = 0. \tag{5.22}$$

Note that $(u^0, v^0) \in W^{1,\infty}(0,1) \times L^\infty(0,1)$ if and only if $A_0 \in L^\infty(0,1)$, and $(u^0, v^0) \in V \times L^2(0,1)$ if and only if $A_0 \in L^2(0,1)$. Similarly

$$E_u(2n) = \int_{2n-1}^{2n+1} f'(s)^2 \, ds = \int_{-1}^{1} A_n(s)^2 \, ds.$$

Then the reasoning applied to prove Proposition 3.2 shows that under the assumption (3.10) we can find $\gamma(s) \in [0, \beta]$ such that

$$|A_n(s)| = u_n(s) \geq h\left(\frac{\gamma(s)}{n^k}\right).$$

Since (u^0, v^0) is not equal to zero, the function γ is bounded from below by some positive constant γ_0 on some subinterval J of $(-1,1)$. Then

$$E_u(2n) \geq |J| \, h\left(\frac{\gamma_0}{n^k}\right)^2,$$

where $|J|$ denotes the length of J. □

5.4. Optimality of estimate (3.1) for a class of nonpolynomial feedbacks.

Proof of Proposition 3.3. We still use the notations of the proof of Proposition 3.1 and we study the sequence $(u_n)_{n\geq n_0} = (|A_n|)_{n\geq n_0}$ that satisfies (5.21). There exists $r_n \in [u_{n+1}, u_n]$ such that

$$\begin{aligned} g(\alpha u_{n+1}) &= g(\alpha u_n) + \alpha(u_{n+1} - u_n) g'(\alpha r_n) \\ &\qquad\geq g(\alpha u_n) - \alpha g(2u_n) g'(\alpha r_n) \geq g(\alpha u_n) - \alpha g(2u_n) g'(\alpha u_n). \end{aligned}$$

Note that (3.12) gives that $g(\alpha u_n) - \alpha g(2u_n) g'(\alpha u_n) > 0$. We deduce that

$$\begin{aligned} \frac{1}{g(\alpha u_{n+1})} - \frac{1}{g(\alpha u_n)} &\leq \frac{1}{g(\alpha u_n)} \left(\frac{1}{1 - \frac{\alpha g(2u_n) g'(\alpha u_n)}{g(\alpha u_n)}} - 1 \right) \\ &\leq \frac{1}{g(\alpha u_n)} \left(1 + 2\alpha \frac{g(2u_n) g'(\alpha u_n)}{g(\alpha u_n)} - 1 \right) \\ &\leq 2\alpha \frac{g(2u_n) g'(\alpha u_n)}{g(\alpha u_n)^2} \leq M. \end{aligned}$$

Thus

$$\frac{1}{g(\alpha u_n)} \leq Mn + C,$$

where $C > 0$ is a constant, which proves the result. □

5.5. Remark concerning (2.2). Let v be the solution of (3.15). Note that

$$v(n+1) - v(n) = v'(\alpha_n) = -q(v(\alpha_n)),$$

with $\alpha_n \in [n, n+1]$. Since q is increasing and v nonincreasing, we have

$$-q(v(n)) \leq v(n+1) - v(n) \leq -q(v(n+1)).$$

Thus, under the hypotheses of Theorem 3.1 and Propositions 3.1 and 3.2 and with similar proofs, we can obtain similar lower bounds of the energy.

5.6. Decay rate when the feedback is weak at infinity. Consider now the special function q defined by

$$\begin{cases} \forall |s| \leq 2, \ q(s) = \frac{s}{2}, \\ \forall |s| \geq 2, \ q(s) = \text{ sgn } (s) 1. \end{cases}$$

We solve (5.22) to deduce that

$$\begin{cases} \text{if } |A_n(s)| \geq \frac{3}{2}, \text{ then } |A_{n+1}(s)| = |A_n(s)| - 1, \\ \text{if } |A_n(s)| \leq \frac{3}{2}, \text{ then } |A_{n+1}(s)| = k|A_n(s)| \text{ with } k := \frac{1}{3}. \end{cases}$$

Set $s \in (0, 1)$ and let $p(s)$ be the smallest nonnegative integer such that

$$|A_0(s)| \leq p(s) + \frac{3}{2}.$$

We deduce easily that

$$\left\{\begin{array}{l} \text{if } q \leq p(s), \text{ then } |A_q(s)| = |A_0(s)| - q, \\ \text{if } q \geq p(s), \text{ then } |A_q(s)| = k^{q-p(s)}|A_{p(s)}(s)|. \end{array}\right. \tag{5.23}$$

Exponential decay of strong solutions: set $(u^0, v^0) \in W^{1,\infty}(0,1) \times L^\infty(0,1)$, then $A_0 \in L^\infty(-1,1)$, and so the energy of the solution u decays exponentially to zero. Indeed, there exists p_0 such that

$$\|A_0\|_\infty \leq p_0 + \frac{3}{2};$$

then $|A_{p_0}(s)| \leq \frac{3}{2}$ and

$$\forall n \geq p_0,\ E_u(2n) = \int_{-1}^{1} A_n(s)^2\, ds \leq \frac{9}{2} k^{2(n-p_0)}. \qquad \square$$

The strong solutions do not decay uniformly exponentially to zero: consider the sequence of initial conditions

$$\left\{\begin{array}{l} A_0^{(m)}(s) = m \text{ if } s \in (-\frac{1}{2m^2}, \frac{1}{2m^2}), \\ A_0^{(m)}(s) = 0 \text{ if } s \in (-1,1) \setminus (-\frac{1}{2m^2}, \frac{1}{2m^2}). \end{array}\right.$$

All these initial conditions satisfy $E^{(m)}(0) = 1$. We easily deduce from (5.23) that

$$\forall r \in \mathbb{N},\ E^{(m)}(2m + 2r) = \frac{1}{9m^2}\frac{1}{9^r}.$$

Thus we cannot expect that there exist $C > 0$ and $\omega > 0$ that do not depend on m such that all strong solutions satisfy

$$\forall t \geq 0,\ E(t) \leq CE(0)e^{-\omega t}.$$

Indeed, that would mean

$$\forall m \in \mathbb{N},\ \frac{1}{9m^2} = E^{(m)}(2m) \leq Ce^{-2\omega m}. \qquad \square$$

The study of the decrease of the energy of weak solutions: first we prove that weak solutions do not decay uniformly polynomially to zero. Fix $\varepsilon \in (0,1)$ and consider the following initial conditions:

$$\left\{\begin{array}{l} \forall s \in (0,1),\ A_0(s) = \frac{1}{s^{(1-\varepsilon)/2}}, \\ \forall s \in (-1,0),\ A_0(s) = 0. \end{array}\right.$$

Then we deduce from (5.23) that

$$\text{if } A_0(s) \geq n + \frac{1}{2}, \text{ then } A_n(s) = A_0(s) - n.$$

Set $a := \frac{1-\varepsilon}{2}$. Thus

$$E_u(2n) = \int_0^1 A_n(s)^2\, ds \geq \int_0^{(n+1)^{-1/a}} A_n(s)^2\, ds = \int_0^{(n+1)^{-1/a}} \left(\frac{1}{s^a} - n\right)^2 ds.$$

Easy computations lead to

$$\int_0^{(n+1)^{-1/a}} \left(\frac{1}{s^a} - n\right)^2 ds \underset{n\to+\infty}{\sim} \frac{2a^2}{(1-2a)(1-a)} \frac{1}{n^{(1-2a)/a}}.$$

Since $\frac{1-2a}{a} = \frac{2\varepsilon}{1-\varepsilon}$ can be chosen as close to zero as we want, the energy of weak solutions does not decay uniformly polynomially to zero.

We need only to refine this study to prove (4.4): fix $p \geq 2$ and consider

$$\forall s \in (0, T_p^{-1}),\ A_0(s) = \left(\frac{1}{s \ln_1(s^{-1}) \ln_2(s^{-1}) \cdots \ln_{p-1}(s^{-1})(\ln_p(s^{-1}))^2}\right)^{1/2},$$

and $A_0(s) = 0$ in $(-1,0) \cup (T_p^{-1}, 1)$. We verify that $A_0 \in L^2(0,1)$: we use several times the change of variables $z = \ln s$ to get

$$\begin{aligned}\int_{-1}^1 A_0(s)^2\, ds &= \int_0^{T_p^{-1}} \frac{ds}{s \ln_1(s^{-1}) \ln_2(s^{-1}) \cdots \ln_{p-1}(s^{-1})(\ln_p(s^{-1}))^2} \\ &= \int_{\ln T_p}^{+\infty} \frac{dz}{z \ln_1(z) \ln_2(z) \cdots \ln_{p-2}(z)(\ln_{p-1}(z))^2} \\ &= \cdots = \int_{\ln_p(T_p)}^{+\infty} \frac{dz}{z^2} = \frac{1}{\ln_p(T_p)}.\end{aligned}$$

Next we use the same strategy: set $\alpha > 0$ and define for n large enough

$$s_n = \frac{\alpha}{n^2 \ln_1(n) \ln_2(n) \cdots \ln_{p-1}(n)(\ln_p(n))^2}.$$

We easily see that

$$A_0(s_n)^2 \underset{n\to+\infty}{\sim} \frac{n^2}{2\alpha}.$$

Therefore, if we choose, for example, $\alpha := \frac{1}{4}$, for n large enough we have

$$A_0(s_n) \geq n + \frac{1}{2}.$$

Therefore

$$\begin{aligned}E_u(2n) = \int_{-1}^1 A_n(s)^2\, ds \geq \int_0^{s_n} A_n(s)^2\, ds &= \int_0^{s_n} (A_0(s) - n)^2\, ds \\ &\geq \frac{1}{2} \int_0^{s_n} A_0(s)^2\, ds - n^2 s_n \\ &\geq \frac{1}{2} \frac{1}{\ln_p(s_n^{-1})} - n^2 s_n \underset{n\to+\infty}{\sim} \frac{1}{2} \frac{1}{\ln_p(s_n^{-1})}.\end{aligned}$$

Thus, we obtain that for n large enough,

$$E_u(2n) \geq \frac{1}{8} \frac{1}{\ln_p(n)} \geq \frac{1}{\ln_{p-1}(n)}. \qquad \square$$

Acknowledgments. The authors are grateful to F. Conrad, A. Haraux, and V. Komornik for their questions and comments.

REFERENCES

[1] M. Aassila, *On a quasilinear wave equation with a strong damping*, Funkcial Ekvac., 41 (1998), pp. 67–78.

[2] A. Carpio, *Sharp estimates of the energy for the solutions of some dissipative second order evolution equations*, Potential Anal., 1 (1992), pp. 265–289.

[3] P. Cannarsa, V. Komornik, and P. Loreti, *Well posedness and control of semilinear wave equations with iterated logarithms*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 37–56.

[4] F. Conrad, J. Leblond, and J. P. Marmorat, *Stabilization of second order evolution equations by unbounded nonlinear feedbacks*, in Proceedings of the Fifth International Federation of Automatic Control Symposium on Control of Distributed Parameter Systems, Perpignan, 1989, pp. 101–116.

[5] F. Conrad and B. Rao, *Decay of solutions of the wave equation in a star-shaped domain with nonlinear boundary feedback*, Asymptot. Anal., 7 (1993), pp. 159–177.

[6] A. Haraux, *Comportement à l'infini pour une équation des ondes non linéaire dissipative*, C. R. Acad. Sci. Paris Sér. A, 287 (1978), pp. 507–509.

[7] A. Haraux, *Systèmes dynamiques dissipatifs et applications*, Masson, Paris, 1991.

[8] A. Haraux, *L^p Estimates of Solutions to Some Nonlinear Wave Equations in One Space Dimension*, Publications du laboratoire d'analyse numérique, Université Pierre et Marie Curie, CNRS, Paris, 1995.

[9] A. Haraux and E. Zuazua, *Decay estimates for some semilinear damped hyperbolic problems*, Arch. Rational Mech. Anal., 100 (1988), pp. 191–206.

[10] V. Komornik, *On the nonlinear boundary stabilization of the wave equation*, Chinese Ann. Math., 14 (1993), pp. 153–164.

[11] V. Komornik, *Decay estimates for the wave equation with internal damping*, in Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena, Internat. Ser. Numer. Math. 118, Birkhäuser, Basel, 1994, pp. 253–266.

[12] S. Kouémou-Patcheu, *On the decay of solutions of some semilinear hyperbolic problems*, Panamer. Math. J., 6 (1996), pp. 69–82.

[13] I. Lasiecka I. and D. Tataru, *Uniform Boundary Stabilization of Semilinear Wave Equation with Nonlinear Boundary Damping*, Lectures Notes in Pure and Appl. Math. 142, Dekker, New York, 1993.

[14] W.-J. Liu and E. Zuazua, *Decay rates for dissipative wave equation*, Ric. Mat., 48 (1999), pp. 61–75.

[15] P. Martinez, *A new method to obtain decay rate estimates for dissipative systems*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 419–444.

[16] P. Martinez and J. Vancostenoble, *Exponential stability for the wave equation with weak nonmonotone damping*, Portugal. Math., to appear.

[17] M. Nakao, *Asymptotic stability of the bounded or almost periodic solution of the wave equation with a nonlinear dissipative term*, J. Math. Anal. Appl., 58 (1977), pp. 336–343.

[18] M. Nakao, *Energy decay for the wave equation with a nonlinear weak dissipation*, Differential Integral Equations, 8 (1995), pp. 681–688.

[19] P. Souplet, *Propriétés globales de quelques équations d'évolution non linéaires du second ordre*, Thèse de doctorat de l'Université de Paris VI, 1994.

[20] L. R. Tcheugoué Tébou, *Stabilization of the wave equation with localized nonlinear damping*, J. Differential Equations, 145 (1998), pp. 502–524.

[21] J. Vancostenoble, *Optimalité d'estimations d'énergie pour une équation des ondes amortie*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 777–782.

[22] H. K. Wang and G. Chen, *Asymptotic behavior of solutions of the one-dimensional wave equation with a nonlinear boundary stabilizer*, SIAM J. Control Optim., 27 (1989), pp. 758–775.

[23] E. Zuazua, *Uniform stabilization of the wave equation by nonlinear boundary feedback*, SIAM J. Control Optim., 28 (1990), pp. 466–477.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 798–811

SUB-FINSLERIAN METRIC ASSOCIATED TO AN OPTIMAL CONTROL SYSTEM*

C. LÓPEZ† AND E. MARTÍNEZ†

Abstract. The problem of minimizing the cost functional of an optimal control system through the use of constrained variational calculus is a generalization of the geodetic problem in Riemannian geometry. In the framework of a geometric formulation of optimal control, we define a metric structure associated to the optimal control system on the enlarged space of state and time variables, such that the minimal length curves of the metric are the optimal solutions of the system. A twofold generalization of metric structure is applied, considering Finslerian-type metrics as well as allowed and forbidden directions (like in sub-Riemannian geometry). Free (null Hamiltonian) or fixed final parameter problems are identified with constant *energy* leaves, and the restriction of the metric to these leaves gives way to a family of metric structures on the usual state manifold.

Key words. optimal control for ordinary differential equations, Finsler metrics

AMS subject classifications. 49J15, 53B40

PII. S0363012999357562

1. Introduction.

A positive definite symmetric tensor g on a manifold M is named a Riemannian metric structure (M, g). It allows us to measure lengths of curves by first calculating the norm of the velocity vector associated to a parametrized curve, and then integrating it along the curve

$$\ell(\gamma) = \int_{t_0}^{t_1} \sqrt{g(\gamma^c, \gamma^c)}dt,$$

where $\gamma^c(t)$ is the lifting of the curve $\gamma\colon [t_0, t_1] \to M$ to TM, the tangent bundle. In local coordinates, if $\gamma(t) = (x^i(t))$, and g is given by $g_{ij}(x)dx^i dx^j$, the norm of the velocity vector is $|\gamma^c(t)| = \sqrt{g_{ij}(x(t))\frac{dx^i}{dt}\frac{dx^j}{dt}}$.

Length has the property of being invariant under reparametrization of the curve, because the norm is a homogeneous positive function on the velocities. Curves of minimal length joining two points are said to be geodesics, and, through classical variational calculus, a set of necessary conditions is obtained for these curves, the geodetic equations or Euler–Lagrange equations for the Lagrangian $L = \sqrt{g(v, v)}$. But (M, g) has much more structural content: associated connection, scalar product of vectors, and the whole exhaustively studied machinery of Riemannian geometry.

If we are just interested in measuring the length of curves, we only need a homogeneous positive function F in TM (not necessarily the square root of a quadratic function) to be integrated along the curves, giving an invariant under reparametrization length. This defines a Finslerian metric structure on (M, F), which was, in fact, the original theme of work in the early study by Riemann (see [5] and references therein), and many properties of Riemannian geometry can be generalized to this

*Received by the editors June 2, 1999; accepted for publication (in revised form) June 9, 2000; published electronically October 4, 2000. This research was partially supported by DGES PB96-0717.

http://www.siam.org/journals/sicon/39-3/35756.html

†Departamento de Matemática Aplicada, C.P.S. Universidad de Zaragoza, María de Luna 3, E-50015 Zaragoza, Spain, (carlopez@posta.unizar.es, emf@posta.unizar.es).

case. In this paper, the extra conditions of convexity, $\frac{\partial^2 F}{\partial v \partial v} > 0$, and central symmetry $F(-v) = F(v)$ (see, for example, [6]), will not be imposed, so that our Finslerian structure will be more general than usual.

Another simple generalization of the Riemannian metric structure is given by constraining the set of allowed directions for the curves to be measured, i.e., by considering a cone of directions $D \subset TM$ or, equivalently, a subset of the positive projective space $SM \equiv TM/\Delta$, where Δ is the Liouville or dilation vector field. Only curves whose lifting to TM lies in D can be measured, and we have a Riemannian metric g defined on D. This is the case of sub-Riemannian geometry [2, 9], where a linear distribution E (usually completely nonintegrable) is considered as the set of allowed directions on M. (M, E, g) is a sub-Riemannian structure, and the length of curves γ whose lifting lies in E is given by the usual integral functional. The unexpected existence of abnormal geodesics has been clarified by using the optimal control machinery, identifying those abnormal curves with the solutions of the problem with $p_0 = 0$. Optimal control, as a generalization of classical variational calculus, is therefore an appropriate framework for the study of metric structures.

The joining of both ideas, that of Finslerian functions and that of a constrained directional set, would give way to sub-Finslerian metric structures (M, D, F), where D is a cone of allowed directions and F is a homogeneous positive function on D. Of course, in this generalization process we have lost properties, but the basic one of measuring lengths of curves is preserved. The visualization of the indicatrix (set of vectors with norm 1) helps to understand the differences between metrics; the indicatrix is a hyperellipsoid centered at the origin for Riemannian metrics, a general convex hypersurface around the origin for Finslerian metrics, and a very general set (neither convex nor intersecting all directional rays) for a more general sub-Finslerian metric. Our aim in this paper is to point out that optimal control systems can be understood as metric structures of this generalized sub-Finslerian type, and to show how typical properties of optimal control, as bang-bang phenomena or allowed but never optimal directions, can be illustrated through the indicatrix of the associated metric structure.

A typical system where allowed velocities are constrained is a control system of ordinary differential equations. A control system is defined through a set of differential equations (usually in normal form) $\frac{dx^i}{dt} = f^i(x, u)$, where the u^a variables are named controls. In geometric terms [11], it corresponds to a bundle map $X : U \to TM$ over the identity in the state manifold M from the control bundle (U, η, M) to the tangent bundle (TM, τ, M), where the image of the map X is the set of allowed velocities. If we want to minimize a cost functional $C(\gamma) = \int f^0(x(t), u(t))dt$ for allowed curves of the control system joining two points, we face an optimal control problem [8, 12]. In other words, we are given a way to measure the cost for allowed curves. It is like a generalized metric structure as presented before, but we do not have the invariance under reparametrization property because the cost functional is not obtained from a Finslerian function.

Using the machinery of optimal control (the Pontryagin maximum principle [12]), we can associate to the optimal control problem a sub-Finslerian structure, i.e., a cone of directions D and a Finslerian function F, such that the cost value of a curve solution of the control system is just the length of the curve (or of any reparametrization of the curve). Therefore, the geodetic curves of the metric system are the curves making minimal the cost, the solutions of the optimal control problem. This metric structure is defined in an enlarged space of states and time (parameter of evolution) $N = M \times \mathbb{R}$

for the general case; for free or fixed final parameter problems, we obtain through restriction and projection a family of metric structures on M, parametrized by the energy, in classical mechanical language.

The paper is organized as follows. Section 2 is devoted to present a geometric description of optimal control systems based on the geometry of control systems as presented in [1, 11], and a geometric version of Pontryagin maximum principle, which gives way to a Dirac type constraint algorithm [7]. In section 3 the metric structure associated to a time optimal problem (the simplest case) is defined on M (the state space), and academic examples are presented where some graphic representation of the indicatrix helps to understand the basic ideas. Section 4 considers the general optimal control problem, defining a metric structure on the manifold $N = M \times \mathbb{R}$ of states and time; the generalization is based on a simple reparametrization of the system which transforms it into a time optimal problem. As an illustrative example, the Finslerian metric associated to a mechanical Lagrangian system is presented. The restriction of the metric to certain energy leaves, associated to free or fixed final parameter problems, is developed in section 5. In the former example of Lagrangian mechanics, this restriction gives way to the well-known Jacobi metrics. A final section with conclusions and possible applications is also included.

2. Geometry of optimal control. A (continuous) control system of (autonomous) ordinary differential equations is a family of differential equations in normal form $\frac{dx^i}{dt} = f^i(x,u)$, where x^i are named state variables, t is the parameter of evolution (usually the time), and u^a are the controls. From the geometric point of view [11], it can be understood as a fibered mapping $X\colon U \to TM$ from a control fiber bundle (U, η, M) over the state manifold M to the tangent bundle (TM, τ, M). Using local coordinates $(x^i)_{i=1}^n$ in M, adapted coordinates $(x^i, u^a)_{a=1}^k$ in U, and natural coordinates (x^i, v^i) in TM, the coordinate expression for X is $X(x,u) = f^i(x,u)\frac{\partial}{\partial x^i}$, or $v^i = f^i(x,u)$, the family of control equations. Allowed curves of the control system are curves $\rho\colon I \subset \mathbb{R} \to U$ such that $(\eta \circ \rho)^c = X \circ \rho$, where γ^c is the natural lifting to TM of a curve γ in M. In local coordinates, $\rho(t) = (x^i(t), u^a(t))$ is a solution integral curve of the control system if $\frac{dx^i}{dt} = f^i(x(t), u(t))$, i.e., if $(\eta \circ \rho)^c = (x^i(t), \frac{dx^i}{dt})$ equals $X \circ \rho = (x^i(t), f^i(x(t), u(t)))$. Note that the evolution on the state manifold M is totally characterized by the image set $S = \mathrm{Im}(X) \subset TM$, while the map $X_S\colon U \to S$ can be understood as a parametrization (perhaps redundant) of S. A particular problem of the control system can always be solved in two steps; the first solves the problem in the *basic* control system $i\colon S \to TM$, obtained from the natural injection of S into TM, and the second determines a particular inverse of X over the solution. In the usual control system language, we are just selecting the *essential* controls (all of them whenever X_S is a diffeomorphism [14]). Therefore, we sometimes will understand (x^i, u^a) as a nonnatural system of coordinates in the allowed velocities space $S \subset TM$, and we will identify the control set $U_x = \eta^{-1}(x)$ for a fixed $x \in M$ with the set of allowed velocities $S_x = S \cap \tau^{-1}(x)$ on x.

In optimal control theory [8, 12], a cost functional $C(\rho) = \int f^0(x(t), u(t))dt$ is given, and the problem is to obtain allowed curves of the control system satisfying some boundary conditions (e.g., $x(0) = x_0$, $x(T) = x_1$) and minimizing the cost functional. It is therefore a classical variational problem with nonintegrable constraints defined by the control equations. Pontryagin's maximum principle [12] gives a set of necessary conditions for a curve $(x(t), u(t))$ to be optimal; introducing a Hamiltonian function

$$H(x, p_0, p, u) = p_0 f^0(x, u) + p_i f^i(x, u), \tag{2.1}$$

where the variables (p_0, p_i) are momenta coordinates, the optimal curves $(x(t), u^*(t))$ must satisfy the control system equations

$$\frac{dx^i}{dt} = \frac{\partial H}{\partial p_i} = f^i(x(t), u^*(t)), \tag{2.2}$$

and there must exist a solution curve for the adjoint differential equations

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial x^i} = -p_0 \frac{\partial f^0}{\partial x^i}\Big|_{(x(t),u^*(t))} - p_j \frac{\partial f^j}{\partial x^i}\Big|_{(x(t),u^*(t))} \tag{2.3}$$

with the optimal control u^* satisfying the algebraic condition of maximality

$$H(x, p, u^*) \geq H(x, p, u) \qquad \forall u \in U_x = \eta^{-1}(x). \tag{2.4}$$

Here p_0 is a nonpositive constant, the case $p_0 = 0$ is named abnormal, and for $p_0 < 0$ it can be fixed to -1 by homogeneity of the adjoint equations and the Hamiltonian. Moreover, the Hamiltonian vanishes when the final parameter T is not fixed.

This set of Hamiltonian differential equations and the algebraic condition of maximality are the cornerstone of optimal control theory. The so-called transversality conditions on the momenta must be added to the boundary conditions when the initial and final endpoints are not fixed but are restricted to belong to some subsets of the state space. When the set of controls U is a manifold with boundary, the algebraic condition of maximality can be achieved in an interior point (and a weaker algebraic equation could be used, $\frac{\partial H}{\partial u^a} = 0$ for stationary points) or in the boundary. Sometimes the values u^{*a} can be determined explicitly as $u^{*a}(x, p)$, which can be understood as a dynamical feedback, and we reduce the problem to a set of $2n$ differential equations with boundary conditions, but this is not the case in many complex situations. It is interesting to observe that the former Hamiltonian is of a mixed type, with explicit dependence on velocity (through the control variables) as well as momenta coordinates. This is similar to the approach used in [13] to study the relationship between Lagrangian and Hamiltonian formulations of classical mechanics: given a Lagrangian function $L(x, v)$ on the tangent bundle, a Hamiltonian function $H(x, v, p) = \langle p, v\rangle - L(x, v)$ (where $\langle p, v\rangle$ is the natural pairing between covectors and vectors) is defined on the Whitney sum of tangent and cotangent bundles $W_M = TM \times_M T^*M$. The Hamiltonian equations (obtained through a natural presymplectic form on W)

$$\frac{dx}{dt} = \frac{\partial H}{\partial p} = v, \qquad \frac{dp}{dt} = -\frac{\partial H}{\partial x} = \frac{\partial L}{\partial x}$$

and algebraic condition

$$\frac{\partial H}{\partial v} = 0 = p - \frac{\partial L}{\partial v}$$

are equivalent to the Euler–Lagrange equations for the Lagrangian L. Only for regular Lagrangians, when the velocities $v(x, p)$ can be determined uniquely through the algebraic condition $p - \frac{\partial L}{\partial v} = 0$, the usual Hamiltonian $H_0(x, p) = \langle p, v(x, p)\rangle - L(x, v(x, p))$

living on T^*M gives way to an equivalent system of differential equations, but this is not the case for singular Lagrangians.

Therefore, a geometric transcription of Pontryagin's maximum principle is naturally developed in the framework of the Whitney sum $W = U \times_M T^*M$ (or the subset $S \times_M T^*M$ of W_M for basic systems) for the normal case $p_0 = -1$, the one presented here for simplicity. The Hamiltonian function $H(x, p, u) = \langle p, X(x, u)\rangle - f^0(x, u)$ is defined on W, and the pull-back of the canonical symplectic form $\omega = dx^i \wedge dp_i$ on T^*M to W, $\Omega = \mathrm{pr}_2^*(\omega)$, with $\mathrm{pr}_2\colon W \to T^*M$ the natural projection, determines a presymplectic Hamiltonian structure (W, Ω, H). The algebraic condition of maximality defines a subset $W_1 \subset W$ by taking on every fiber of pr_2 the point (or points) where H is maximal

$$W_1 = \{(x, p, u^*) \in W | H(x, p, u^*) \geq H(x, p, u) \quad \forall (x, p, u) \in \mathrm{pr}_2^{-1}(x, p), (x, p) \in T^*M\}.$$

The optimal vector field Γ solution of the optimal control system is determined by

$$i(Z)(i(\Gamma)\Omega - dH)\Big|_{W_1} \geq 0$$

for every arbitrary *allowed* vector Z in $T_{W_1}W$, i.e., an arbitrary vector in an interior point (in this case the condition is simply $(i(\Gamma)\Omega - dH)|_{W_1} = 0$) or a vector tangent to the boundary or pointing to the interior of W for a point in the boundary of W. The optimal curves are integral curves of Γ. In local coordinates, the points (x, p, u^*) of W_1 are just those satisfying the algebraic condition of maximality (2.4); given $\Gamma = a^i \frac{\partial}{\partial x^i} + b_i \frac{\partial}{\partial p_i} + c^a \frac{\partial}{\partial u^a}$, we get

$$i(\Gamma)\Omega - dH = \left(-b_i - \frac{\partial H}{\partial x^i}\right) dx^i + \left(a^i - \frac{\partial H}{\partial p_i}\right) dp_i - \frac{\partial H}{\partial u^a} du^a.$$

Therefore, $i(Z)(i(\Gamma)\Omega - dH)|_{W_1} \geq 0$ gives way to

$$a^i = \frac{\partial H}{\partial p_i}, \quad b_i = -\frac{\partial H}{\partial x^i},$$

and to the condition of local maximum in the fiber for H, which is certainly fulfilled in the global maximum point of W_1. A similar geometric description can be obtained to include the abnormal case $p_0 = 0$ and to formulate nonautonomous systems, by taking into account the cost and/or the time variables in an enlarged state manifold $M \times \mathbb{R}$ or $M \times \mathbb{R}^2$, with extra coordinates x^0 (the cost variable) and $x^{n+1} \equiv t$, and extra control equations

$$\frac{dx^0}{dt} = f^0(x, u), \qquad \frac{dx^{n+1}}{dt} = 1.$$

We present in the next section the simpler case of autonomous time optimal problems, where $f^0(x, u) = 1$, and those extra variables are not necessary to define the associated metric.

Notice that the solution Γ is not necessarily unique (the c^a components can be underdetermined) and it should be tangent to W_1. This compatibility condition is solved by applying a constraint algorithm similar to the one developed in [7] for presymplectic Hamiltonian systems, defining a chain of constraint submanifolds. In many cases the algorithm is trivial and W_1 is the final constraint submanifold, but sometimes the compatibility condition can be nontrivial, as it is in the case of sub-Riemannian geometry for the abnormal solutions.

3. Time optimal problems and the associated metric. For a time optimal problem we try to minimize the final time T (we take $t = 0$ as initial value), so that $f^0(x,u) = 1$, and the Hamiltonian (for the normal case) becomes

$$H(x,p,u) = p_i f^i(x,u) - 1 = \langle p, X(x,u)\rangle - 1. \tag{3.1}$$

The final parameter is obviously free (it is the unknown cost), and therefore the optimal curves take values on the subset $H = 0$ of W. In order to define a sub-Finslerian metric structure on M we need a set of allowed directions, a cone bundle $D \subset TM$, and a Finslerian (homogeneous) function defined on D. The construction is made pointwise, i.e., fixing an $x \in M$ and defining $D_x = D \cap \tau^{-1}(x)$ and $F_x\colon D_x \to \mathbb{R}$.

Once fixed x, the algebraic condition of maximality (2.4) determines a subset U_x^* of $U_x = \eta^{-1}(x)$ of possible optimal controls

$$U_x^* = \{(x,u_o) \in U_x \,|\, \exists (x,p) \in T_x^*M \quad \text{such that} \quad H(x,p,u_o) \geq H(x,p,u) \quad \forall u \in U_x\},$$

from which the actual optimal control $u^*(x,p)$ is determined for every p. Note that $U^* = \cup_{x\in M} U_x^*$ is just the projection of W_1 to U. The image subset $S_x^* = X(U_x^*) \subset S_x$ is made of some of the longest allowed velocity vectors, because, for a given momentum covector $p_i dx^i$, the Hamiltonian is maximal for the vector $f^i(x,u)\frac{\partial}{\partial x^i}$ with the greatest projection, $\max\langle p, f(x,u)\rangle$, i.e., the longest in some particular ray direction. (Note that for the actual solutions of the adjoint equations the projection is always positive by the nullity of the Hamiltonian, $\langle p, f(x,u)\rangle = 1$, and therefore the shortest vector in a given direction is never optimal.) So we have

$$S_x^* \subset S_x^o = \{(x,v) \in T_xM | v \in S_x \quad \text{and} \quad \lambda v \notin S_x \quad \forall \lambda > 1\}. \tag{3.2}$$

Now, the invariance under reparametrization needed to define a metric can be obtained by considering the cone D_x of rays generated by elements of S_x^o (the curves can be followed at arbitrary positive speed)

$$D_x = \{(x,v) \,|\, \exists (x,v_o) \in S_x^o \quad \text{such that} \quad v = \lambda v_o, \quad \lambda > 0\}. \tag{3.3}$$

The *norm* of the velocity is now defined to obtain a length equal to the original cost; we associate to every $(x,v) \in D_x$ the norm λ, which is the factor between it and the element of S_x^o in the same ray

$$F_x\colon D_x \to \mathbb{R}^+, \qquad F_x(v) = \lambda, \quad \text{where } v = \lambda v_o, \text{ with } v_o \in S_x^o. \tag{3.4}$$

An optimal curve $\rho(t) = (x(t), u^*(t))$, with cost $\int_0^T dt = T$, will have, once arbitrarily reparametrized by $t(\tau)$ ($\frac{dt}{d\tau} > 0$), the length

$$\int_{\tau_i}^{\tau_f} F(x(\tau), v(\tau)) d\tau = \int_{\tau_i}^{\tau_f} \lambda(\tau) d\tau = \int_{\tau_i}^{\tau_f} \frac{dt}{d\tau} d\tau = \int_0^T dt = T,$$

because

$$v(\tau) = \frac{dx}{d\tau} = \frac{dx}{dt}\frac{dt}{d\tau} = v_o \frac{dt}{d\tau}$$

so that $F(x(\tau), v(\tau)) = \lambda(\tau) = \frac{dt}{d\tau}$. Note that $S_x^o \subset D_x$ is the indicatrix of the defined metric, the set of unit norm velocities.

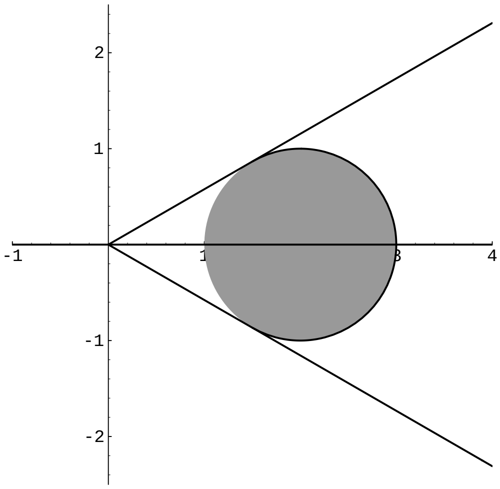


FIG. 1. *Allowed velocities, optimal velocities, and the cone of directions.*

For example, let us consider in $\mathbb{R}^2$ the control system

$$\frac{dx}{dt} = 2 + u^1 \cos(u^2), \quad \frac{dy}{dt} = u^1 \sin(u^2), \qquad 0 \leq u^1 \leq 1, \quad u^2 \in S^1,$$

for which, in a particular point $(x_0, y_0) \in \mathbb{R}^2$, the subset $S_{(x_0,y_0)}$ is the unit disk centered at $(2, 0)$. In a time optimal problem for this control system, the subset of possible optimal allowed velocities $S^*_{(x_0,y_0)}$ ($\equiv S^o_{(x_0,y_0)}$ in this case) is determined by $u^1 = 1$ and $\frac{-2\pi}{3} \leq u^2 \leq \frac{2\pi}{3}$, the boundary of the disk made of the longest vectors. (See Figure 1.)

$D_{(x_0,y_0)}$ is determined by the conditions $|\frac{v_y}{v_x}| \leq \frac{1}{\sqrt{3}}$ and $v_x > 0$. The Finslerian function, the factor between elements of $D_{(x_0,y_0)}$ and the element of $S^o_{(x_0,y_0)}$ on the same ray, is explicitly given by

$$F_{(x_0,y_0)}(v_x, v_y) = \frac{v_x^2 + v_y^2}{2v_x + \sqrt{v_x^2 - 3v_y^2}},$$

a not very simple homogeneous function defined on $D_{(x_0,y_0)}$.

Next we present another illustrative example of a time optimal problem, where the bang-bang phenomenon is clearly illustrated through the indicatrix of the associated metric. Given the system of control equations

$$\frac{dx}{dt} = u_1 |\cos u_2| \cos u_2, \qquad \frac{dy}{dt} = u_1 |\cos u_2| \sin u_2,$$

with $0 \leq u_1 \leq 2$, $0 \leq u_2 < 2\pi$, the problem is to find the curve solution of the system joining two fixed endpoints with minimum parameter increasing ($(x, y)(0) = (x_1, y_1)$, $(x, y)(T) = (x_2, y_2)$, T minimum). It is a typical optimal control problem, defined in this case through the map $X \colon R \times \mathbb{R}^2 \to T\mathbb{R}^2$ given by

$$X(x, y, u_1, u_2) = u_1 |\cos u_2| \cos u_2 \frac{\partial}{\partial x} + u_1 |\cos u_2| \sin u_2 \frac{\partial}{\partial y},$$

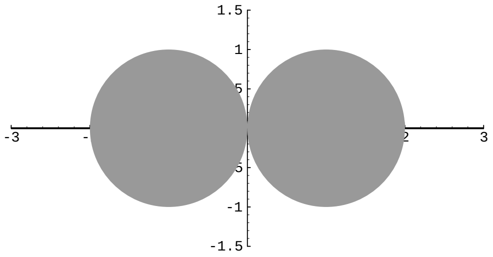


FIG. 2. *The set S in a generic $T_{(x,y)}\mathbb{R}^2$.*

where R is the direct product $[0, 2] \times S^1$. The image set $\mathrm{Im}(X) = S$ is given at every point of the plane by the union of two unit disks centered at $(1, 0)$ and $(-1, 0)$, respectively (see Figure 2). The family of functions (x, y, u_1, u_2) is a system of coordinates for $S \subset T\mathbb{R}^2$ which are related to the natural coordinates (x, y, v_x, v_y) by

$$v_x = u_1|\cos u_2|\cos u_2, \qquad v_y = u_1|\cos u_2|\sin u_2,$$

with inverse transformation

$$u_1 = \frac{v_x^2 + v_y^2}{|v_x|}, \qquad u_2 = \arctan\left(\frac{v_y}{v_x}\right).$$

Being a minimum time problem, the Hamiltonian of the optimal control system, defined on $R \times T^*\mathbb{R}^2$, is given by

$$H = (p_x \cos u_2 + p_y \sin u_2)u_1|\cos u_2| - 1.$$

In a first step, maximality is obtained on every direction of S for $u_1 = 2$, i.e., S^o is the boundary of the disks (see Figure 3), where coordinates (x, y, u_2) can be used. The cone of allowed ray directions is $D = \{v \in T\mathbb{R}^2; v_x \neq 0\}$, and every vector (v_x, v_y) in D is positive proportional to a vector (v_{xo}, v_{yo}) in S^o (see Figure 3) by a factor λ which determines the metric function F:

$$F(x, y, v_x, v_y) = \frac{v_x^2 + v_y^2}{2|v_x|}.$$

F is homogeneous positive of degree 1, and note that $F = 1$ reproduces the set S^o, the indicatrix. In a usual metric notation, we have defined, associated to the optimal control system, the Finslerian metric $ds = \frac{(dx)^2+(dy)^2}{2|dx|}$ on the allowed directions $\langle dx, v\rangle \neq 0$.

Now, by concavity it is clear in Figure 4 that for $u_2 \in (\pi/4, 3\pi/4) \cup (5\pi/4, 7\pi/4)$ the corresponding direction is never a maximum of the Hamiltonian (the movable normal line to the momentum vector never first contacts the indicatrix in this interval), and we have as optimal set S^* of allowed directions those of slope $(v_y/v_x)^2 \leq 1$ on S^o. Geodetic curves of the metric are segments of straight lines. (This fact can be easily determined by obtaining the Euler–Lagrange equations for the singular Lagrangian $L = \frac{v_x^2+v_y^2}{2|v_x|}$, which are equivalent to the condition v_y/v_x constant.)

Therefore, although we can measure lengths of curves with slope $(v_y/v_x)^2 > 1$, those curves will never be geodesics and, if the initial and final points are, for example, $(0, 0)$ and $(2, 1)$, the straight segment between them is not the geodesic. The geodesic

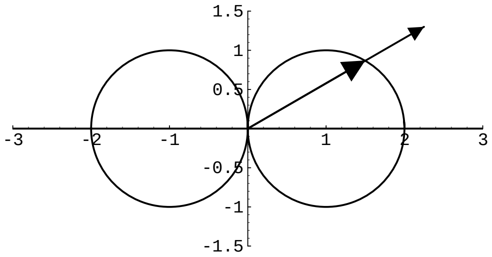


FIG. 3. *The set S^0, i.e., the indicatrix. The length of a vector is the factor between it and the parallel vector on the indicatrix.*

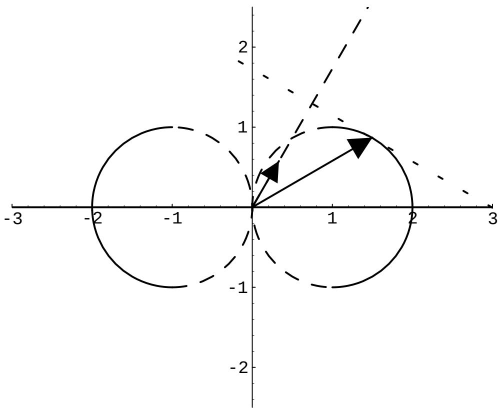


FIG. 4. *A given momentum vector and its corresponding optimal velocity. Dashed directions on the indicatrix can never be optimal and do not belong to S^*.*

(not unique) is a succession of segments with slopes 1 and -1, as, for example, $(0,0) \to (\frac{3}{2}, \frac{3}{2}) \to (2,1)$, and it contains at least one point of discontinuity on the derivative, a bang-bang on the control u_2 between $\pi/4$ and $3\pi/4$. The bang-bang phenomenon is here understood by the concavity-convexity properties of the indicatrix, i.e., by the fact that S^* is a proper subset of S^o.

4. The general case: The metric on state-time space. As it was pointed out in section 2, Pontryagin's maximum principle for the abnormal case as well as for nonautonomous systems can be formulated in the same framework by considering extra variables. Given a nonautonomous control system $\frac{dx^i}{dt} = f^i(x,t,u)$, a cost function $f^0(x,t,u)$, and some boundary conditions (for example, $x(t_0) = x_0$, $x(t_1) = x_1$, t_0, and t_1 fixed), let us consider the enlarged state space $Q = M \times \mathbb{R}^2$ with coordinates $(x^0, x^i|_{i=1,\cdots,n}, x^{n+1})$ and the enlarged (but autonomous) control system

$$\frac{dx^0}{d\tau} = f^0(x, x^{n+1}, u), \qquad \frac{dx^i}{d\tau} = f^i(x, x^{n+1}, u), \qquad \frac{dx^{n+1}}{d\tau} = 1,$$

with boundary conditions

$$x^0(\tau_0) = 0, \quad x^i(\tau_0) = x^i_0, \quad x^i(\tau_1) = x^i_1, \quad x^{n+1}(\tau_0) = t_0 \equiv \tau_0, \quad x^{n+1}(\tau_1) = t_1,$$

with τ_1 not fixed (although it is determined by the last control equation). It is clear that there is a correspondence between allowed curves for both systems, and optimal curves of the first system are curves of the second system with minimal final value $x^0(\tau_1)$. The Hamiltonian function in $U \times_Q T^*Q$ is defined, as before, as

$$H(x^0, x, x^{n+1}, p_0, p, p_{n+1}, u) = p_0 f^0(x, x^{n+1}, u) + p_i f^i(x, x^{n+1}, u) + p_{n+1},$$

and the same machinery works. It is immediate that p_0 is constant because there is no dependence of x^0 on H, and we can consider the abnormal case $p_0 = 0$ on the same foot as the normal one. The final parameter is free, so that $H = 0$ is an extra condition; in the case of autonomous systems, p_{n+1} is also constant by the adjoint equations, and we recover the constant autonomous Hamiltonian (for free t_1 the transversality condition $p_{n+1} = 0$ reproduces the vanishing autonomous Hamiltonian).

From another point of view, we can think of the optimal control system as a Pfaffian problem [6, 14] characterized by the family of the form

$$\mathcal{F} = \{dx^i - f^i(x, x^{n+1}, u)dx^{n+1}, dx^0 - f^0(x, x^{n+1}, u)dx^{n+1}\},$$

where allowed curves are now $\gamma(\epsilon) = (x^0(\epsilon), x(\epsilon), x^{n+1}(\epsilon))$, such that $\gamma^*(\mathcal{F}) = 0$. Tangent vectors to the allowed curves are therefore $\lambda \left(f^0 \frac{\partial}{\partial x^0} + f^i \frac{\partial}{\partial x^i} + \frac{\partial}{\partial x^{n+1}}\right)$ with $\lambda > 0$ arbitrary, i.e., arbitrary positive reparametrizations of the allowed curves for the enlarged control system.

Taking $\lambda = 1/f^0$ and discarding those directions for which $f^0 = 0$, we obtain an equivalent control system, $(f^i/f^0)\frac{\partial}{\partial x^i} + (1/f^0)\frac{\partial}{\partial x^{n+1}} + \frac{\partial}{\partial x^0}$. It is just a reparametrization of the allowed curves by the cost variable x^0, and for this new control system, the Hamiltonian is $H_c = p_i g^i + p_{n+1} g^{n+1} + p_0$, with $g^{n+1} = 1/f^0$ and $g^i = f^i/f^0$, defined again in $U \times_Q T^*Q$. This new optimal control system is the translation of the original optimal control system to a x^0-time optimal problem. Therefore, the construction of the metric can be made similarly to the one presented in section 3 for time optimal problems. In particular, we restrict ourselves to the normal case $p_0 = -1$, so that the explicit coordinate x^0 is not needed, and our state space will be $N = M \times \mathbb{R}$. The control system is $Y : U \to TN$

$$\frac{dx^i}{ds} = g^i(x, x^{n+1}, u), \quad \frac{dx^{n+1}}{ds} = g^{n+1}(x, x^{n+1}, u), \tag{4.1}$$

where s is the new parameter (the arc-length parameter in metric notation) and the Hamiltonian is defined on $W = U \times T^*N$

$$H(x, x^{n+1}, u, p, p_{n+1}) = p_i g^i(x, x^{n+1}, u) + p_{n+1} g^{n+1}(x, x^{n+1}, u) - 1 = \langle p, Y \rangle - 1.$$

Note that this construction is valid for an original autonomous system, where there is not dependence of the g on x^{n+1}, and p_{n+1} is a constant.

Once we fix a point (x_0, x_0^{n+1}) in N, we find a set of possible optimal controls $U^*_{(x_0, x_0^{n+1})}$, such that its image $Y(U^*) = S^*$ is made of some of the longest vectors in the set S of allowed velocities:

$$S^o = \{(x, x^{n+1}, v, v^{n+1}) \in S \mid \quad \lambda(v, v^{n+1}) \notin S \quad \forall \lambda > 1\}. \tag{4.2}$$

The metric is defined on the cone of directions

$$D = \{(v, v^{n+1}) | (v, v^{n+1}) = \lambda(v_o, v_o^{n+1}) \quad \text{with} \quad \lambda > 0 \quad \text{and} \quad (v_o, v_o^{n+1}) \in S^o\},$$

and the Finslerian function is $F(v, v^{n+1}) = \lambda$. For the actual solution, the curves are constrained to satisfy the condition $H = 0$; the algebraic condition of maximality determines the same velocities for (p_i, p_{n+1}) that for $\mu(p_i, p_{n+1})$, $\mu > 0$ (it is just the maximal projection $\max\langle(p_i, p_{n+1}), (v^i, v^{n+1})\rangle$ into the momentum ray). But the actual momentum vector of the ray is determined by

$$\langle (p_i, p_{n+1})^*, (v^i, v^{n+1})\rangle = 1, \qquad (v^i, v^{n+1}) \in S^*. \tag{4.3}$$

Let us develop as an example the classical Lagrangian mechanics, where the control space is the whole tangent space $S = TM$ and $f^0 = L = \frac{1}{2}g_{ij}(x)u^i u^j - V(x)$. We have the control equations $\frac{dx^i}{dt} = u^i$, and the cost function is the Lagrangian L. The associated optimal control system lives on $N = M \times \mathbb{R}$ of local coordinates $(x^i, x^{n+1} \equiv t)$. For the transformed minimal s-time problem, we have the Hamiltonian

$$H = p_i \frac{u^i}{L} + p_{n+1}\frac{1}{L} - 1$$

living in $W_N = TN \times_N T^*N$.

There is a unique vector on every allowed direction $(v_o^i, v_o^{n+1}) = (u^i/L, 1/L)$, and the points of S_x^* satisfy, eliminating the parameters u^i, the equation

$$\frac{1}{2}g_{ij}v_o^i v_o^j - V(x)(v_o^{n+1})^2 - v_o^{n+1} = 0,$$

which is an ellipsoid $V(x) < 0$ (e.g., the Kepler problem), paraboloid $V(x) = 0$ (free particle), or hyperboloid $V(x) > 0$ (harmonic oscillator; we must take the future branch of the hyperboloid for time increasing curves), always containing the origin.

The metric function F is obtained by looking for the factor between (v^i, v^{n+1}) $(v^{n+1} > 0)$ and the parallel element of S^*,

$$F(v^i, v^{n+1}) = v^{n+1}L = v^{n+1}\left[\frac{1}{2}g_{ij}\left(\frac{v^i}{v^{n+1}}\right)\left(\frac{v^j}{v^{n+1}}\right) - V(x)\right],$$

which is clearly homogeneous. It is simply the Lagrangian under an arbitrary change of parameter,

$$\int L(x,\dot{x})dt = \int L\frac{dt}{d\tau}d\tau = \int F d\tau,$$

with $\frac{dt}{d\tau} = v^{n+1}$ and $\frac{dx^i}{d\tau} = v^i$, i.e., $\dot{x} = \frac{dx^i}{dt} = \frac{v^i}{v^{n+1}}$. The consequence of this example is that the trajectories of a Lagrangian mechanical system can be seen as the geodesics of a Finslerian metric on space-time. We analyze in the next section the possibility of restricting the metric to constant energy leaves (by means of seeing it as an isoperimetric problem), and projecting the restricted metric to the state space; we obtain in this way a family of metric structures parametrized by the energy.

5. Restrictions of the metric. Let us note that the vanishing of the Hamiltonian on $W = U \times_N T^*N$ determines $-p_{n+1} = p_i f^i - f^0$, similar to the usual energy function in Lagrangian mechanics. For autonomous systems $\frac{\partial H}{\partial x^{n+1}} = 0$ gives $p_{n+1} = -E$ constant, usually nonvanishing for fixed final time problems, and $E = 0$ for free final time problems by transversality. We now restrict our attention to autonomous systems, and we try to find the restriction of the metric to a constant energy leaf. As we have pointed out before, the algebraic condition of optimality (2.4) for s-time optimal problems determines not only the elements (v_o^i, v_o^{n+1}) on S^o, the longest allowed velocity vectors, by

$$\max H(x,p,u) \equiv \max\langle(p_i, p_{n+1}), (v^i, v^{n+1})\rangle \qquad \forall u \in U_x,$$

once fixed $x \in M$ and a ray $\mu(p_i, p_{n+1})$ ($\mu > 0$), but also the particular momentum element $(p_i, p_{n+1})^*$ of the ray fulfilling the condition $\langle (p_i, p_{n+1})^*, (v_o^i, v_o^{n+1})\rangle = 1$. Let us denote by $P^o \subset T^*N$ this set of momentum vectors; it is made of one element on every ray of T^*N with positive projection over S^o. We can find the elements of P^o following the next steps.

1. Take a ray $\mu(p_i, p_{n+1})$ ($\mu > 0$) such that there are elements $(v^i, v^{n+1}) \in S$ for which $\langle (p_i, p_{n+1}), (v^i, v^{n+1})\rangle$ is positive.

2. Find the optimal velocity, the element $(v_o^i, v_o^{n+1}) \in S^o$, with maximal projection over the ray $\mu(p_i, p_{n+1})$.

3. Find the element $(p_i, p_{n+1})^*$ on the ray such that $\langle (p_i, p_{n+1})^*, (v_o^i, v_o^{n+1})\rangle = 1$.

There is generically a one-to-one correspondence between elements of S^o and elements of P^o, except in bang-bang situations, where a momentum vector is associated to two (or more) possible velocities. For the restriction to energy E leaves on S^o, we can begin by considering the subset P^E of P^o, made by the elements with $p_{n+1} = -E$, $P^E = \{(p_i, p_{n+1}) \in P^o \,|\, p_{n+1} = -E\}$. Its corresponding elements on S^o define the subset S^E of allowed optimal velocities with energy E. They are obtained by taking every element $p^E \in P^E$ and looking for the $g_E \in S^o$ with maximal projection; the condition $\langle p^E, g_E\rangle = 1$ is automatic by definition.

The restriction of the metric structure (N, D, F) to the energy leaves is obtained by defining a subset $D^E \subset D$ of the cone of allowed directions, and a Finslerian function F_E on D^E. D^E is the cone generated by the velocities in S^E

$$D^E = \{v = \lambda g_E | \quad \forall g_E \in S^E \quad \forall \lambda > 0\},$$

while $F_E\colon D^E \to \mathbb{R}$ is, as usual, $F_E(v) = \lambda$, the restriction of F to D^E.

This metric can be projected to the state space M by simply projecting S^E to a subset of velocities $T^E \subset TM$ through the natural projection $\mathrm{pr}_1\colon N = M \times \mathbb{R} \to M$, $T^E = \mathrm{pr}_{1*}(S^E)$, and considering the cone of directions C^E generated by T^E, $C^E = \mathrm{pr}_{1*}(D^E)$. Again, the Finslerian function (which will be denoted with the same letter F_E) is given by $F_E(v) = \lambda$ for $v \in C^E$ and $v = \lambda v_E$, $v_E \in T^E$.

As we said, the constancy of $p_{n+1} = -E$ is associated to fixed final parameter T. Therefore, we can think of the problem of restriction as an isoperimetric problem of minimizing the s-length among the curves with fixed t-length. Using a Lagrange multiplier α for the constraint $\int_0^T dt = T$, the complete restricted metric will be $G_E = F_E + \alpha F_t$, where F_E has already been defined and F_t is the t-metric, i.e., the metric whose associated length is the time increase of the curve measured. The Finslerian function F_t is simply given by $F_t(v, v^{n+1}) = v^{n+1}$ so that

$$\int_{\tau_i}^{\tau_f} F_t(x(\tau), v(\tau))d\tau = \int_{\tau_i}^{\tau_f} v^{n+1} d\tau = \int_{\tau_i}^{\tau_f} \frac{dt}{d\tau} d\tau = T.$$

A careful view to the Hamiltonian form $Hdt = p_i dx^i - (ds - p_{n+1}dt)$ shows that the value of the Lagrange multiplier is $\alpha = -p_{n+1} = E$, so that our restricted metric on D^E is given by the Finslerian function $G_E = F_E + EF_t$, and similarly for its projection to M. (We must project the function $F_t = v^{n+1}$ to C^E similarly to the defined projection of F_E.)

We have built a family of metric structures (M, C^E, G_E) on the state space, parametrized by the energy. Geodetic curves of these metrics are reparametrizations of the optimal solution curves of the optimal control system for different final parameters; i.e., fixed x_0 and x_1 on M, the family of optimal curves with boundary conditions

$x(0) = x_0$, $x(T) = x_1$, parametrized by T, is identified with the family of geodesics for the metrics G_E with those fixed endpoints, parametrized by E.

Continuing with the example of a Lagrangian mechanical system, let us consider a Lagrangian with kinetic term derived from a Riemannian metric $\frac{1}{2}g_{ij}v^iv^j$, and negative potential function $V(x) < 0$ (so that we have negative as well as positive energies), as, for example, the Kepler problem. Remember that the indicatrix in space-time is in this case an ellipsoid on the future semispace containing the origin. If we restrict our curves to the zero level energy (free final parameter), then covectors are $(p_i, 0)^* \in P^{E=0}$, optimality is obtained on the *equator* of the ellipsoid $v_o^{n+1} = \frac{-1}{2V(x)}$, and projection to the space manifold generates the metric

$$F_{E=0}(v) = \sqrt{-2V(x)g_{ij}v^iv^j}.$$

For another energy level E, optimality associated to the covector $(p_i, -E)^* \in P^E$ determines a closed set on the ellipsoid, south ($E > 0$) or north ($E < 0$) of the equator, determined by the equation of the ellipsoid, the condition $(p_i, -E)\cdot(v^i, v^{n+1})$ maximal and vanishing of the Hamiltonian, giving $g^{ij}p_ip_j = g_{ij}u^iu^j = 2(E - V)$, i.e., $v_o^{n+1} = 1/(E - 2V)$, intersection of the ellipsoid with a horizontal hyperplane. Projection to the spatial directions determines a particular metric F_E,

$$F_E(v) = \frac{E - 2V}{\sqrt{2(E - V)}}\sqrt{g_{ij}v^iv^j}.$$

Now, the t-metric $F_t(v) = v^{n+1}$ is projected to M giving

$$F_t(v) = v^{n+1} = F_E v_o^{n+1} = \frac{\sqrt{g_{ij}v^iv^j}}{\sqrt{2(E - V)}}.$$

Finally, the complete metric in the state space, obtained from F_E and F_t with Lagrange multiplier the energy, is

$$G_E = F_E + EF_t = \sqrt{2(E - V(x))}\sqrt{g_{ij}(x)v^iv^j},$$

and its geodesic curves are the trajectories of energy E for the Lagrangian system, i.e., the geodesics of the classical Jacobi metrics.

6. Conclusions and outlook. We have seen the relationship between a natural generalization of metric structure, using Finslerian functions and constrained directions, and time optimal problems in control systems. This relationship is clearly stated through the indicatrix of the metric (a simple way to define metric structures is to consider the set of unit norm velocities [4]), which is identified with the set of longest (optimal) velocities for the control system. The bang-bang phenomenon is, for example, associated to *holes* on the indicatrix. A reparametrization for general optimal control problems, using the control parameter as arc-length, allows us to extend the relationship to this general case where the metric is now defined on state-time space. The trajectories of a Lagrangian mechanical system are understood in this way as the geodesics of a Finslerian metric in space-time. A process of restriction to constant energy leaves determines a family of metric structures on state space parametrized by the energy, whose geodesics are the solutions of the optimal control problem for different final times. In the example of classical mechanics, this family is nothing but the Jacobi metrics associated to the Lagrangian, whose geodesics are the classical trajectories of a given energy. In some other analyzed examples [3, 10], the metric point of view for the optimal control problems has also been useful, at least qualitatively. The generalization of some concepts in Riemannian geometry to a Finslerian metric structure has been a field of research during recent years [6, 14, 4]. The link with optimal control problems allows us to think of giving one more step and try to generalize the study to metrics with restricted allowed directions. For example, the abnormal solutions of minimal length in sub-Riemannian geometry are naturally included on the set of geodesics [1, 2, 9] as the solutions with $p_0 = 0$ when the problem is seen from the optimal control point of view.

REFERENCES

[1] A. A. Agrachev, *Methods of control theory in non-holonomic geometry*, in Proceedings of the International Congress of Mathematicians, 1, 2, Zurich, Switzerland, 1994, Birkhäuser, Basel, Switzerland, 1995, pp. 1473–1483.

[2] A. A. Agrachev and A. V. Sarychev, *Abnormal sub-Riemannian geodesics: Morse index and rigidity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 635–690.

[3] A. Badía, C. López, and J. L. Giordano, *Optimal control model for the critical state in superconductors*, Phys. Rev. B, 58 (1998), pp. 9440–9449.

[4] R. L. Bryant, *Finsler manifolds with constant flag curvature*, in Proceedings of the Chern Symposium, 1998, http://www.msri.org/publications/video/spring98/chern.html.

[5] S. Chern, *Finsler geometry is just Riemannian geometry without the quadratic restriction*, Notices Amer. Math. Soc., 43 (1996), pp. 959–963.

[6] R. B. Gardner and G. R. Wilkens, *A pseudo-group isomorphism between control systems and certain generalized Finsler structures,* in Finsler Geometry (Seattle, WA, 1995), Contemp. Math. 196, Amer. Math. Soc., Providence, RI, 1996, pp. 231–243.

[7] M. J. Gotay and J. M. Nester, *Pre-symplectic Lagrangian systems* I: *The constraint algorithm and the equivalence theorem*, Ann. Inst. H. Poincaré A30 (1979), pp. 129–142.

[8] G. Leitmann, *The Calculus of Variations and Optimal Control*, Math. Concepts Methods Sci. Engrg. 24, A. Miele, ed., Plenum Press, New York, 1981.

[9] W. Liu and H. J. Sussman, *Shortest path for sub-Riemannian metrics on rank-two distributions*, Mem. Amer. Math. Soc. 118 (1995).

[10] C. López and V. Camarena, *Minimum Cost Transfer Orbits on the Kepler Problem: A Metric Point of View*, preprint, Universidad de Zaragoza, Zaragoza, Spain, 2000.

[11] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[12] L. S. Pontryagin, V. G. Bolyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[13] R. Skinner and R. Rusk, *Generalized Hamiltonian dynamics* I: *Formulation on* $T^*Q \oplus TQ$, J. Math. Phys., 24 (1983), pp. 2589–2594.

[14] G. R. Wilkens, *Finsler geometry in low dimensional control theory*, in Finsler Geometry (Seattle, WA, 1995), Contemp. Math. 196, Amer. Math. Soc., Providence, RI, 1996, pp. 245–257.

SIAM J. Control Optim.
Vol. 39, No. 3, pp. 812–834

EXPLICIT OBSERVABILITY INEQUALITIES FOR THE WAVE EQUATION WITH LOWER ORDER TERMS BY MEANS OF CARLEMAN INEQUALITIES[*]

XU ZHANG[†]

Abstract. In this paper, by means of Carleman estimates and the usual energy estimate, we obtain directly two observability inequalities for the linear wave equation with time-variant nonsmooth lower order terms. We do not need any unique continuation property of the linear equation a priori, since this is actually one of the by-products of our analysis. Furthermore, the constant in the observability inequality is estimated by an explicit function of the norm of the involved coefficients in the equation. Also, we apply our observability estimates to exact controllability for wave equations.

Key words. Carleman estimate, energy estimate, observability, controllability, wave equation

AMS subject classifications. 35B60, 35L05, 35Q60, 93B05, 93B07

PII. S0363012999350298

1. Introduction.

Let us consider the following wave equation:

$$\begin{cases} w_{tt} - \Delta w = q_1(t,x)w + q_2(t,x)w_t + \langle\, q_3(t,x), \nabla w\,\rangle & \text{in } Q,\\ w = 0 & \text{on } \Sigma,\\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in } \Omega. \end{cases} \tag{1.1}$$

In (1.1), $Q \overset{\triangle}{=} (0,T)\times\Omega$, $\Sigma \overset{\triangle}{=} (0,T)\times\Gamma$, $T>0$, $\Omega\subset\mathbb{R}^n$ is a bounded domain with a C^2 boundary $\Gamma \overset{\triangle}{=} \partial\Omega$, $q_i(\cdot)$ $(i=1,2,3)$ are given functions allowed to be *time-variant and nonsmooth.* We are concerned about the following two observability problems.

Problem 1. Given an open subset Γ_0 of Γ and a $T>0$, find (if possible) a constant $\mathcal{C}>0$ such that

$$|w_0|^2_{H^1_0(\Omega)} + |w_1|^2_{L^2(\Omega)} \le \mathcal{C}\left|\frac{\partial w}{\partial\nu}\right|^2_{L^2(\Sigma_0)} \tag{1.2}$$

for all weak solutions $w\in C([0,T];H^1_0(\Omega))\cap C^1([0,T];L^2(\Omega))$ of (1.1). Here, ν is the unit outward normal vector of Ω on Γ, $\Sigma_0 \overset{\triangle}{=} (0,T)\times\Gamma_0$.

Problem 2. Given an open subset Γ_0 of Γ and a $T>0$, find (if possible) a constant $\mathcal{C}>0$ such that

$$|w_0|^2_{L^2(\Omega)} + |w_1|^2_{H^{-1}(\Omega)} \le \mathcal{C}\left|\frac{\partial w}{\partial\nu}\right|^2_{H^{-1}(\Sigma_0)} \tag{1.3}$$

for all weak solutions $w\in C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))$ (in the sense of transposition) of (1.1) with $q_2=0$ and $q_3=0$.

[*] Received by the editors January 4, 1999; accepted for publication (in revised form) March 27, 2000; published electronically October 4, 2000.
http://www.siam.org/journals/sicon/39-3/35029.html

[†] Institute of Mathematics, Fudan University, Shanghai 200433, People's Republic of China. Present address: School of Mathematics, Sichuan University, Chengdu 610064, Sichuan Province, People's Republic of China (xuzhang@fudan.edu). This research was partially supported by NSF of China under grant 19901024 and the Chinese Education Ministry Science Foundation.

Note that the constant $\mathcal{C}$ in (1.2) and/or (1.3) depends on the lower order term coefficients $q_i(\cdot)$ $(i = 1, 2, 3)$ in (1.1). The explicit estimate of $\mathcal{C}$ via the norm of coefficients is a part of the problem and it is the main novelty of this paper. We remark that explicit observability estimates are crucial for some problems (see, for example, [19] and [32] and so on).

It is well known that the above observability estimates are closely related to the exact controllability of the following wave equation:

$$\begin{cases} y_{tt} - \Delta y = p_1(t,x)y + p_2(t,x)y_t + \langle p_3(t,x), \nabla y \rangle & \text{in } Q, \\ y = u\chi_{\Sigma_0}(t,x) & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } \Omega, \end{cases} \tag{1.4}$$

where χ_{Σ_0} stands for the characteristic function of set Σ_0. The exact controllability problem may be formulated as follows: For any (y_0, y_1), $(z_0, z_1) \in L^2(\Omega) \times H^{-1}(\Omega)$, find a control $u \in L^2(\Sigma_0)$ such that the weak solution of (1.4) satisfies

$$y(T) = z_0, \quad y_t(T) = z_1. \tag{1.5}$$

If $p_i(t,x) \equiv p_i(x)$ $(i = 1, 2, 3)$, system (1.4) is a linear time-invariant system, and the study of exact controllability for this case seems to be complete, especially for the case $p_i(\cdot) \equiv 0$ $(i = 1, 2, 3)$ (see [2, 3, 7, 16, 17, 30]). Two classical references in the theory of controllability are Russell's work [21] and Lions's monograph [17]. Other interesting related works can be found in [8, 11, 12, 15, 18, 22, 23, 26, 31, 32, 33] and the references cited therein.

By duality arguments [7, 16, 17, 30], we know that the exact controllability of a linear system can be reduced to the observability estimate of its dual system. On the other hand, by Zuazua's method in [31, 32], we know that the exact controllability of the semilinear system can be reduced to such a sort of estimate, provided also we know how the observability constant depends on the coefficients in the "linearized" systems. Thus, one of the main problems in the theory of exact controllability is how to construct the observability estimate for the linear system.

In the literature, one can find two important methods to derive observability estimates. The first one combines multiplier techniques and compactness-uniqueness arguments (see [2, 4, 7, 16, 17, 26, 30] and the rich references cited therein). The second one replaces the multipliers techniques by Carleman-type inequalities (see, [3, 11, 12, 22, 23][1]). We remark that, to the best of our knowledge, both of these methods do not give any estimate on the observability constant $\mathcal{C}$ in (1.2) or (1.3) because of the contradiction argument that is needed to absorb lower order terms. Furthermore, we note that, except for some special cases, both of these methods depend one way or another on some sort of uniqueness (i.e., unique continuation property (UCP for short) of the involved linear system. By this, for example, for system (1.1) we mean that if whenever $w \in H^1(Q)$ satisfies (1.1) and $\partial w/\partial\nu\big|_{\Sigma_0} = 0$, w is then identically zero in Q). There are many related UCP results in the literature (see [5, 20, 24, 25] and the rich references cited therein). However, to our knowledge, these results may not be applied directly to the exact controllability of system (1.4). For example, Ruiz's UCP theorem [20] applies only to the case $\Gamma_0 \equiv \Gamma$. On the other hand, the results in [24, 25] are of a *local* nature, compared with the *global* UCP property defined

[1] In [3], Fursikov and Imanuvilov derived the observability estimate for the parabolic equations without using the compact-uniqueness argument. However, they have used such a sort of argument when they constructed the observability estimate for the hyperbolic equations.

above; furthermore, the results in [24, 25] need at least *partial analyticity.* It would be natural to expect that the local UCP can be iterated or "patched" together to produce global ones. However, this remains to be done. On the other hand, it was showed in [1] that local UCP may fail for the wave equations with *time-variant nonanalytic* coefficients.

In this paper, we will use a different method combining Carleman and energy estimates to derive the desired observability estimates (1.2) and (1.3) directly. Our method was stimulated by the work of [6] and [14], which has the following advantages: (1) We can give explicit estimates on the constants $\mathcal{C}$ in (1.2) and (1.3); (2) We do not need a priori any UCP for (1.1). Indeed, in our approach, UCP is a by-product consequence of the observability inequality.

In what follows, we will denote the norms in $L^s(Q)$ and in $W^{1,s}(Q)$ ($s \in [1,\infty]$) simply by $|\cdot|_s$ and $|\cdot|_{1,s}$, respectively. Furthermore, we will use C to denote a generic positive constant which may change from line to line.

The rest of this paper is organized as follows. In section 2, we collect some preliminaries. The main results are stated and proved in section 3. In section 4, we apply our observability estimates to exact controllability problem related to wave equations.

2. Preliminaries. We need the following preliminaries, which are essentially known although most of which are not explicitly listed in the literature (but follow directly from the known references).

First, let us consider

$$\begin{cases} w_{tt} - \Delta w = F & \text{in } Q, \\ w = g & \text{on } \Sigma, \\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in } \Omega. \end{cases} \tag{2.1}$$

The following result is well known (see [9, Theorem 2.1 and Remark 2.2]).

LEMMA 2.1. *Let $T > 0$ be given. Suppose that*

$$\begin{cases} F \in L^1(0,T;L^2(\Omega)), \\ g \in H^1(\Sigma), \\ w_0 \in H^1(\Omega), \ w_1 \in L^2(\Omega), \end{cases} \tag{2.2}$$

satisfying the compatibility condition $g\big|_{t=0} = w_0\big|_\Gamma$. Then the unique weak solution w of (2.1) *satisfies*

$$\begin{cases} w \in C([0,T];H^1(\Omega)) \cap C^1([0,T];L^2(\Omega)), \\ \dfrac{\partial w}{\partial \nu} \in L^2(\Sigma). \end{cases} \tag{2.3}$$

Furthermore, there is a constant $C = C(T,\Omega) > 0$ such that

$$\begin{aligned} &|w|_{C([0,T];H^1(\Omega))\cap C^1([0,T];L^2(\Omega))} + \left|\frac{\partial w}{\partial \nu}\right|_{L^2(\Sigma)} \\ &\le C\Big(|F|_{L^1(0,T;L^2(\Omega))} + |g|_{H^1(\Sigma)} + |w_0|_{H^1(\Omega)} + |w_1|_{L^2(\Omega)}\Big). \end{aligned} \tag{2.4}$$

Combining the results in [9] and [17], one can get the following lemma.

Lemma 2.2. *Let $T > 0$ be given. Suppose that*

$$\begin{cases} F \in L^1(0,T;H^{-1}(\Omega)), \\ g \in L^2(\Sigma), \\ w_0 \in L^2(\Omega), \; w_1 \in H^{-1}(\Omega). \end{cases} \tag{2.5}$$

Then the unique weak solution w of (2.1) *satisfies*

$$\begin{cases} w \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega)), \\ \dfrac{\partial w}{\partial \nu} \in H^{-1}(\Sigma). \end{cases} \tag{2.6}$$

Furthermore, there is a constant $C = C(T,\Omega) > 0$ such that

$$\begin{aligned} |w|_{C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))} + \Big|\frac{\partial w}{\partial \nu}\Big|_{H^{-1}(\Sigma)} & \\ \le C\Big(|F|_{L^1(0,T;H^{-1}(\Omega))} + |w_0|_{L^2(\Omega)} & \\ +|w_1|_{H^{-1}(\Omega)} + |g|_{L^2(\Sigma)}\Big). & \end{aligned} \tag{2.7}$$

We refer to Appendix A at the end of this paper for the proof of Lemma 2.2.

Next, let us consider

$$\begin{cases} w_{tt} - \Delta w = p_1(t,x)w + p_2(t,x)w_t + \langle\, p_3(t,x), \nabla w\,\rangle + g_1 & \text{in } Q, \\ w = g_2 & \text{on } \Sigma, \\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in } \Omega. \end{cases} \tag{2.8}$$

Using the usual energy estimate and noting the time reversibility of (2.8), proceeding as in [32], one can obtain the following result.

Lemma 2.3. *We have the following two conclusions:*

(1) *Let $T > 0$, $p_1 \in L^1(0,T;L^n(\Omega))$, $p_2 \in L^\infty(Q)$, $p_3 \in L^\infty(Q;\mathbb{R}^n)$, $g_1 \in L^1(0,T;L^2(\Omega))$, $g_2 = 0$, $w_0 \in H_0^1(\Omega)$, and $w_1 \in L^2(\Omega)$. Then* (2.8) *admits a unique weak solution $w(\cdot) \in C([0,T];H_0^1(\Omega)) \cap C^1([0,T];L^2(\Omega))$, which satisfies*

$$\mathcal{E}(t) \le C(\mathcal{E}(s) + |g_1|^2_{L^1(0,T;L^2(\Omega))})e^{C(|p_1|^{1/2}_{L^1(0,T;L^n(\Omega))}+|p_2|_\infty+|p_3|_\infty)} \quad \forall\, t,s \in [0,T] \tag{2.9}$$

for some constant $C = C(T,\Omega) > 0$, where

$$\mathcal{E}(t) \overset{\triangle}{=} |w_t(t,\cdot)|^2_{L^2(\Omega)} + |w(t,\cdot)|^2_{H_0^1(\Omega)}. \tag{2.10}$$

(2) *Let $T > 0$, $p_1 \in L^1(0,T;L^n(\Omega))$, $p_2 \in W^{1,\infty}(Q)$, $p_3 \in W^{1,\infty}(Q;\mathbb{R}^n)$, $g_1 = 0$, $g_2 = 0$, $w_0 \in L^2(\Omega)$, and $w_1 \in H^{-1}(\Omega)$. Then* (2.8) *admits a unique weak solution $w(\cdot) \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))$, which satisfies*

$$E(t) \le CE(s)e^{C(|p_1|_{L^1(0,T;L^n(\Omega))}+|p_2|_{1,\infty}+|p_3|_{1,\infty})} \quad \forall\, t,s \in [0,T] \tag{2.11}$$

for some constant $C = C(T,\Omega) > 0$, where

$$E(t) \overset{\triangle}{=} |w_t(t,\cdot)|^2_{H^{-1}(\Omega)} + |w(t,\cdot)|^2_{L^2(\Omega)}. \tag{2.12}$$

Further, suppose $p_2 = 0$ and $p_3 = 0$. Then there is a constant $C = C(T, \Omega) > 0$ such that

$$E(t) \le CE(s)e^{C|p_1|^{1/2}_{L^1(0,T;L^n(\Omega))}} \quad \forall\ t, s \in [0, T]. \tag{2.13}$$

We also need the following result, which is a simple consequence of Lemmas 2.1 and 2.3.

LEMMA 2.4. *Let $T > 0$, $p_1 \in L^1(0, T; L^n(\Omega))$, $p_2 \in L^\infty(Q)$, $p_3 \in L^\infty(Q; \mathbb{R}^n)$, $g_1 \in L^1(0, T; L^2(\Omega))$, $g_2 = 0$, $w_0 \in H_0^1(\Omega)$, and $w_1 \in L^2(\Omega)$. Then the unique weak solution $w \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of* (2.8) *satisfies*

$$\frac{\partial w}{\partial \nu} \in L^2(\Sigma). \tag{2.14}$$

Furthermore, there is a constant $C = C(T, \Omega, p_1, p_2, p_3) > 0$ such that

$$\begin{aligned} &|w|_{C([0,T];H_0^1(\Omega))\cap C^1([0,T];L^2(\Omega))} + \left|\frac{\partial w}{\partial \nu}\right|_{L^2(\Sigma)} \\ &\le C\Big(|g_1|_{L^1(0,T;L^2(\Omega))} + |w_0|_{H_0^1(\Omega)} + |w_1|_{L^2(\Omega)}\Big). \end{aligned} \tag{2.15}$$

Further, we need the following result, which is a simple consequence of Lemma 2.2.

THEOREM 2.5. *Let $T > 0$, $p_1 \in L^1(0, T; L^n(\Omega))$, $p_2 \in W^{1,\infty}(Q)$, $p_3 \in W^{1,\infty}(Q; \mathbb{R}^n)$, $g_1 = 0$, $g_2 \in L^2(\Sigma)$, $w_0 \in L^2(\Omega)$, and $w_1 \in H^{-1}(\Omega)$. Then system* (2.8) *admits a unique weak solution $w \in C([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-1}(\Omega))$. Furthermore, there is a constant $C = C(T, \Omega, p_1, p_2, p_3) > 0$ such that*

$$\begin{aligned} &|w|_{C(0,T;L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))} + \left|\frac{\partial w}{\partial \nu}\right|_{H^{-1}(\Sigma)} \\ &\le C(|w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)} + |g_2|_{L^2(\Sigma)}). \end{aligned} \tag{2.16}$$

Remark 2.6. Take $p_2 \equiv 0$ and $p_i(t, x) \equiv p_i(x)$ $(i = 1, 3)$ in Theorem 2.5, and thus we obtain the main result in [10].

Proof of Theorem 2.5. Let us use the transposition method [17]. First of all, we note that without loss of generality, we may assume $p_2 \equiv 0$. In fact, one can always reduce the general problems to such a special case by means of the following simple transformation:

$$\bar{w}(t, x) \overset{\triangle}{=} w(t, x)e^{-\frac{1}{2}\int_0^t p_2(s,x)ds}, \qquad (t, x) \in Q. \tag{2.17}$$

We define a linear functional $L(f)$ on the space $L^1(0, T; L^2(\Omega))$ by

$$\begin{aligned} L(f) \overset{\triangle}{=} \langle\, w_1, \theta(0)\, \rangle_{H^{-1}(\Omega),H_0^1(\Omega)} - \int_\Omega w_0(x)\theta_t(0, x)dx - \int_\Sigma g_2 \frac{\partial \theta}{\partial \nu} d\Sigma& \\ \forall\, f \in L^1(0, T; L^2(\Omega)),& \end{aligned} \tag{2.18}$$

where $\theta \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ solves

$$\begin{cases} \theta_{tt} - \Delta\theta - p_1\theta + \nabla \cdot (p_3\theta) = f & \text{in } Q, \\ \theta = 0 & \text{on } \Sigma, \\ \theta(T) = \theta_t(T) = 0 & \text{in } \Omega. \end{cases} \tag{2.19}$$

We assert that the functional $L(f)$ is bounded. In fact, by (2.18) and Lemma 2.4, and noting the time reversibility of (2.19), one gets

$$\begin{aligned}|L(f)| &\leq |\langle\, w_1, \theta(0)\,\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}| + \left|\int_\Omega w_0\theta_t(0)dx\right| + \left|\int_\Sigma g_2 \frac{\partial\theta}{\partial\nu} d\Sigma\right| \\ &\leq C(|w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)} + |g_2|_{L^2(\Sigma)}) \\ &\quad \times \left(|\theta|_{C([0,T];H_0^1(\Omega))\cap C^1([0,T];L^2(\Omega))} + \left|\frac{\partial\theta}{\partial\nu}\right|_{L^2(\Sigma)}\right) \\ &\leq C(|w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)} + |g_2|_{L^2(\Sigma)})|f|_{L^1(0,T;L^2(\Omega))}.\end{aligned} \tag{2.20}$$

It is known that any linear bounded functional on the space $L^1(0,T;L^2(\Omega))$ can be written as

$$L(f) = \int_Q wf dxdt, \tag{2.21}$$

where w is some function from the space $L^\infty(0,T;L^2(\Omega))$. Thus, system (2.8) admits a weak solution $w \in L^\infty(0,T;L^2(\Omega))$, which satisfies

$$|w|_{L^\infty(0,T;L^2(\Omega))} \leq C(|w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)} + |g_2|_{L^2(\Sigma)}). \tag{2.22}$$

Now, denote $F \overset{\triangle}{=} p_1 w + \langle\, p_3, \nabla w\,\rangle \equiv (p_1 - \nabla\cdot p_3)w + \nabla\cdot(p_3 w)$ (recall that without loss of generality, we have assumed $p_2 \equiv 0$). Then, by $w \in L^\infty(0,T;L^2(\Omega))$, we see that $F \in L^1(0,T;H^{-1}(\Omega))$. Thus, the desired result follows from Lemma 2.2 and (2.22) immediately. □

In what follows, we use the notation

$$f_i = f_i(x) \overset{\triangle}{=} \frac{\partial f(x)}{\partial x_i}, \qquad i = 1,2,\ldots,n; \qquad \sum_i \overset{\triangle}{=} \sum_{i=1}^n$$

(on the other hand, x_i is always the ith coordinate of the point x).

Finally, we need the following known pointwise estimate, which is a special case of [14, Lemma 1, p. 124] and [19, Lemma 5.1].

LEMMA 2.7. *Let $\lambda > 0$ and $\alpha_1, \alpha_2 \in (0,1)$ be constant. Let $x_0 \in \mathbb{R}^n$, $T > 0$, and*

$$\begin{cases} \psi(t,s,x) = [|x-x_0|^2 - \alpha_1(t-T/2)^2 - \alpha_2(s-T/2)^2]/2, \\ \ell = \lambda\psi, \\ \beta \overset{\triangle}{=} \min(n+\alpha_1-1, n+\alpha_2-1), \quad \Psi = \beta\lambda. \end{cases} \tag{2.23}$$

Let $z = z(t,s,x) \in C^2(\mathbb{R}^{2+n})$. Denote

$$v \overset{\triangle}{=} \theta z \quad \textit{with} \quad \theta = e^\ell. \tag{2.24}$$

Then

$$
\begin{aligned}
&\theta^2|z_{tt}+z_{ss}-\Delta z|^2\\
&\geq \left[-2\ell_t\left(v_t^2-v_s^2+\sum_j v_j^2\right)-4\ell_s v_t v_s+4\sum_j(\ell_j v_t v_j)+2\Psi v_t v-2\ell_t(A+\Psi)v^2\right]_t\\
&\quad+\left[-2\ell_s\left(v_s^2-v_t^2+\sum_j v_j^2\right)-4\ell_t v_t v_s+4\sum_j(\ell_j v_s v_j)+2\Psi v_s v-2\ell_s(A+\Psi)v^2\right]_s\\
&\quad-2\sum_j\left[2\sum_i(\ell_i v_i v_j)-\ell_j\sum_i v_i^2-2\ell_t v_t v_j-2\ell_s v_s v_j\right.\\
&\qquad\qquad\left.+\Psi v_j v+\ell_j(v_t^2+v_s^2)-(A+\Psi)\ell_j v^2\right]_j\\
&\quad+2(n-\alpha_1+\alpha_2-\beta)\lambda v_t^2+2(n+\alpha_1-\alpha_2-\beta)\lambda v_s^2\\
&\quad+2(2-n-\alpha_1-\alpha_2+\beta)\lambda\sum_j v_j^2+Bv^2,
\end{aligned}
\tag{2.25}
$$

where

$$
A=\lambda^2\Big[\alpha_1^2(t-T/2)^2+\alpha_2^2(s-T/2)^2-|x-x_0|^2\Big]+(n+\alpha_1+\alpha_2-\beta)\lambda \tag{2.26}
$$

and

$$
\begin{aligned}
B=&2\lambda^3\Big[(2+n-\beta+\alpha_1+\alpha_2)|x-x_0|^2\\
&+\alpha_1^2(\beta-n-3\alpha_1-\alpha_2)(t-T/2)^2+\alpha_2^2(\beta-n-3\alpha_2-\alpha_1)(s-T/2)^2\Big]\\
&-\Big[2(\beta-n)(n-\beta+\alpha_1+\alpha_2)-2(\alpha_1+\alpha_2)(n+\alpha_1+\alpha_2)+\beta^2-2n\beta\Big]\lambda^2.
\end{aligned}
\tag{2.27}
$$

3. Statement and proof of the main results.

3.1. Statement of the main results. We need the following assumption:

(H) *Let* Γ_0 *be given by*

$$
\Gamma_0=\left\{x\in\Gamma\;\Big|\;(x-x_0)\cdot\nu(x)>0\right\},
$$

where $x_0\in\mathbb{R}^n\setminus\overline{\Omega}$ *is a fixed point,* $\nu(x)$ *denotes the unit outward normal vector of* Ω *at* $x\in\Gamma$.

The main results in this paper can be stated as follows.

Theorem 3.1. *Let* (H) *hold,* $T>2\max_{x\in\Omega}|x-x_0|$, $q_1\in L^{n+1}(Q)$, $q_2\in L^\infty(Q)$, *and* $q_3\in L^\infty(Q;\mathbb{R}^n)$. *Then for any weak solution* $w\in C([0,T];H_0^1(\Omega))\cap C^1([0,T];L^2(\Omega))$ *of* (1.1), *it holds that*

$$
|w_0|^2_{H_0^1(\Omega)}+|w_1|^2_{L^2(\Omega)}\leq\mathcal{C}(r)\Big|\frac{\partial w}{\partial\nu}\Big|^2_{L^2(\Sigma_0)}\qquad\forall\;(w_0,w_1)\in H_0^1(\Omega)\times L^2(\Omega) \tag{3.1}
$$

for some constant $\mathcal{C}(r)>0$ *with* $r\overset{\triangle}{=}|q_1|_{n+1}+|q_2|_\infty+|q_3|_\infty$. *Furthermore, the constant* $\mathcal{C}(r)$ *in* (3.1) *may be bounded as*

$$
\mathcal{C}(r)=C\exp(Cr^2) \tag{3.2}
$$

for some constant $C = C(T, \Omega) > 0$.

THEOREM 3.2. *Let* (H) *hold,* $T > 2\max_{x\in\Omega}|x - x_0|$, $q_1 \in L^\infty(Q)$, $q_2 = 0$, *and* $q_3 = 0$. *Then for any weak solution* $w \in C([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))$ *of* (1.1), *it holds that*

$$(3.3)\, |w_0|^2_{L^2(\Omega)} + |w_1|^2_{H^{-1}(\Omega)} \leq \mathcal{C}(h)\left|\frac{\partial w}{\partial \nu}\right|^2_{H^{-1}(\Sigma_0)} \qquad \forall\ (w_0, w_1) \in L^2(\Omega) \times H^{-1}(\Omega)$$

for some constant $\mathcal{C}(h) > 0$ *with* $h \overset{\triangle}{=} |q_1|_\infty$. *Furthermore, the constant* $\mathcal{C}(h)$ *in* (3.3) *may be bounded as*

$$(3.4) \qquad \mathcal{C}(h) = C\exp(Ch^2)$$

for some constant $C = C(T, \Omega) > 0$.

The proof of Theorems 3.1 and 3.2 will be given in the next two subsections. Now several remarks are in order.

Remark 3.3. In (H), we assume a technical condition $x_0 \notin \overline{\Omega}$. This is a key condition in our approach (to guarantee a controlled "right" sign of the lower-order term Bv^2 in (3.19) and (3.48)). In the case $x_0 \in \Omega$, proceeding as in [19], one can prove the same results as in Theorems 3.1–3.2 provided Σ_0 in (3.1) and (3.3) is replaced by $(0,T) \times \big(\mathcal{O}_\delta(\Gamma_0) \cap \Gamma\big)$, where $\delta > 0$ is any given constant and $\mathcal{O}_\delta(\Gamma_0) \overset{\triangle}{=} \{y \in \mathbb{R}^n \,\big|\, |y - x| < \delta \text{ for some } x \in \Gamma_0\}$.

Remark 3.4. From the proofs of Theorems 3.1–3.2 in the next two subsections, we see that the above results remain valid if (1.1) is replaced by the following inequality:

$$(1.1)' \qquad \begin{cases} |w_{tt} - \Delta w| \leq |q_1(t,x)w + q_2(t,x)w_t + \langle\, q_3(t,x), \nabla w\,\rangle\,| & \text{in } Q, \\ w = 0 & \text{on } \Sigma, \\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in } \Omega. \end{cases}$$

Thus, by taking $q_2 \equiv 0$, $q_1 \in L^\infty(Q)$, and $w \in H^2(Q)$ in Theorem 3.1, we can obtain the main result in [6]. Furthermore, we see that the scalar function w can be replaced by a vector-valued function, i.e., (1.1) can be replaced by a system with *coupled* lower order terms.

Remark 3.5. The originality of our method consists in the fact that we can give explicit estimates (3.2) and (3.4) of the constant $\mathcal{C}(r)$ in (3.1) and $\mathcal{C}(h)$ in (3.3), respectively, via the norm of the coefficients of (1.1). To the best of our knowledge, such sorts of estimates for dimensions $n \geq 2$ are not available in the literature. For the case $n = 1$, Zuazua [32] obtained a similar estimate, and such an estimate played a crucial role in the proof of his main result on exact controllability for the subcritical semilinear wave equations in one space dimension. However, we would like to point out that estimate (3.2) (resp., (3.4)) is not sharp. In fact, one may expect an estimate of the order of $e^{Cr^{1/2}}$ (resp., $e^{Ch^{1/2}}$), as indicated by [32] for the case $n = 1$.

Remark 3.6. Take $q_1 \in L^\infty(Q)$ in Theorem 3.1 (and noting Remark 3.4), we may obtain some of the main results of Lasiecka and Triggiani in [11]. On the other hand, Tataru [22, 23] has obtained observability results for a large class of partial differential equations without explicit estimates on the observability constant like that of (3.2) or (3.4) in our Theorems 3.1 and 3.2.

Remark 3.7. Our method is rather general, and it can be applied to internal observability estimate [19, 28], and/or other boundary conditions [13], and/or other equations [29].

3.2. Proof of Theorem 3.1.

Proof of Theorem 3.1. We divide the proof into several steps.

Step 1. The main idea of our proof is to use the pointwise estimate (2.25) in Lemma 2.7 (this idea is borrowed from [6]). For this purpose, we need to choose a suitable pseudoconvex function ψ, that is, to choose x_0, α_1, and α_2.

In order to choose x_0, α_1, and α_2, we proceed as in [6]. Put

$$R_0 \overset{\triangle}{=} \min_{x\in\Omega} |x - x_0|, \quad R_1 \overset{\triangle}{=} \max_{x\in\Omega} |x - x_0|, \tag{3.5}$$

where x_0 is given in (H). Then $R_0 > 0$ and $T > 2R_1$. Thus we can choose a constant $\alpha \in (0, 1)$ (close to 1) such that

$$R_1^2 < \alpha T^2/4. \tag{3.6}$$

Having chosen x_0 and α as above, we next introduce the desired pseudoconvex function ψ by setting

$$\psi = \psi(t, x) \overset{\triangle}{=} [|x - x_0|^2 - \alpha(t - T/2)^2]/2. \tag{3.7}$$

Step 2. We need the following notations. First, denote

$$\Lambda_j \overset{\triangle}{=} \{(t, x) \in Q \mid 2\psi(t, x) > R_0^2/(j + 2)\}, \tag{3.8}$$

where $j = 0, 1, 2$. Next, denote

$$\begin{cases} T_i \overset{\triangle}{=} T/2 - \varepsilon_i T, \quad T_i' \overset{\triangle}{=} T/2 + \varepsilon_i T, \\ Q_i \overset{\triangle}{=} (T_i, T_i') \times \Omega, \end{cases} \tag{3.9}$$

where $i = 0, 1$; $0 < \varepsilon_0 < \varepsilon_1 < 1/2$ will be given below.

In order to determine ε_i $(i = 0, 1)$, we need an idea in [11] (see also [13]). First of all, by (3.5)–(3.7), one sees that

$$\psi(0, x) = \psi(T, x) = (R_1^2 - \alpha T^2/4)/2 < 0 \quad \forall\, x \in \Omega. \tag{3.10}$$

Thus, one can find an $\varepsilon_1 \in (0, 1/2)$ (close to 1/2) such that (recall (3.8)–(3.9) for Λ_2, Q_1, T_1, and T_1')

$$\Lambda_2 \subset Q_1 \tag{3.11}$$

and for any $(t, x) \in ((0, T_1) \cup (T_1', T)) \times \Omega$ it holds that

$$\psi(t, x) < 0. \tag{3.12}$$

Next, noting that since $\{T/2\} \times \Omega \subset \Lambda_0$, one can find a small $\varepsilon_0 \in (0, \varepsilon_1)$ such that (recall (3.8) and (3.9) for Λ_0 and Q_0, respectively)

$$Q_0 \subset \Lambda_0. \tag{3.13}$$

Now, we note that (recall (2.27) for B)

$$B = B\chi_{\Lambda_2}(t, x) + B\chi_{Q\setminus\Lambda_2}(t, x). \tag{3.14}$$

By (3.7), we see that (recall (2.23) for α_1, α_2, and β)

$$\alpha_1 = \alpha, \quad \alpha_2 = 0, \quad \beta = n - 1, \tag{3.15}$$

where α is given in (3.6). Thus, by (2.27), (3.8), and (3.15), one sees easily that there exists a constant $\lambda_1 > 1$ such that for any $\lambda > \lambda_1$, it holds that

$$B\chi_{\Lambda_2}(t,x) \geq c_0\lambda^3\chi_{\Lambda_2}(t,x) \tag{3.16}$$

and

$$\left|B\chi_{Q\setminus\Lambda_2}(t,x)\right| \leq C\lambda^3 \tag{3.17}$$

for some constants $c_0 > 0$ and $C > 0$, which depend only on T and Ω.

Step 3. Let us use Lemma 2.7. For any given $\tau \in (0, T_1)$ and $\tau' \in (T_1', T)$ (recall (3.9) for T_1 and T_1'), denote

$$Q_\tau^{\tau'} \overset{\triangle}{=} (\tau, \tau') \times \Omega. \tag{3.18}$$

Let us observe (2.25), where $z = z(t,s,x)$ is replaced by $w = w(t,x)$, and ψ is given by (3.7). Integrating (2.25) on $Q_\tau^{\tau'}$, using integration by parts, and taking (1.1) into account, we arrive at (noting that by (2.24), $v = \theta w$)

$$\begin{aligned}
&2(1-\alpha)\lambda\int_{Q_\tau^{\tau'}}\left(v_t^2 + \sum_i v_i^2\right)dxdt + \int_{Q_\tau^{\tau'}} Bv^2dxdt\\
&\leq \int_Q \theta^2|q_1w + q_2w_t + \langle\, q_3, \nabla w\,\rangle\,|^2dxdt + \int_{\Sigma_0}\left|\frac{\partial v}{\partial\nu}\right|^2 d\Sigma_0\\
&\quad + C\lambda^3\left[\int_\Omega\Big(|v(\tau,x)|^2 + |v_t(\tau,x)|^2 + \sum_i|v_i(\tau,x)|^2\right.\\
&\qquad\left. + |v(\tau',x)|^2 + |v_t(\tau',x)|^2 + \sum_i|v_i(\tau',x)|^2\Big)dx\right] \qquad \forall\,\lambda > 1.
\end{aligned} \tag{3.19}$$

However, by $v = \theta w$ and $\theta = e^\ell$, by (2.23), (3.7), and (3.12), we get

$$\begin{aligned}
&\int_\Omega\left(|v(\tau,x)|^2 + |v_t(\tau,x)|^2 + \sum_i|v_i(\tau,x)|^2\right.\\
&\qquad\left. + |v(\tau',x)|^2 + |v_t(\tau',x)|^2 + \sum_i|v_i(\tau',x)|^2\right)dx\\
&\leq C\lambda^2\left[\int_\Omega\left(|w(\tau,x)|^2 + |w_t(\tau,x)|^2 + \sum_i|w_i(\tau,x)|^2\right.\right.\\
&\qquad\left.\left. + |w(\tau',x)|^2 + |w_t(\tau',x)|^2 + \sum_i|w_i(\tau',x)|^2\right)dx\right].
\end{aligned} \tag{3.20}$$

Further, by (3.7)–(3.8), (2.23)–(2.24), (3.14), and (3.16)–(3.18), we get

$$\begin{aligned}
\int_{Q_\tau^{\tau'}} Bv^2dxdt &= \int_{Q_\tau^{\tau'}\cap\Lambda_2} Bv^2dxdt + \int_{Q_\tau^{\tau'}\setminus\Lambda_2} Bv^2dxdt\\
&\geq c_0\lambda^3\int_{Q_\tau^{\tau'}\cap\Lambda_2} v^2dxdt - C\lambda^3e^{R_0^2\lambda/4}\int_Q w^2dxdt \quad \forall\,\lambda > \lambda_1.
\end{aligned} \tag{3.21}$$

Note that by (3.8), (3.11), and (3.18), we have $Q_\tau^{\tau'} \supset \Lambda_1$. Thus, by (3.21), for any $\lambda > \lambda_1$, we have

$$\begin{aligned}
&2(1-\alpha)\lambda \int_{Q_\tau^{\tau'}} \left(v_t^2 + \sum_i v_i^2 \right) dxdt + \int_{Q_\tau^{\tau'}} Bv^2 dxdt \\
&\geq c_1 \left[\lambda \int_{\Lambda_1} \left(v_t^2 + \sum_i v_i^2 \right) dxdt + \lambda^3 \int_{\Lambda_1} v^2 dxdt \right] \qquad (3.22)\\
&\quad - C\lambda^3 e^{R_0^2\lambda/4} \int_Q w^2 dxdt,
\end{aligned}$$

where $c_1 > 0$ and $C > 0$ are two constants which depend only on T and Ω.

Now, combining (3.19)–(3.20) and (3.22), we conclude that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}
&\int_{\Lambda_1} \left(v_t^2 + \sum_i v_i^2 \right) dxdt + \lambda^2 \int_{\Lambda_1} \theta^2 v^2 dxdt \\
&\leq C\lambda^{-1} \left\{ \int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt + \int_{\Sigma_0} \left|\frac{\partial v}{\partial \nu}\right|^2 d\Sigma_0 \right. \\
&\quad + \lambda^5 \left[\int_\Omega \left(|w(\tau,x)|^2 + |w_t(\tau,x)|^2 + \sum_i |w_i(\tau,x)|^2 \right.\right. \qquad (3.23)\\
&\quad \left.\left. + |w(\tau',x)|^2 + |w_t(\tau',x)|^2 + \sum_i |w_i(\tau',x)|^2 \right) dx \right] \\
&\quad \left. + \lambda^3 e^{R_0^2\lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned}$$

Integrating (3.23) with respect to τ and τ' from T_2, T_1 and T_1', T_2', respectively, we get

$$\begin{aligned}
&\int_{\Lambda_1} \left(v_t^2 + \sum_i v_i^2 \right) dxdt + \lambda^2 \int_{\Lambda_1} v^2 dxdt \\
&\leq C\lambda^{-1} \left\{ \int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt + \int_{\Sigma_0} \left|\frac{\partial v}{\partial \nu}\right|^2 d\Sigma_0 \right. \qquad (3.24)\\
&\quad \left. + \lambda^5 \int_Q \left(w^2 + w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^3 e^{R_0^2\lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned}$$

Consequently, by (2.22)–(2.23) and (3.7), recalling that $w = \theta^{-1} v$ with $\theta = e^\ell$, and using (3.24) and (1.1), we see that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}
&\int_{\Lambda_1} \theta^2 \left(w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^2 \int_{\Lambda_1} \theta^2 w^2 dxdt \\
&\leq C\lambda^{-1} \left\{ \int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt + \int_{\Sigma_0} \theta^2 \left|\frac{\partial w}{\partial \nu}\right. |^2 d\Sigma_0 \right. \qquad (3.25)\\
&\quad \left. + \lambda^5 \int_Q \left(w^2 + w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^3 e^{R_0^2\lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned}$$

Step 4. Let us estimate "$\int_Q \theta^2|q_1 w + q_2 w_t + \langle\, q_3, \nabla w\,\rangle\,|^2 dxdt$." By the Hölder inequality, the Sobolev embedding theorem, and the Poincaré inequality, we get (recalling $r \overset{\triangle}{=} |q_1|_{n+1} + |q_2|_\infty + |q_3|_\infty$)

$$\begin{aligned}
&\int_Q \theta^2|q_1 w + q_2 w_t + \langle\, q_3, \nabla w\,\rangle\,|^2 dxdt \\
&\qquad = \left\{\int_{\Lambda_1} + \int_{Q\setminus\Lambda_1}\right\} \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w\,\rangle\,|^2 dxdt \\
&\le C\left\{\int_{\Lambda_1} \theta^2 q_1^2 w^2 dxdt + \int_{\Lambda_1} \theta^2 (q_2^2 + |q_3|^2)(w_t^2 + |\nabla w|^2) dxdt \right.\\
&\qquad \left. + e^{R_0^2\lambda/3} \int_Q (q_1^2 w^2 + q_2^2 w_t^2 + |q_3|^2|\nabla w|^2) dxdt \right\}\\
&\le C\left\{ |q_1|_{n+1}^2 |\theta w|^2_{L^{2(n+1)/(n-1)}(\Lambda_1)} + \int_{\Lambda_1} \theta^2 (q_2^2 + |q_3|^2)(w_t^2 + |\nabla w|^2) dxdt \right.\\
&\qquad \left. + e^{R_0^2\lambda/3} \int_Q (q_1^2 w^2 + q_2^2 w_t^2 + |q_3|^2|\nabla w|^2) dxdt \right\}\\
&\le Cr^2 \left[|\theta w|^2_{H^1(\Lambda_1)} + \int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt + e^{R_0^2\lambda/3} \int_Q (w_t^2 + |\nabla w|^2) dxdt \right]\\
&\le Cr^2 \left[\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt + (1+\lambda^2) \int_{\Lambda_1} \theta^2 w^2 dxdt + e^{R_0^2\lambda/3} \int_Q (w_t^2 + |\nabla w|^2) dxdt \right].
\end{aligned}$$

(3.26)

Thus, combining (3.25) and (3.26), we see that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}
&\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt + \lambda^2 \int_{\Lambda_1} \theta^2 w^2 dxdt \\
&\le C_1 \lambda^{-1} \left\{ r^2 \left[\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt \right.\right.\\
&\qquad\qquad \left. + \lambda^2 \int_{\Lambda_1} \theta^2 w^2 dxdt + e^{R_0^2\lambda/3} \int_Q (w_t^2 + |\nabla w|^2) dxdt \right]\\
&\qquad \left. + \int_{\Sigma_0} \theta^2 \left|\frac{\partial w}{\partial \nu}\right|^2 d\Sigma_0 + \lambda^5 \int_Q \left(w^2 + w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^3 e^{R_0^2\lambda/4} \int_Q w^2 dxdt \right\},
\end{aligned}$$

(3.27)

where $C_1 > 0$ is a constant. Now, taking

$$\lambda_2 \overset{\triangle}{=} \max(\lambda_1, 2 + C_1 r^2), \tag{3.28}$$

by (3.27)–(3.28), we see that for any $\lambda > \lambda_2$ it holds that

$$\begin{aligned}
\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt \le C\lambda^{-1} &\left\{ \int_{\Sigma_0} \theta^2 \left|\frac{\partial w}{\partial \nu}\right|^2 d\Sigma_0 + r^2 e^{R_0^2\lambda/3} \int_Q (w_t^2 + |\nabla w|^2) dxdt \right.\\
&\left. + \lambda^5 \int_Q \left(w^2 + w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^3 e^{R_0^2\lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned}$$

(3.29)

Note that by (3.8) and (3.13), we have

$$\begin{aligned}\int_{\Lambda_1}\theta^2(w_t^2+|\nabla w|^2)dxdt &\geq \int_{\Lambda_0}\theta^2(w_t^2+|\nabla w|^2)dxdt\\ &\geq e^{R_0^2\lambda/2}\int_{Q_0}(w_t^2+|\nabla w|^2)dxdt.\end{aligned} \tag{3.30}$$

Thus, by (3.29)–(3.30), we see that for any $\lambda > \lambda_2$, it holds that

$$\begin{aligned}&\int_{Q_0}(|w_t|^2+|\nabla w|^2)dxdt\\ &\quad\leq C\lambda^{-1}\left\{e^{C\lambda}\int_{\Sigma_0}\left|\frac{\partial w}{\partial\nu}\right|^2 d\Sigma_0 + r^2e^{-R_0^2\lambda/6}\int_Q(|w_t|^2+|\nabla w|^2)dxdt\right.\\ &\qquad\left.+\lambda^5e^{-R_0^2\lambda/2}\int_Q\left(w^2+w_t^2+\sum_i w_i^2\right)dxdt+\lambda^3e^{-R_0^2\lambda/4}\int_Q w^2dxdt\right\}.\end{aligned} \tag{3.31}$$

Step 5. Let us complete the proof of Theorem 3.1. By (3.31), (2.10), and (3.28), using Poincaré inequality, we conclude that there is a constant $\lambda_3 > 0$, which depends only on T and Ω, such that (recall (3.9) for T_0 and T_0')

$$\int_{T_0}^{T_0'}\mathcal{E}(t)dt \leq C\left\{e^{C\lambda}\int_{\Sigma_0}\left|\frac{\partial w}{\partial\nu}\right|^2 d\Sigma_0+e^{-R_0^2\lambda/8}\int_0^T\mathcal{E}(t)dt\right\} \qquad \forall\,\lambda>\lambda_2+\lambda_3. \tag{3.32}$$

On the other hand, by (2.9) in Lemma 2.3 and (3.32), we arrive at

$$\mathcal{E}(0)\leq C_2\left\{e^{C_2\lambda}\int_{\Sigma_0}\left|\frac{\partial w}{\partial\nu}\right|^2 d\Sigma_0+e^{-R_0^2\lambda/8+C_2r}\mathcal{E}(0)\right\} \qquad \forall\,\lambda>\lambda_2+\lambda_3, \tag{3.33}$$

where $C_2 = C_2(T,\Omega)$ is a positive constant. However, it is easy to find a constant $\lambda_4 = \lambda_4(R_0, C_2) > 0$ such that

$$C_2e^{-R_0^2\lambda_4/8+C_2r}\leq 1/2. \tag{3.34}$$

Thus, by (3.33)–(3.34), one gets

$$\mathcal{E}(0)\leq Ce^{C\lambda}\int_{\Sigma_0}\left|\frac{\partial w}{\partial\nu}\right|^2 d\Sigma_0 \qquad \forall\,\lambda>\max(\lambda_2+\lambda_3,\lambda_4), \tag{3.35}$$

which is exactly the desired inequality (3.1). On the other hand, the explicit estimate (3.2) follows from (3.28) and (3.34)–(3.35) immediately. □

3.3. Proof of Theorem 3.2.

Proof of Theorem 3.2. The main idea of our proof is similar to that of Theorem 3.1. Also, the proof is divided into several steps.

Step 1. Let us introduce some notations. First, denote R_0 and R_1 as in (3.5) and choose α as in (3.6). Then we introduce the desired pseudoconvex function ψ by setting

$$\psi=\psi(t,s,x)\overset{\triangle}{=}[|x-x_0|^2-\alpha(t-T/2)^2-\alpha(s-T/2)^2]/2, \tag{3.36}$$

where x_0 is given in (H).

Next, denote

$$\begin{cases} \mathcal{Q} \overset{\triangle}{=} (0,T)\times(0,T)\times\Omega, \quad \mathcal{S} \overset{\triangle}{=} (0,T)\times(0,T)\times\Gamma, \quad \mathcal{S}_0 \overset{\triangle}{=} (0,T)\times(0,T)\times\Gamma_0, \\ T_i \overset{\triangle}{=} T/2-\varepsilon_i T, \quad T_i' \overset{\triangle}{=} T/2+\varepsilon_i T, \quad \mathcal{Q}_i \overset{\triangle}{=} (T_i,T_i')\times(T_i,T_i')\times\Omega \end{cases}$$

(3.37)

and

$$\Xi_j \overset{\triangle}{=} \left\{(t,s,x)\in\mathcal{Q} \,\middle|\, 2\psi(t,s,x) \geq R_0^2/(j+2)\right\}, \tag{3.38}$$

where $i=0,1,2$; $j=0,1,2$ and $0<\varepsilon_0<\varepsilon_1<\varepsilon_2<1/2$ will be given below.

In order to determine ε_i $(i=0,1,2)$, we proceed as in (3.10)–(3.13). First of all, by (3.5)–(3.6) and (3.36), one gets

$$\psi(0,s,x)=\psi(T,s,x)=(R_1^2-\alpha T^2/4)/2<0 \quad \forall\,(s,x)\in Q. \tag{3.39}$$

Thus, one can find an $\varepsilon_1\in(0,1/2)$ (close to $1/2$) such that (recall (3.37)–(3.38) for Ξ_2, $\mathcal{Q}_1$, T_1 and T_1')

$$\Xi_2\subset\mathcal{Q}_1 \tag{3.40}$$

and for any $(t,s,x)\in((0,T_1)\cup(T_1',T))\times Q$ and any $(s,t,x)\in((0,T_1)\cup(T_1',T))\times Q$ it holds that

$$\psi(t,s,x)<0. \tag{3.41}$$

Next, noting that since $\{T/2\}\times\{T/2\}\times\Omega\subset\Xi_0$, one can find a small $\varepsilon_0\in(0,\varepsilon_1)$ such that (recall (3.38) and (3.37) for Ξ_0 and $\mathcal{Q}_0$, respectively)

$$\mathcal{Q}_0\subset\Xi_0. \tag{3.42}$$

Finally, we fix any constant $\varepsilon_2\in(\varepsilon_1,1/2)$.

Now, let us observe B defined by (2.27), where x_0, α_1, α_2, and β are given in (3.36) and (2.23). Similar to (3.16)–(3.17), by (2.27) and (3.38), one can find a constant $\lambda_1>1$ such that for any $\lambda>\lambda_1$, it holds that

$$B\chi_{\Xi_2}(t,s,x)\geq c_0\lambda^3\chi_{\Xi_2}(t,s,x) \tag{3.43}$$

and

$$\left|B\chi_{\mathcal{Q}\setminus\Xi_2}(t,s,x)\right|\leq C\lambda^3 \tag{3.44}$$

for some constants $c_0>0$ and $C>0$, which depend only on T and Ω.

Finally, put

$$z(t,s,x)\overset{\triangle}{=}\int_s^t w(\xi,x)d\xi \qquad \forall\,(t,s,x)\in\mathcal{Q}, \tag{3.45}$$

where w is the weak solution of (1.1). One sees that z satisfies (recall $q_2=0$ and $q_3=0$)

$$\begin{cases} z_{tt}+z_{ss}-\Delta z=\displaystyle\int_s^t q_1(\xi,x)z_t(\xi,s,x)d\xi & \text{in } \mathcal{Q}, \\ z=0 & \text{on } \mathcal{S}. \end{cases} \tag{3.46}$$

Step 2. Let us use Lemma 2.7. We proceed as in [19]. For any given $\tau \in (T_2, T_1)$ and $\tau' \in (T_1', T_2')$ (recall (3.37) for T_i and T_i'), denote

$$\mathcal{Q}_\tau^{\tau'} \overset{\triangle}{=} (\tau, \tau') \times (\tau, \tau') \times \Omega. \tag{3.47}$$

Let us observe (2.25), where $z = z(t, s, x)$ is given by (3.45) and ψ is given by (3.36). Integrating (2.25) on $\mathcal{Q}_\tau^{\tau'}$, using integration by parts, and taking (3.46) into account, we arrive at (noting that by (2.24), $v = \theta z$)

$$\begin{aligned}
&2(1-\alpha)\lambda \int_{\mathcal{Q}_\tau^{\tau'}} \left(v_t^2 + v_s^2 + \sum_i v_i^2 \right) dxdtds + \int_{\mathcal{Q}_\tau^{\tau'}} Bv^2 dxdtds \\
&\le \int_{\mathcal{Q}} \theta^2 \left| \int_s^t q_1(\xi, x) z_t(\xi, s, x) d\xi \right|^2 dxdtds + \int_{\mathcal{S}_0} \left| \frac{\partial v}{\partial \nu} \right|^2 d\mathcal{S}_0 \\
&\quad + C\lambda^3 \Bigg[\int_{T_2}^{T_2'} \int_\Omega \Bigg(|v(\tau, s, x)|^2 + |v_t(\tau, s, x)|^2 + |v_s(\tau, s, x)|^2 + \sum_i |v_i(\tau, s, x)|^2 \\
&\quad + |v(\tau', s, x)|^2 + |v_t(\tau', s, x)|^2 + |v_s(\tau', s, x)|^2 + \sum_i |v_i(\tau', s, x)|^2 \Bigg) dxds \\
&\quad + \int_{T_2}^{T_2'} \int_\Omega \Bigg(|v(t, \tau, x)|^2 + |v_t(t, \tau, x)|^2 + |v_s(t, \tau, x)|^2 + \sum_i |v_i(t, \tau, x)|^2 \\
&\quad + |v(t, \tau', x)|^2 + |v_t(t, \tau', x)|^2 + |v_s(t, \tau', x)|^2 + \sum_i |v_i(t, \tau', x)|^2 \Bigg) dxdt \Bigg] \qquad \forall\, \lambda > 1.
\end{aligned} \tag{3.48}$$

However, recalling $v = \theta z$ with $\theta = e^\ell$, by (2.23), (3.36), and (3.41), we get

$$\begin{aligned}
&\int_{T_2}^{T_2'} \int_\Omega \Bigg(|v(\tau, s, x)|^2 + |v_t(\tau, s, x)|^2 + |v_s(\tau, s, x)|^2 + \sum_i |v_i(\tau, s, x)|^2 \\
&\quad + |v(\tau', s, x)|^2 + |v_t(\tau', s, x)|^2 + |v_s(\tau', s, x)|^2 + \sum_i |v_i(\tau', s, x)|^2 \Bigg) dxds \\
&\quad + \int_{T_2}^{T_2'} \int_\Omega \Bigg(|v(t, \tau, x)|^2 + |v_t(t, \tau, x)|^2 + |v_s(t, \tau, x)|^2 + \sum_i |v_i(t, \tau, x)|^2 \\
&\quad + |v(t, \tau', x)|^2 + |v_t(t, \tau', x)|^2 + |v_s(t, \tau', x)|^2 + \sum_i |v_i(t, \tau', x)|^2 \Bigg) dxdt \\
&\le C\lambda^2 \Bigg[\int_{T_2}^{T_2'} \int_\Omega \Bigg(|z(\tau, s, x)|^2 + |z_t(\tau, s, x)|^2 + |z_s(\tau, s, x)|^2 + \sum_i |z_i(\tau, s, x)|^2 \\
&\quad + |z(\tau', s, x)|^2 + |z_t(\tau', s, x)|^2 + |z_s(\tau', s, x)|^2 + \sum_i |z_i(\tau', s, x)|^2 \Bigg) dxds \\
&\quad + \int_{T_2}^{T_2'} \int_\Omega \Bigg(|z(t, \tau, x)|^2 + |z_t(t, \tau, x)|^2 + |z_s(t, \tau, x)|^2 + \sum_i |z_i(t, \tau, x)|^2 \\
&\quad + |z(t, \tau', x)|^2 + |z_t(t, \tau', x)|^2 + |z_s(t, \tau', x)|^2 + \sum_i |z_i(t, \tau', x)|^2 \Bigg) dxdt \Bigg].
\end{aligned} \tag{3.49}$$

Further, similar to (3.21), by (3.38), (2.23)–(2.24), (3.36), and (3.43)–(3.44), we get

$$\begin{aligned}\int_{\mathcal{Q}_\tau^{\tau'}} Bv^2 dxdtds &= \int_{\mathcal{Q}_\tau^{\tau'}\cap\Xi_2} Bv^2 dxdtds + \int_{\mathcal{Q}_\tau^{\tau'}\setminus\Xi_2} Bv^2 dxdtds \\ &\geq c_0\lambda^3 \int_{\mathcal{Q}_\tau^{\tau'}\cap\Xi_2} v^2 dxdtds - C\lambda^3 e^{R_0^2\lambda/4}\int_{\mathcal{Q}} z^2 dxdtds \quad \forall\,\lambda > \lambda_1.\end{aligned} \tag{3.50}$$

Note that by (3.38), (3.40), and (3.47) we have $\mathcal{Q}_\tau^{\tau'} \supset \Xi_1$. Thus, by (3.50), for any $\lambda > \lambda_1$, we have

$$\begin{aligned}&2(1-\alpha)\lambda\int_{\mathcal{Q}_\tau^{\tau'}}\left(v_t^2+v_s^2+\sum_i v_i^2\right)dxdtds + \int_{\mathcal{Q}_\tau^{\tau'}} Bv^2 dxdtds \\ &\geq c_1\left[\lambda\int_{\Xi_1}\left(v_t^2+v_s^2+\sum_i v_i^2\right)dxdtds + \lambda^3\int_{\Xi_1} v^2 dxdtds\right] - C\lambda^3 e^{R_0^2\lambda/4}\int_{\mathcal{Q}} z^2 dxdtds,\end{aligned} \tag{3.51}$$

where $c_1 > 0$ and $C > 0$ are two constants that depend only on T and Ω.

Now, combining (3.48)–(3.49) and (3.51), we conclude that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}&\int_{\Xi_1}\left(v_t^2+v_s^2+\sum_i v_i^2\right)dxdtds + \lambda^2\int_{\Xi_1}\theta^2 v^2 dxdtds \\ &\leq C\lambda^{-1}\left\{\int_{\mathcal{Q}}\theta^2\left|\int_s^t q_1(\xi,x)z_t(\xi,s,x)d\xi\right|^2 dxdtds + \int_{\mathcal{S}_0}\left|\frac{\partial v}{\partial\nu}\right|^2 d\mathcal{S}_0\right. \\ &\quad + \lambda^5\left[\int_{T_2}^{T_2'}\int_\Omega\left(|z(\tau,s,x)|^2+|z_t(\tau,s,x)|^2+|z_s(\tau,s,x)|^2+\sum_i|z_i(\tau,s,x)|^2\right.\right. \\ &\quad \left. + |z(\tau',s,x)|^2+|z_t(\tau',s,x)|^2+|z_s(\tau',s,x)|^2+\sum_i|z_i(\tau',s,x)|^2\right)dxds \\ &\quad + \int_{T_2}^{T_2'}\int_\Omega\left(|z(t,\tau,x)|^2+|z_t(t,\tau,x)|^2+|z_s(t,\tau,x)|^2+\sum_i|z_i(t,\tau,x)|^2\right. \\ &\quad \left.\left. + |z(t,\tau',x)|^2+|z_t(t,\tau',x)|^2+|z_s(t,\tau',x)|^2+\sum_i|z_i(t,\tau',x)|^2\right)dxdt\right] \\ &\quad \left. + \lambda^3 e^{R_0^2\lambda/4}\int_{\mathcal{Q}} z^2 dxdtds\right\}.\end{aligned} \tag{3.52}$$

Integrating (3.52) with respect to τ and τ' from T_2 to T_1 and from T_1' to T_2', respectively, we get

$$\begin{aligned}&\int_{\Xi_1}\left(v_t^2+v_s^2+\sum_i v_i^2\right)dxdtds + \lambda^2\int_{\Xi_1}\theta^2 v^2 dxdtds \\ &\leq C\lambda^{-1}\left\{\int_{\mathcal{Q}}\theta^2\left|\int_s^t q_1(\xi,x)z_t(\xi,s,x)d\xi\right|^2 dxdtds + \int_{\mathcal{S}_0}\left|\frac{\partial v}{\partial\nu}\right|^2 d\mathcal{S}_0\right. \\ &\quad \left. + \lambda^5\int_{\mathcal{Q}_2}\left(z^2+z_t^2+z_s^2+\sum_i z_i^2\right)dxdtds + \lambda^3 e^{R_0^2\lambda/4}\int_{\mathcal{Q}} z^2 dxdtds\right\}.\end{aligned} \tag{3.53}$$

Consequently, by (2.23)–(2.24) and (3.36), recalling that $z = \theta^{-1}v$ with $\theta = e^{\ell}$, and using (3.53), we see that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}
&\int_{\Xi_1}\theta^2\left(z_t^2+z_s^2+\sum_i z_i^2\right)dxdtds+\lambda^2\int_{\Xi_1}\theta^2 z^2dxdtds\\
&\le C\lambda^{-1}\left\{\int_{\mathcal{Q}}\theta^2\left|\int_s^t q_1(\xi,x)z_t(\xi,s,x)d\xi\right|^2dxdtds+e^{C\lambda}\int_{\mathcal{S}_0}\left|\frac{\partial z}{\partial\nu}\right|^2d\mathcal{S}_0\right.\\
&\quad\left.+\lambda^5\int_{\mathcal{Q}_2}\left(z^2+z_t^2+z_s^2+\sum_i z_i^2\right)dxdtds+\lambda^3e^{R_0^2\lambda/4}\int_{\mathcal{Q}}z^2dxdtds\right\}.
\end{aligned}\tag{3.54}$$

Step 3. Let us now estimate "$\int_{\mathcal{Q}_2}\theta^2\ \ |\int_s^t q_1(\xi,x)z_t(\xi,s,x)d\xi|^2dxdtds$" and "$\int_{\mathcal{Q}_2}\sum_i z_i^2dxdtds$." First, similar to [28], we get (recalling $h\overset{\triangle}{=}|q_1|_\infty$)

$$\begin{aligned}
\int_{\mathcal{Q}}\theta^2\left|\int_s^t q_1(\xi,x)z_t(\xi,s,x)d\xi\ \right|^2dxdtds&\le Ch^2\int_{\mathcal{Q}}\theta^2(z_t^2+z_s^2)dxdtds\\
&=Ch^2\left(\int_{\Xi_1}+\int_{\mathcal{Q}\setminus\Xi_1}\right)\theta^2(z_t^2+z_s^2)dxdtds\\
&\le Ch^2\left[\int_{\Xi_1}\theta^2(z_t^2+z_s^2)dxdtds\right.\\
&\qquad\left.+e^{R_0^2\lambda/3}\int_{\mathcal{Q}}(z_t^2+z_s^2)dxdtds\ \right].
\end{aligned}\tag{3.55}$$

Next, denote

$$\eta=\eta(t,s)\overset{\triangle}{=}t(T-t)s(T-s).\tag{3.56}$$

Multiplying the first equation of (3.46) by ηz, integrating it on $\mathcal{Q}$, using integration by parts, by (3.56) and noting that

$$\eta(t,s)\ge(T_2-T)^2(T-T_2)^2\quad\forall\, t\,,s\in(T_2,T_2'),$$

we get (recalling $h\overset{\triangle}{=}|q_1|_\infty$)

$$\int_{\mathcal{Q}_2}\sum_i z_i^2dxdtds\le C\left[\int_{\mathcal{Q}}(z_t^2+z_s^2+z^2)dxdtds+h\int_{\mathcal{Q}}(z_t^2+z_s^2+z^2)dxdtds\right].\tag{3.57}$$

Now, combining (3.54)–(3.55) and (3.57), we conclude that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}
&\int_{\Xi_1}\theta^2\left(z_t^2+z_s^2+\sum_i z_i^2\right)dxdtds+\lambda^2\int_{\Xi_1}\theta^2z^2dxdtds\\
&\le C_1\lambda^{-1}\left\{e^{C_1\lambda}\int_{\mathcal{S}_0}\left|\frac{\partial z}{\partial\nu}\right|^2d\mathcal{S}_0\right.\\
&\qquad+h^2\left[\int_{\Xi_1}\theta^2(z_t^2+z_s^2)dxdtds+e^{R_0^2\lambda/3}\int_{\mathcal{Q}}(z_t^2+z_s^2)dxdtds\right]\\
&\quad+\lambda^5\left[\int_{\mathcal{Q}}(z_t^2+z_s^2+z^2)dxdtds+h\int_{\mathcal{Q}}(z_t^2+z_s^2+z^2)dxdtds\right]\\
&\quad\left.+\lambda^3e^{R_0^2\lambda/4}\int_{\mathcal{Q}}z^2dxdtds\ \right\},
\end{aligned}\tag{3.58}$$

where $C_1 = C_1(T, \Omega) > 0$ is a constant.

Let us take

$$\lambda_2 \overset{\triangle}{=} \max(\lambda_1, 2 + C_1 h^2). \tag{3.59}$$

Then by (3.58)–(3.59), we see that for any $\lambda > \lambda_2$ it holds that

$$\begin{aligned}&\int {}_{\Xi_1}\theta^2(z_t^2 + z_s^2)dxdtds \\ &\leq C\lambda^{-1}\left\{e^{C\lambda}\int_{\mathcal{S}_0}\left|\frac{\partial z}{\partial\nu}\right|^2 d\mathcal{S}_0 + \lambda^5 e^{R_0^2\lambda/3}\int_{\mathcal{Q}}(z_t^2 + z_s^2 + z^2)dxdtds\right\}.\end{aligned} \tag{3.60}$$

Note that by (3.38) and (3.42), we have

$$\int_{\Xi_1}\theta^2(z_t^2 + z_s^2)dxdt \geq \int_{\Xi_0}\theta^2(z_t^2 + z_s^2)dxdt \geq e^{R_0^2\lambda/2}\int_{\mathcal{Q}_0}(z_t^2 + z_s^2)dxdt. \tag{3.61}$$

Thus, by (3.60)–(3.61), we conclude that for any $\lambda > \lambda_2$ it holds that

$$\int_{\mathcal{Q}_0}(z_t^2+z_s^2)dxdt \leq C\left\{e^{C\lambda}\int_{\mathcal{S}_0}\left|\frac{\partial z}{\partial\nu}\right|^2 d\mathcal{S}_0 + \lambda^5 e^{-R_0^2\lambda/6}\int_{\mathcal{Q}}(z_t^2 + z_s^2 + z^2)dxdtds\right\}. \tag{3.62}$$

Step 4. Let us complete the proof of Theorem 3.2. By (3.62) and (3.45), we get (recalling (3.37) for T_0 and T_0')

$$\int_{T_0}^{T_0'}\int_\Omega w^2 dxdt \leq C\left\{e^{C\lambda}\left|\frac{\partial w}{\partial\nu}\right|^2_{H^{-1}(\Sigma_0)} + \lambda^5 e^{-R_0^2\lambda/6}\int_0^T E(t)dt\right\} \qquad \forall\, \lambda > \lambda_2, \tag{3.63}$$

where $E(t)$ is defined by (2.12).

Fix $S_0 \in (T_0, T/2)$ and $S_0' \in (T/2, T_0')$; then it is easy to check that

$$\int_{S_0}^{S_0'} E(t)dt \leq C(1+h)\int_{T_0}^{T_0'}\int_\Omega w^2 dxdt. \tag{3.64}$$

Thus, by (3.63)–(3.64) and (3.59), one can find a constant $\lambda_3 = \lambda_3(R_0) > 0$ such that

$$\int_{S_0}^{S_0'} E(t)dt \leq C\left\{e^{C\lambda}\left|\frac{\partial w}{\partial\nu}\right|^2_{H^{-1}(\Sigma_0)} + e^{-R_0^2\lambda/8}\int_0^T E(t)dt\right\} \qquad \forall\, \lambda > \lambda_2 + \lambda_3. \tag{3.65}$$

Finally, by (2.13) (in Lemma 2.3) and (3.65), one gets

$$E(0) \leq C_2\left\{e^{C_2\lambda}\left|\frac{\partial w}{\partial\nu}\right|^2_{H^{-1}(\Sigma_0)} + e^{-R_0^2\lambda/8 + C_2\sqrt{h}}E(0)\right\} \qquad \forall\, \lambda > \lambda_2 + \lambda_3, \tag{3.66}$$

where $C_2 = C_2(T, \Omega)$ is a positive constant. However, it is easy to find a constant $\lambda_4 = \lambda_4(R_0, C_2)$ such that

$$C_2 e^{-R_0^2\lambda_4/8 + C_2\sqrt{h}} \leq 1/2. \tag{3.67}$$

Thus, by (3.66)–(3.67), we see that for any $\lambda > \max(\lambda_2 + \lambda_3, \lambda_4)$ it holds that

$$E(0) \leq Ce^{C\lambda}\left|\frac{\partial w}{\partial\nu}\right|^2_{H^{-1}(\Sigma_0)}. \tag{3.68}$$

Equation (3.68) is exactly the desired result. Thus, the proof of Theorem 3.2 is completed. □

4. Exact controllability of the linear and semilinear wave equations. In this section, we apply our observability estimates (3.1) and (3.3) to exact controllability for wave equations. First of all, let us consider the exact controllability of system (1.4). We have the following result.

THEOREM 4.1. *Let* (H) *hold,* $T > 2\max_{x\in\Omega}|x-x_0|$. *Let* $p_1 \in L^{n+1}(Q)$, $p_2 \in C^1(\overline{Q})$, *and* $p_3 \in C^1(\overline{Q};\mathbb{R}^n)$. *Then for any given* (y_0,y_1), $(z_0,z_1) \in L^2(\Omega)\times H^{-1}(\Omega)$, *there is a control* $u \in L^2(\Sigma_0)$ *such that the weak solution* y *of* (1.4) *satisfies* (1.5). *Furthermore, concerning the control* u, *we have the following estimate:*

$$|u|_{L^2(\Sigma_0)} \le \mathcal{C}(r_1)(|y_0|_{L^2(\Omega)} + |y_1|_{H^{-1}(\Omega)} + |z_0|_{L^2(\Omega)} + |z_1|_{H^{-1}(\Omega)}), \tag{4.1}$$

where $\mathcal{C}(r_1)$ *is given by* (3.2) *with* r *replaced by* $r_1 \overset{\triangle}{=} |p_1|_{n+1} + |p_2|_{1,\infty} + |p_3|_{1,\infty}$.

Proof. Let us use Lions's Hilbert uniqueness method (see [7, 16, 17, 30]). First, we solve

$$\begin{cases} v_{tt} - \Delta v = p_1 v + p_2 v_t + \langle p_3, \nabla v\rangle & \text{in } Q,\\ v = 0 & \text{on } \Sigma,\\ v(T) = z_0, \quad v_t(T) = z_1 & \text{in } \Omega. \end{cases} \tag{4.2}$$

Next, for any $(\varphi_0,\varphi_1) \in X \overset{\triangle}{=} H_0^1(\Omega)\times L^2(\Omega)$, we solve

$$\begin{cases} \varphi_{tt} - \Delta\varphi = [p_1 - (p_2)_t - \nabla\cdot p_3]\varphi - p_2\varphi_t - \langle p_3, \nabla\varphi\rangle & \text{in } Q,\\ \varphi = 0 & \text{on } \Sigma,\\ \varphi(0) = \varphi_0, \quad \varphi_t(0) = \varphi_1 & \text{in } \Omega \end{cases} \tag{4.3}$$

and

$$\begin{cases} \eta_{tt} - \Delta\eta = p_1\eta + p_2\eta_t + \langle p_3, \nabla\eta\rangle & \text{in } Q,\\ \eta = \big(\partial\varphi/\partial\nu\big)\chi_{\Sigma_0}(t,x) & \text{on } \Sigma,\\ \eta(T) = 0, \quad \eta_t(T) = 0 & \text{in } \Omega. \end{cases} \tag{4.4}$$

Then, we define a linear and continuous operator $\Lambda : X \to X'(\equiv H^{-1}(\Omega)\times L^2(\Omega))$ by

$$\Lambda(\varphi_0,\varphi_1) = \big(p_2(0)\eta(0) - \eta_t(0), \eta(0)\big), \tag{4.5}$$

where $\eta \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))$ is the weak solution of (4.4). It is sufficient to prove the existence of some $(\varphi_0,\varphi_1) \in X$ such that

$$\Lambda(\varphi_0,\varphi_1) = \Big(p_2(0)(y_0 - v(0)) - y_1 + v_t(0),\, y_0 - v(0)\Big), \tag{4.6}$$

where $v \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))$ is the weak solution of (4.2). In order to solve (4.6), we observe that (by (4.3)–(4.4))

$$\langle \Lambda(\varphi_0,\varphi_1), (\varphi_0,\varphi_1)\rangle_{X',X} = \int_{\Sigma_0} \left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\Sigma_0. \tag{4.7}$$

However, by Theorem 3.1 and (4.7), we have

$$|(\varphi_0,\varphi_1)|_X^2 \le \mathcal{C}(r_1)\langle \Lambda(\varphi_0,\varphi_1), (\varphi_0,\varphi_1)\rangle_{X',X} \qquad \forall\, (\varphi_0,\varphi_1) \in X\,, \tag{4.8}$$

where $\mathcal{C}(r_1)$ is given by (3.2) with r replaced by $r_1 \overset{\triangle}{=} |p_1|_{n+1} + |q_2|_{1,\infty} + |q_3|_{1,\infty}$. Therefore $\Lambda : X \to X'$ is an isomorphism. Thus (4.6) admits a unique solution $(\varphi_0, \varphi_1) \in X$ and

$$u = \partial\varphi/\partial\nu \tag{4.9}$$

is the desired control such that the weak solution of (1.4) satisfies (1.5).

Now, let us prove (4.1). Concerning (4.2), by Lemma 2.3, we obtain that

$$|(v_t(0), v(0))|_{X'} \leq Ce^{Cr_1}|(z_1, z_0)|_{X'}. \tag{4.10}$$

Consequently, by (4.6)–(4.9) and (4.10), we get

$$\begin{aligned} |u|_{L^2(\Sigma_0)} &\leq \mathcal{C}(r_1)\big|\big(p_2(0)(y_0 - v(0)) - y_1 + v_t(0),\, y_0 - v(0)\big)\big|_{X'} \\ &\leq \mathcal{C}(r_1)(|(y_1, y_0)|_{X'} + |(z_1, z_0)|_{X'}), \end{aligned} \tag{4.11}$$

which proves (4.1). □

Next, let us consider the exact controllability of the following semilinear wave equation:

$$\begin{cases} y_{tt} - \Delta y = f(y) & \text{in } Q, \\ y = u\chi_{\Sigma_0}(t, x) & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } \Omega. \end{cases} \tag{4.12}$$

We have the following result.

Theorem 4.2. *Let Γ_0 satisfy* (H). *Let $f(\cdot) : \mathbb{R}^1 \to \mathbb{R}^1$ be of class $C^1(\mathbb{R})$ with $f'(\cdot) \in L^\infty(\mathbb{R})$ and $T > 2\max_{x\in\Omega}|x - x_0|$. Then, for any $s \in (0, 1)$, the semilinear wave equation* (4.12) *is exactly controllable in $H_0^s(\Omega) \times H^{s-1}(\Omega)$ at time T with control u in $H^s(0, T; L^2(\Gamma_0))$.*

Proof. By [31, Remark 2.3] and by our Theorem 3.2 (which implies a UCP for the wave equations), we obtain Theorem 4.2 immediately. □

Appendix A. Proof of Lemma 2.2.

This appendix is devoted to giving a proof of Lemma 2.2. For this purpose, we need the following known result (see [17, p. 46]).

Lemma A.1. *Let $T > 0$ and $F = 0$. Suppose that*

$$(g, w_0, w_1) \in L^2(\Sigma) \times L^2(\Omega) \times H^{-1}(\Omega). \tag{A.1}$$

Then the unique weak solution w of (2.1) *satisfies*

$$w \in C([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-1}(\Omega)). \tag{A.2}$$

Furthermore, there is a constant $C = C(T, \Omega) > 0$ such that

$$\begin{aligned} |w|_{C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))} \leq &C\big(|g|_{L^2(\Sigma)} + |w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)}\big) \\ &\forall\, (g, w_0, w_1) \in L^2(\Sigma) \times L^2(\Omega) \times H^{-1}(\Omega). \end{aligned} \tag{A.3}$$

Now, let us prove Lemma 2.2. We divide the proof into two steps.

Step 1. We decompose the solution w of (2.1) as

$$w = \xi + \eta, \tag{A.4}$$

with ξ and η, respectively, solutions of

$$\text{(A.5)}\qquad \begin{cases} \xi_{tt} - \Delta\xi = 0 & \text{in } Q, \\ \xi = g & \text{on } \Sigma, \\ \xi(0) = w_0, \quad \xi_t(0) = w_1 & \text{in } \Omega \end{cases}$$

and

$$\text{(A.6)}\qquad \begin{cases} \eta_{tt} - \Delta\eta = F & \text{in } Q, \\ \eta = 0 & \text{on } \Sigma, \\ \eta(0) = 0, \quad \eta_t(0) = 0 & \text{in } \Omega. \end{cases}$$

First of all, by Lemma A.1, we see that $\xi \in C([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))$; furthermore, there is a constant $C = C(T, \Omega) > 0$ such that

$$\text{(A.7)}\qquad |\xi|_{C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))} \leq C\big(|g|_{L^2(\Sigma)} + |w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)}\big).$$

Next, denote

$$\text{(A.8)}\qquad \begin{cases} \mathcal{A} \overset{\triangle}{=} -\Delta, \\ D(\mathcal{A}) = H^2(\Omega) \cap H_0^1(\Omega). \end{cases}$$

Put

$$\text{(A.9)}\qquad Z \overset{\triangle}{=} \mathcal{A}^{-1/2}\eta.$$

Then, by (A.6) and (A.9), we see that

$$\text{(A.10)}\qquad \begin{cases} Z_{tt} - \Delta Z = \mathcal{A}^{-1/2}F & \text{in } Q, \\ Z = 0 & \text{on } \Sigma, \\ Z(0) = 0, \quad Z_t(0) = 0 & \text{in } \Omega. \end{cases}$$

However, by (2.5), we have $\mathcal{A}^{-1/2}F \in L^1(0,T; L^2(\Omega))$. Thus, by (A.9)–(A.10) and Lemma 2.1, we see that $\eta \in C([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))$; furthermore, there is a constant $C = C(T, \Omega) > 0$ such that

$$\text{(A.11)}\qquad |\eta|_{C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))} \leq C|F|_{L^1(0,T;H^{-1}(\Omega))}.$$

Combining (A.4), (A.7), and (A.11), we see that $w \in C([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))$; furthermore, there is a constant $C = C(T, \Omega) > 0$ such that

$$\text{(A.12)}\qquad \begin{aligned} &|w|_{C([0,T];L^2(\Omega))\cap C^1([0,T];H^{-1}(\Omega))} \\ &\leq C\big(|F|_{L^1(0,T;H^{-1}(\Omega))} + |g|_{L^2(\Sigma)} + |w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)}\big). \end{aligned}$$

Step 2. It remains to prove $\partial w/\partial\nu \in H^{-1}(\Sigma)$ and

$$\text{(A.13)}\qquad \Big|\frac{\partial w}{\partial \nu}\Big|_{H^{-1}(\Sigma)} \leq C\Big(|F|_{L^1(0,T;H^{-1}(\Omega))} + |g|_{L^2(\Sigma)} + |w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)}\Big).$$

This is essentially known (see [9, Theorem 2.3]), but we give the proof for the readers' convenience. Let W be the solution of

$$\text{(A.14)}\qquad \begin{cases} W_{tt} - \Delta W = 0 & \text{in } Q, \\ W = h & \text{on } \Sigma, \\ W(T) = 0, \quad W_t(T) = 0 & \text{in } \Omega, \end{cases}$$

where h is given such that

$$h \in H^1(\Sigma), \quad h(0) = h(T) = 0 \text{ on } \Gamma. \tag{A.15}$$

Then, by Lemma 2.1, we obtain

$$|W|_{C([0,T];H^1(\Omega))\cap C^1([0,T];L^2(\Omega))} + \left|\frac{\partial W}{\partial \nu}\right|_{L^2(\Sigma)} \leq C|h|_{H^1(\Sigma)} \tag{A.16}$$

for some constant $C = C(T, \Omega) > 0$. Assuming all data are smooth (we then extend by continuity), multiplying the first equation of (A.14) by w, integrating it on Q, and using integration by parts, by (A.14) and (2.1), we get

$$\int_\Sigma \frac{\partial w}{\partial \nu} h d\Sigma = -(F, W)_Q + \int_\Sigma g \frac{\partial W}{\partial \nu} d\Sigma - (w_1, W(0))_\Omega + (w_0, W_t(0))_\Omega. \tag{A.17}$$

Now, by (A.16)–(A.17), we see that there is a constant $C = C(T, \Omega) > 0$ such that

$$\left|\int_\Sigma \frac{\partial w}{\partial \nu} h d\Sigma\right| \leq C|h|_{H^1(\Sigma)}\big(|F|_{L^1(0,T;H^{-1}(\Omega))} + |g|_{L^2(\Sigma)} + |w_0|_{L^2(\Omega)} + |w_1|_{H^{-1}(\Omega)}\big), \tag{A.18}$$

which implies (A.13) immediately. □

Acknowledgments. The author heartily thanks Professor J. E. Lagnese (Georgetown University), Professors I. Lasiecka and R. Triggiani (University of Virginia), Professor M. Yamamoto (The University of Tokyo), and Professor E. Zuazua (Universidad Complutense) for fruitful discussions and useful help. This paper is an improved version of one chapter of the author's Ph.D. thesis [27] accomplished at Fudan University under the guidance of Professors X. Li and J. Yong. The author would like to take this opportunity to thank them deeply for their encouragement and suggestions. Also, the author gratefully acknowledges the anonymous referees for their criticism and suggestions which led to this improved version.

REFERENCES

[1] S. Alinhac and M. S. Baouendi, *A nonuniqueness result for operators of principal type*, Math. Z., 220 (1995), pp. 561–568.

[2] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Series 34, Research Institute of Mathematics, Seoul National University, Seoul, Korea, 1994.

[4] L. F. Ho, *Observabilité frontiére de l'équation des ondes*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 443–446.

[5] L. Hörmander, *The Analysis of Linear Partial Differential Operators.* III. *Pseudo-differential Operators*, Springer-Verlag, Berlin, 1985.

[6] M. A. Kazemi and M. V. Klibanov, *Stability estimates for ill-posed Cauchy problems involving hyperbolic equations and inequalities*, Appl. Anal., 50 (1993), pp. 93–102.

[7] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, John Wiley, Chichester, UK, Masson, Paris, 1995.

[8] J. Lagnese, *The Hilbert uniqueness method: A retrospective*, in Optimal Control of Partial Differential Equations, Lecture Notes in Control and Inform. Sci. 149, Springer-Verlag, New York, 1990, pp. 158–181.

[9] I. Lasiecka, J.-L. Lions, and R. Triggiani, *Nonhomogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 69 (1986), pp. 149–192.

[10] I. Lasiecka and R. Triggiani, *Regularity of hyperbolic equations under $L_2(0,T;L_2(\Gamma))$-boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275–286.

[11] I. Lasiecka and R. Triggiani, *Carleman estimates and exact boundary controllability for a system of coupled, nonconservative second-order hyperbolic equations*, in Partial Differential Equations Methods in Control and Shape Analysis, Lecture Notes in Pure and Appl. Math. 188, Marcel Dekker, New York, 1997, pp. 215–243.

[12] I. Lasiecka, R. Triggiani, and P. F. Yao, *Exact controllability for second-order hyperbolic equations with variable coefficient-principle part and first-order terms*, Nonlinear Anal., 30 (1997), pp. 111–122.

[13] I. Lasiecka, R. Triggiani, and X. Zhang, *Nonconservative wave equations with purely Neumann B.C.: Global uniqueness and observability in one shot*, in Contemp. Math., to appear.

[14] M. M. Lavrentév, V. G. Romanov, and S. P. Shishataskii, *Ill-Posed Problems of Mathematics Physics and Analysis*, Transl. Math. Monogr. 64, AMS, Providence, RI, 1986.

[15] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Boston, MA, 1995.

[16] J.-L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[17] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabalisation de systémes distribués, tome* 1, Rech. Math. Appl. 8, Masson, Paris, 1988.

[18] K. Liu, *Locally distributed control and damping for the conservative systems*, SIAM J. Control Optim, 35 (1997), pp. 1574–1590.

[19] A. López, X. Zhang, and E. Zuazua, *Null controllability of the heat equation as singular limit of the exact controllability of dissipative wave equations*, J. Math. Pures Appl., to appear.

[20] A. Ruiz, *Unique continuation for weak solutions of the wave equation plus a potential*, J. Math. Pures Appl., 71 (1992), pp. 455–467.

[21] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[22] D. Tataru, *Boundary controllability for conservative PDEs*, Appl. Math. Optim., 31 (1995), pp. 257–295.

[23] D. Tataru, *Boundary observability and controllability for evolution governed by higher order PDE*, J. Math. Anal. Appl., 193 (1995), pp. 632–658.

[24] D. Tataru, *Unique continuation for solutions to PDE's; between Hörmander's theorem and Holmgren's theorem*, Comm. Partial Differential Equations, 20 (1995), pp. 855–884.

[25] D. Tataru, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl., 75 (1996), pp. 367–408.

[26] P. F. Yao, *On the observability inequalities for exact controllability of wave equations with variable coefficients*, SIAM J. Control Optim., 37 (1999), pp. 1568–1599.

[27] X. Zhang, *Exact Controllability of the Semilinear Distributed Parameter System and Some Related Problems*, Ph.D. thesis, Fudan University, Shanghai, People's Republic of China, 1998.

[28] X. Zhang, *Explicit observability estimate for the wave equation with potential and its application*, Proc. Roy. Soc. London Sect. A: Mathematical, Physical and Engineering Sciences, 456 (2000), pp. 1101–1115.

[29] X. Zhang, *Exact controllability of the semilinear plate equations*, Asymptot. Anal., submitted.

[30] E. Zuazua, *An Introduction to Exact Controllability for Distributed Systems*, Textos e Notas 44, Centrode de Matemática e Aplicações Fundamêntais, Universidades de Lisboa, Lisbon, Portugal, 1990.

[31] E. Zuazua, *Exact boundary controllability for the semilinear wave equation*, in Nonlinear Partial Differential Equations and Their Applications, Vol. 10, H. Brezis and J.-L. Lions, eds., Longman, Harlow, UK, 1991, pp. 357–391.

[32] E. Zuazua, *Exact controllability for semilinear wave equations in one space dimension*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 109–129.

[33] E. Zuazua, *Some problems and results on the controllability of partial differential equations*, European Congress of Mathematics, Vol. II, Progr. Math. 169, Birkhäuser-Verlag, Basel, Switzerland, 1998, pp. 276–311.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 835–851

OPTIMAL CONTROL OF A CLASS OF LINEAR HYBRID SYSTEMS WITH SATURATION*

BART DE SCHUTTER†

Abstract. We consider a class of first order linear hybrid systems with saturation. A system that belongs to this class can operate in several modes or phases; in each phase each state variable of the system exhibits a linear growth until a specified upper or lower saturation level is reached, and after that the state variable stays at that saturation level until the end of the phase. A typical example of such a system is a traffic signal controlled intersection. We develop methods to determine optimal switching time sequences for first order linear hybrid systems with saturation that minimize criteria such as average queue length, worst case queue length, average waiting time, and so on. First we show how the extended linear complementarity problem (ELCP), which is a mathematical programming problem, can be used to describe the set of system trajectories of a first order linear hybrid system with saturation. Optimization over the solution set of the ELCP then yields an optimal switching time sequence. Although this method yields globally optimal switching time sequences, it is not feasible in practice due to its computational complexity. Therefore, we also present some methods to compute suboptimal switching time sequences. Furthermore, we show that if there is no upper saturation, then for some objective functions the globally optimal switching time sequence can be computed very efficiently. We also discuss some approximations that lead to suboptimal switching time sequences that can be computed very efficiently. Finally, we use these results to design optimal switching time sequences for traffic signal controlled intersections.

Key words. hybrid systems, control, nonlinear optimization, extended linear complementarity problem

AMS subject classifications. 93C10, 49N99, 90C33, 90B22

PII. S0363012999354648

1. Introduction.

Hybrid systems arise from the interaction between continuous variable systems[1] and discrete event systems.[2] In general we could say that a hybrid system can be in one of several modes whereby in each mode the behavior of the system can be described by a system of difference or differential equations, and that the system switches from one mode to another due to the occurrence of an event. There are many frameworks to model, analyze, and control hybrid systems (see, e.g., [1, 11, 12, 17] and the references cited therein). An important trade-off in this context is that of modeling power versus decision power: the more accurate the model is the less we can analytically say about its properties. Furthermore, many analysis and control problems lead to computationally hard problems for even the most elementary hybrid systems [2]. Therefore, we focus on a specific class of hybrid systems that can be analyzed using a mathematical programming problem that is called the extended linear complementarity problem (ECLP). More specifically, we study the design of optimal switching time sequences for a class of first order linear hybrid systems subject to saturation.

*Received by the editors April 5, 1999; accepted for publication (in revised form) June 14, 2000; published electronically October 11, 2000.
http://www.siam.org/journals/sicon/39-3/35464.html

†Control Laboratory, Faculty of Information Technology and Systems, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands (b.deschutter@its.tudelft.nl).

[1] Continuous variable systems are systems that can be described by a difference or differential equation.

[2] Discrete event systems are asynchronous systems where the state transitions are initiated by events; in general the time instants at which these events occur are not equidistant.

This work is an extension of the work reported in [8] in which we developed some algorithms to design optimal traffic signal switching schemes for single intersections. In [8] we considered only fixed amber durations and we could efficiently compute only suboptimal switching schemes for *approximations* of the real objective functions. Now we allow variable durations for the amber phases, and we show that if there is no upper saturation, then for certain objective functions the optimal switching scheme can be computed very efficiently without making any approximations. Furthermore, in this paper we also consider a more general class of systems than the traffic signal controlled intersections of [8].

The work reported here is closely related to optimal traffic signal control (see, e.g., [10, 14, 15, 16]). The main difference between the model presented in this paper applied to traffic signal optimization and the models used by most other researchers is that in our approach the length of the green-amber-red cycles may vary from cycle to cycle, i.e., we optimize over a fixed number of switch-overs instead of over a fixed number of time steps. This allows us to optimize not only the split but also the cycle time with continuous optimization variables (usually the optimization of split and cycle time is performed using boolean variables at each time step, each variable corresponding to the decision of switching or not the traffic signals as in UTOPIA, OPAC, SCOOT, or SCATS). Our method adds an extra degree of freedom, which will in general lead to a more optimal switching scheme.

This paper is organized as follows. In section 2 we discuss model predictive control, which is the framework in which our approach can be embedded. Next we give the definition and a brief description of the ELCP. In section 3 we introduce a class of first order linear hybrid systems with saturation. We show that computing the optimal switching time instants in general leads to a nonconvex optimization problem or to an optimization problem over the solution set of an extended linear complementarity problem. In section 4 we show that if there is no upper saturation, then for some objective functions the feasible set of the optimal switching problem can be replaced by a convex set without changing the optimum. In that case the optimal switching time sequence can be computed very efficiently. Furthermore, by making some approximations the problem becomes a linear programming problem. These results will be illustrated in section 5 in which we compute optimal traffic signal switching time sequences for traffic signal controlled intersections.

2. Preliminaries.

2.1. Notation. Let a and b be vectors with n components. The ith component of a is denoted by a_i or $(a)_i$. We use $a \geqslant b$ to indicate that $a_i \geqslant b_i$ for all i. The maximum operator on vectors is defined as follows: $\big(\max(a,b)\big)_i = \max(a_i, b_i)$ for all i. The minimum operator on vectors is defined analogously. The zero vector with n components is denoted by 0_n, or by 0 if the dimension is clear from the context. The n by n identity matrix is denoted by I_n, or by I if the dimension is clear from the context. The set of the real numbers is denoted by $\mathbb{R}$.

2.2. Model predictive control. Model predictive control (MPC) [3, 4, 9] is a very popular controller design method in the process industry. An important advantage of MPC is that it allows the inclusion of constraints on the inputs and outputs, and that it can handle changes in the system parameters by using a moving horizon approach, in which the model and the control strategy are continuously updated. We will use the MPC framework to design optimal switching schemes for a class of hybrid systems. In general the resulting optimization problem is nonlinear and nonconvex.

However, if the control objective and the constraints depend monotonically on the outputs of the system, the MPC problem can be recast as problem with a convex feasible set. As a consequence, the problem can be solved very efficiently so that on-line computation is feasible.

In each step of the conventional MPC algorithm for discrete-time systems an optimal input sequence is computed that minimizes a given cost criterion over a given prediction horizon N_{p}. Furthermore, for the optimization the control input u is taken to be constant from a certain point on: $u(k+j) = u(k+N_{\mathrm{c}}-1)$ for $j = N_{\mathrm{c}}, N_{\mathrm{c}}+1, \ldots, N_{\mathrm{p}}-1$, where N_{c} is the control horizon and where k is the first sampling index of the period under consideration. MPC uses a receding horizon principle: after computation of the optimal control sequence $u(k), u(k+1), \ldots, u(k+N_{\mathrm{c}}-1)$, only the first control input sample $u(k)$ will be implemented; subsequently the horizon is shifted one sample, the estimates of the state and the parameters of the system are updated using information coming from new measurements, and the optimization is restarted. Note that the continuous updating of the model and of the estimates of the states also introduces a kind of feedback in the control system. In general feedback is necessary to obtain good performance and tracking in most control applications (see, e.g., [13] for applications of feedback control in traffic).

The parameters N_{p} and N_{c} are the basic tuning parameters of the MPC algorithm:

(i) In general the prediction horizon N_{p} is selected such that the time interval $[k, k+N_{\mathrm{p}}-1]$ contains the crucial dynamics of the process.

(ii) An important effect of a small control horizon N_{c} is the smoothing of the control signal (because of the emphasis on the average behavior rather than on aggressive noise reduction). The control horizon forces the control signal to a constant value. This also has a stabilizing effect since the output signal is forced to its steady-state value. Another important consequence of decreasing N_{c} is the reduction of the number of optimization variables, which results in a decrease of the computational effort.

2.3. The ELCP. The ELCP is a mathematical programming problem which is defined as follows [7]:

Given $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{q\times n}$, $c \in \mathbb{R}^p$, $d \in \mathbb{R}^q$ and m subsets ϕ_1, ϕ_2, $\ldots$, ϕ_m of $\{1, 2, \ldots, p\}$, find $x \in \mathbb{R}^n$ such that

$$\sum_{j=1}^{m} \prod_{i\in\phi_j} (Ax-c)_i = 0 \tag{1}$$

subject to $Ax \geqslant c$ and $Bx = d$, or show that no such x exists.

The ELCP can be considered as a system of linear equations and inequalities ($Ax \geqslant c$, $Bx = d$), where there are m groups of linear inequalities (one group for each index set ϕ_j) such that in each group at least one inequality should hold with equality. In [7] we have developed an algorithm to compute the complete solution set of an ELCP. In general this solution set consists of the union of a subset of faces of the polyhedron $\mathcal{P}$ defined by the system $Ax \geqslant c$, $Bx = d$ (i.e., the solution set contains all the points of $\mathcal{P}$ that satisfy condition (1)). Our ELCP algorithm yields a compact representation of the solution set of an ELCP by vertices, extreme rays, and a basis of the linear subspace corresponding to the largest affine subspace of the solution set. In [7] we have also shown that the general ELCP is NP-hard.

In the next section we shall show that the ELCP can be used to determine optimal switching time instants for a special class of hybrid systems.

3. Optimal switching time sequences for a class of linear hybrid systems with saturation.

3.1. First order linear hybrid systems with saturation. Consider a system the evolution of which is characterized by consecutive phases. In each phase each state variable of the system exhibits a linear growth or decrease until a certain upper or lower saturation level is reached; then the state variable stays constant until the end of the phase. A system, the behavior of which satisfies this description, will be called a *first order linear hybrid system with saturation.*

A typical example of a first order linear hybrid system with saturation is a traffic signal controlled intersection, provided that we use a continuous approximation for the queue lengths (see section 5 and [8]). The state variables of this system correspond to the queue lengths in the different lanes. For a traffic signal controlled intersection the lower bound for the queue length is equal to 0. The upper bound could correspond to the maximal available storage space due to the distance to the preceding junction or due to the layout of the intersection. We assume that if this upper bound is reached then newly arriving cars take another route to get to their destination. Another example of a first order linear hybrid system with saturation is a system consisting of several fluid containers that are connected by tubes with valves and that have two outlets—one at the bottom (with a tube that leads to another fluid container) and one at the top (so that the fluid level in the containers can never exceed a given level)—provided that we assume that the increase or decrease of the fluid levels is linear if the system is not saturated.

Now we derive the equations that describe the evolution of the state variables in a first order linear hybrid system with saturation. Analogous to a traffic signal controlled intersection, we will use the phrase "queue lengths" to refer to the state variables of the system. Note, however, that our definition of a first order linear hybrid system with saturation is not limited to queuing systems only.

Let M be the number of "queues." The length of queue i at time t is denoted by $q_i(t)$. Let $\alpha_{i,k}$, $b^{\mathrm{ls}}_{i,k}$, and $b^{\mathrm{us}}_{i,k}$ be, respectively, the queue length growth rate for queue i in phase k, the lower saturation bound for the queue length q_i in phase k, and the upper saturation bound for the queue length q_i in phase k. The evolution of the system begins at time t_0. Let $t_1, t_2, t_3, \ldots$ be the switching time instants, i.e., the time instants at which the system switches from one phase to another. Note that in general the sequence $t_0, t_1, t_2, \ldots$ is *not* an equidistant sequence. The length of the kth phase is equal to $\delta_k \overset{\mathrm{def}}{=} t_{k+1} - t_k$. Note that $\delta_k > 0$ for all k. We assume that $0 \leqslant b^{\mathrm{ls}}_{i,k+1} \leqslant q_i(t_{k+1}) \leqslant b^{\mathrm{us}}_{i,k+1}$ for all i, k such that the queue lengths are always nonnegative and such that there are no sudden jumps in the queue lengths due to a change in the saturation level at one of the switching time instants. For queue i we have

$$\frac{dq_i(t)}{dt} = \begin{cases} \alpha_{i,k} & \text{if } b^{\mathrm{ls}}_{i,k} < q_i(t) < b^{\mathrm{us}}_{i,k}, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

for $t \in (t_k, t_{k+1})$. This implies that the evolution of the queue lengths at the switching time instants is given by

$$q_i(t_{k+1}) = \max\big(\min(q_i(t_k) + \alpha_{i,k}\delta_k,\ b^{\mathrm{us}}_{i,k}),\ b^{\mathrm{ls}}_{i,k}\big) \tag{3}$$

for $k = 0, 1, 2, \ldots$. If we define $q_{i,k} = q_i(t_k)$ and

$$q_k = \begin{bmatrix} q_{1,k} \\ q_{2,k} \\ \vdots \\ q_{M,k} \end{bmatrix}, \quad \alpha_k = \begin{bmatrix} \alpha_{1,k} \\ \alpha_{2,k} \\ \vdots \\ \alpha_{M,k} \end{bmatrix}, \quad b_k^{\mathrm{ls}} = \begin{bmatrix} b_{1,k}^{\mathrm{ls}} \\ b_{2,k}^{\mathrm{ls}} \\ \vdots \\ b_{M,k}^{\mathrm{ls}} \end{bmatrix}, \quad b_k^{\mathrm{us}} = \begin{bmatrix} b_{1,k}^{\mathrm{us}} \\ b_{2,k}^{\mathrm{us}} \\ \vdots \\ b_{M,k}^{\mathrm{us}} \end{bmatrix},$$

we obtain the vector equation

$$q_{k+1} = \max\bigl(\min(q_k + \alpha_k \delta_k,\, b_k^{\mathrm{us}}),\, b_k^{\mathrm{ls}}\bigr). \tag{4}$$

If we introduce dummy vectors z_k, then (3) can be rewritten as

$$z_{k+1} = \min(q_k + \alpha_k \delta_k, b_k^{\mathrm{us}}), \tag{5}$$

$$q_{k+1} = \max(z_{k+1}, b_k^{\mathrm{ls}}). \tag{6}$$

3.2. Optimal switching time sequences for linear hybrid systems with saturation. Now we consider the problem of computing an optimal (finite) sequence of switching time instants for a system described by a system of equations of the form (4) using an MPC approach.

We may assume without loss of generality that t_0 will be the first switching time instant in each step of the MPC algorithm. Note that this implies that switching time instant t_1 of the current MPC step will correspond to switching time instant t_0 of the next MPC step. The queue length vector $q_0 = q(t_0)$ at time $t = t_0$ can be measured[3] or estimated. Now we want to determine the optimal switching time sequence $t_0, t_1, \ldots, t_{N_{\mathrm{p}}}$ for a given performance criterion J. For the class of systems we consider it makes more sense to replace the condition that the control input is constant after the control horizon by the condition

$$\delta_k = \delta_{k-K_{\mathrm{c}}} \qquad \text{for } k = N_{\mathrm{c}}, N_{\mathrm{c}}+1, \ldots, N_{\mathrm{p}}-1, \tag{7}$$

where K_{c} is the number of switching phases in one larger cycle of the system (e.g., in traffic signal control for an intersection of two streets K_{c} could be equal to 4 corresponding to the combinations red-green, red-amber, green-red, amber-red for the traffic signals on the crossing roads (see also section 5)). Possible performance criteria are

1. (weighted) average queue length over all queues:

$$J_1 = \sum_{i=1}^{M} w_i \, \frac{1}{t_{N_{\mathrm{p}}} - t_0} \int_{t_0}^{t_{N_{\mathrm{p}}}} q_i(t)\, dt, \tag{8}$$

2. (weighted) average queue length over the worst queue:

$$J_2 = \max_i \left(w_i \, \frac{1}{t_{N_{\mathrm{p}}} - t_0} \int_{t_0}^{t_{N_{\mathrm{p}}}} q_i(t)\, dt \right), \tag{9}$$

3. (weighted) worst case queue length:

$$J_3 = \max_{i,\, t} \bigl(w_i \, q_i(t) \bigr), \tag{10}$$

[3] Note that if we compute the switching time sequence fast enough (i.e., if the computation time is less than $\delta_0 = t_1 - t_0$), we can wait to compute the optimal sequence until after t_0.

4. (weighted) average "waiting time" over all queues:[4]

$$J_4 = \sum_{i=1}^{M} w_i \, \frac{\displaystyle\int_{t_0}^{t_{N_\mathrm{p}}} q_i(t)\,dt}{\displaystyle\sum_{k=0}^{N_\mathrm{p}-1} \alpha_{i,k}^{\mathrm{a}} \delta_k}, \tag{11}$$

5. (weighted) average "waiting time" over the worst queue:

$$J_5 = \max_i \left(w_i \, \frac{\displaystyle\int_{t_0}^{t_{N_\mathrm{p}}} q_i(t)\,dt}{\displaystyle\sum_{k=0}^{N_\mathrm{p}-1} \alpha_{i,k}^{\mathrm{a}} \delta_k} \right), \tag{12}$$

where $w_i > 0$ for all i and $\alpha_{i,k}^{\mathrm{a}}$ is the arrival rate of "customers" for queue i in phase k.

We can impose extra conditions such as minimum or maximum queue lengths (which could be useful in order to prevent saturation at the lower or upper level for some queues), minimum and maximum durations for the switching time intervals, and so on.

This leads to the following optimization problem that should be solved in each MPC step:

$$\operatorname*{minimize}_{\delta_0,\delta_1,\ldots,\delta_{N_\mathrm{c}-1}} \; J \tag{13}$$

subject to

$$\delta_k = \delta_{k-K_\mathrm{c}} \qquad \text{for } k = N_\mathrm{c}, N_\mathrm{c}+1, \ldots, N_\mathrm{p}-1, \tag{14}$$

$$\delta_{\min,k} \leqslant \delta_k \leqslant \delta_{\max,k} \qquad \text{for } k = 0,1,\ldots,N_\mathrm{c}-1, \tag{15}$$

$$q_{\min,k} \leqslant q_{k+1} \leqslant q_{\max,k} \qquad \text{for} k = 0,1,\ldots,N_\mathrm{p}-1, \tag{16}$$

$$z_{k+1} = \min(q_k + \alpha_k \delta_k,\, b_k^{\mathrm{us}}) \qquad \text{for } k = 0,1,\ldots,N_\mathrm{p}-1, \tag{17}$$

$$q_{k+1} = \max(z_{k+1},\, b_k^{\mathrm{ls}}) \qquad \text{for } k = 0,1,\ldots,N_\mathrm{p}-1 \tag{18}$$

with $q_0 = q(t_0)$, and where $\delta_{\min,k}$ and $\delta_{\max,k}$ are, respectively, the minimum and the maximum values of δ_k, and $(q_{\min,k})_i$ and $(q_{\max,k})_i$ are, respectively, the minimum and the maximum queue lengths for queue i at time instant t_{k+1}.

Remark 3.1. We can also use a first order linear hybrid system with saturation as an approximate model if we have a hybrid system with saturation in which the queue length growth or decrease rates are slowly time-varying since in MPC we use a moving horizon approach in which the model of the system and the estimate of the initial condition can be updated at the beginning of each control cycle. This also introduces a feedback into the control system.

[4] The average waiting time is equal to the total waiting time divided by the number of arrivals. If the initial and final queue lengths are 0, then the average waiting time for queue i is given by the fraction in the expression on the right-hand side of (11). So J_4 is in fact an approximation of the (weighted) average waiting time.

3.3. The ELCP and optimal switching time sequences. Now we show that the system (14)–(18) can be reformulated as an ELCP.

Consider (17) for an arbitrary index k. This equation can be rewritten as follows:

$$
\begin{gathered}
z_{k+1} \leqslant q_k + \alpha_k \delta_k, \\
z_{k+1} \leqslant b_k^{\mathrm{us}}, \\
(z_{k+1})_i = (q_k + \alpha_k \delta_k)_i \quad \text{or} \quad (z_{k+1})_i = (b_k^{\mathrm{us}})_i \qquad \text{for } i = 1, 2, \ldots, M,
\end{gathered}
$$

or equivalently

$$q_k + \alpha_k \delta_k - z_{k+1} \geqslant 0, \tag{19}$$

$$b_k^{\mathrm{us}} - z_{k+1} \geqslant 0, \tag{20}$$

$$(q_k + \alpha_k \delta_k - z_{k+1})_i \, (b_k^{\mathrm{us}} - z_{k+1})_i = 0 \qquad \text{for } i = 1, 2, \ldots, M. \tag{21}$$

Since a sum of nonnegative numbers is equal to 0 if and only if all the numbers are equal to 0, (21) is equivalent to

$$\sum_{i=1}^{M} (q_k + \alpha_k \delta_k - z_{k+1})_i \, (b_k^{\mathrm{us}} - z_{k+1})_i = (q_k + \alpha_k \delta_k - z_{k+1})^T \, (b_k^{\mathrm{us}} - z_{k+1}) = 0.$$

Hence, (17) can be rewritten as

$$q_k + \alpha_k \delta_k - z_{k+1} \geqslant 0, \tag{22}$$

$$b_k^{\mathrm{us}} - z_{k+1} \geqslant 0, \tag{23}$$

$$(q_k + \alpha_k \delta_k - z_{k+1})^T \, (b_k^{\mathrm{us}} - z_{k+1}) = 0. \tag{24}$$

We can repeat this reasoning for (18) and for each index k. So if we define

$$x_q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_{N_{\mathrm{p}}} \end{bmatrix}, \quad x_z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{N_{\mathrm{p}}} \end{bmatrix}, \quad x_\delta = \begin{bmatrix} \delta_0 \\ \delta_1 \\ \vdots \\ \delta_{N_{\mathrm{c}}-1} \end{bmatrix},$$

and if we replace all δ_k's with index $k \geqslant N_{\mathrm{c}}$ using (14), we finally get a problem of the form

$$\underset{x_\delta}{\text{minimize}} \; J \tag{25}$$

subject to

$$A x_q + B x_z + C x_\delta + d \geqslant 0, \tag{26}$$

$$E x_q + F x_z + g \geqslant 0, \tag{27}$$

$$H x_q + K x_\delta + l \geqslant 0, \tag{28}$$

$$(A x_q + B x_z + C x_\delta + d)^T (E x_q + F x_z + g) = 0 \tag{29}$$

for appropriately defined matrices A, B, C, E, F, H, K and vectors d, g, l. Equations (26), (27), and (29) correspond to (22), (23), and (24), respectively, and the system of linear inequalities (28) contains the conditions (15) and (16). It is easy to verify that the system (26)–(29) is (a special case of) an ELCP.

The time evolution of the queue lengths in a first order linear hybrid system with saturation is given by piecewise-affine functions. The link between piecewise-affine functions and (ordinary) linear complementarity problems has also been explored by several other authors (see, e.g., [5] and the references therein).

Remark 3.2. If we introduce additional linear equality or inequality constraints on the components of x_δ such as, e.g., a maximum or total duration for the N_p phases ($\delta_0 + \delta_1 + \cdots + \delta_{N_\mathrm{p}} \leqslant T_\mathrm{max}$ or $\delta_0 + \delta_1 + \cdots + \delta_{N_\mathrm{p}} = T_\mathrm{tot}$), maximum or total durations for two or more consecutive phases (e.g., $\delta_{2k} + \delta_{2k+1} \leqslant T_{\mathrm{max},k}$ or $\delta_{2k} + \delta_{2k+1} = T_{\mathrm{tot},k}$), we still obtain an ELCP. The additional linear inequality constraints lead to extra inequalities in (28), and the additional linear equality constraints lead to an extra equation of the form $Px_\delta + q = 0$, which also fits in the ELCP framework.

The ELCP (26)–(29) describes all feasible system trajectories for the first order linear hybrid system with saturation. In order to determine the optimal switching time sequence we could minimize the objective function J over the solution set of the ELCP as follows. If we assume that x_q and x_δ are bounded,[5] then the solution set of the ELCP consists of a union of faces of the (finite and bounded) polytope defined by (26)–(28). Each face of the polyhedron can be represented by its vertices, and the points of the face can be written as convex combinations of these vertices. For each face we could determine for which convex combination of the vertices the objective function J reaches a global minimum over the face and afterwards select the overall minimum.

However, the general ELCP is an NP-hard problem [7]. Furthermore, the algorithm of [7] to compute the solution set of a general ELCP requires exponential execution times. This implies that the ELCP approach sketched above is not feasible if the number of variables is large. Since the number of variables in the ELCP is equal to $2MN_\mathrm{p} + N_\mathrm{c}$, this implies that the ELCP should not be used if M, N_p, or N_c are large.

If the ELCP is not tractable, either we could select lower values for N_c and N_p (which would result in less optimal solutions) or we could use multistart local optimization to determine the optimal switching scheme. For given N_p, N_c, K_c, q_0, $\alpha_{i,k}$'s, $b^\mathrm{ls}_{i,k}$'s, and $b^\mathrm{us}_{i,k}$'s, the evolution of the system to be optimized is uniquely determined by the sequence $\delta_0, \delta_1, \ldots, \delta_{N_\mathrm{c}-1}$ since the remaining δ_k's, the queue lengths $q_i(t)$, and the components of x_q and x_z are given by (7), (2), (5), and (6), respectively. Therefore, we can consider (13)–(18) as a constrained optimization problem in x_δ, where the constraints (16)–(18) are nonlinear constraints. Alternatively, these constraints can be taken into account by adding an extra penalty term to the objective function J if $q_{i,k} < (q_{\mathrm{min},k})_i$ or $q_{i,k} > (q_{\mathrm{max},k})_i$. If we use the penalty function[6] approach, the only remaining constraints on x_δ are the simple upper and lower bound constraints (15). However, the major disadvantage of the multistart local minimization approaches discussed above is that in general the minimization routine will return only a local minimum and that several starting points are necessary to obtain a good approximation to the global optimum. Note that the final solution $x_\delta^\mathrm{opt,curr}$ of the current MPC step can be used to obtain a good initial solution $x_\delta^\mathrm{init,next}$ for the next MPC stepping by setting $\delta_k^\mathrm{init,next} = \delta_{k-1}^\mathrm{opt,curr}$ for $k = 1, 2, \ldots, N_\mathrm{c}$.

Recall that in each MPC step the problem (13)–(18) has to be solved. In order to be able to do this on-line, it is important to have efficient algorithms to solve the

[5] A sufficient condition for this is that $\delta_{\mathrm{min},k}$ and $\delta_{\mathrm{max},k}$ are defined and finite for all k.

[6] Note that in this case a barrier function approach is not advantageous since the optimal solution will often lie on the boundary of the feasible region.

problem. Therefore, we shall now discuss some other approaches to compute solutions very efficiently if there is no saturation at the upper level.

4. Optimal and suboptimal switching time sequences for systems with saturation at a lower level only.

4.1. Optimal switching time sequences.

In this section we consider systems with saturation at the lower level only. So $b_{i,k}^{\mathrm{us}}$ is equal to ∞ for all i, k, or equivalently $(q_{\max,k})_i \leqslant b_{i,k}^{\mathrm{us}}$ for all i, k. We also assume that $q_{\min,k} \leqslant b_k^{\mathrm{ls}}$ for all k, i.e., we do not impose extra lower bound conditions on the queue lengths. The optimal switching problem (13)–(18) then reduces to

$$\underset{x_\delta}{\text{minimize}}\ J \tag{30}$$

subject to

$$\delta_k = \delta_{k-K_{\mathrm{c}}} \quad \text{for } k = N_{\mathrm{c}}, N_{\mathrm{c}}+1, \ldots, N_{\mathrm{p}}-1, \tag{31}$$

$$\delta_{\min,k} \leqslant \delta_k \leqslant \delta_{\max,k} \quad \text{for } k = 0, 1, \ldots, N_{\mathrm{c}}-1, \tag{32}$$

$$q_{k+1} \leqslant q_{\max,k} \quad \text{for } k = 0, 1, \ldots, N_{\mathrm{p}}-1, \tag{33}$$

$$q_{k+1} = \max(q_k + \alpha_k\delta_k,\, b_k^{\mathrm{ls}}) \quad \text{for } k = 0, 1, \ldots, N_{\mathrm{p}}-1. \tag{34}$$

We call this problem $\mathcal{P}$. We define the "relaxed" problem $\tilde{\mathcal{P}}$ corresponding to $\mathcal{P}$ as

$$\underset{x_q, x_\delta}{\text{minimize}}\ J \tag{35}$$

subject to

$$\delta_k = \delta_{k-K_{\mathrm{c}}} \quad \text{for } k = N_{\mathrm{c}}, N_{\mathrm{c}}+1, \ldots, N_{\mathrm{p}}-1, \tag{36}$$

$$\delta_{\min,k} \leqslant \delta_k \leqslant \delta_{\max,k} \quad \text{for } k = 0, 1, \ldots, N_{\mathrm{c}}-1, \tag{37}$$

$$q_{k+1} \leqslant q_{\max,k} \quad \text{for } k = 0, 1, \ldots, N_{\mathrm{p}}-1, \tag{38}$$

$$q_{k+1} \geqslant q_k + \alpha_k\delta_k \quad \text{for } k = 0, 1, \ldots, N_{\mathrm{p}}-1, \tag{39}$$

$$q_{k+1} \geqslant b_k^{\mathrm{ls}} \quad \text{for } k = 0, 1, \ldots, N_{\mathrm{p}}-1. \tag{40}$$

Thus compared to the original problem we have replaced (34) by relaxed equations of the form (22)–(23) without taking (24) into account. As a consequence, x_q and x_δ are not directly coupled anymore. The set of feasible solutions of $\tilde{\mathcal{P}}$ is a convex set, whereas the set of feasible solutions of $\mathcal{P}$ is in general not convex since (34) is a nonconvex constraint. Therefore, the relaxed problem $\tilde{\mathcal{P}}$ will in general be easier to solve than the problem $\mathcal{P}$.

The objective function J is a monotonically nondecreasing function of x_q if for every x_δ and for every $\tilde{x}_q, \hat{x}_q$ with $\tilde{x}_q \leqslant \hat{x}_q$, we have $J(\tilde{x}_q, x_\delta) \leqslant J(\hat{x}_q, x_\delta)$. The following proposition shows that for monotonically nondecreasing objective functions any optimal solution of the relaxed problem $\tilde{\mathcal{P}}$ can be transformed into an optimal solution of the problem $\mathcal{P}$.

PROPOSITION 4.1. *Let the objective function J be a monotonically nondecreasing function of x_q and let (x_q^*, x_δ^*) be an optimal solution of $\tilde{\mathcal{P}}$. If we construct $x_q^\sharp$ such that*

$$q_1^\sharp = \max(q_0 + \alpha_0\delta_0^*, b_0^{\mathrm{ls}}), \tag{41}$$

$$q^\sharp_{k+1} = \max(q^\sharp_k + \alpha_k \delta^*_k, b^{\mathrm{ls}}_k) \qquad \textit{for } k = 1, 2, \ldots, N_\mathrm{p} - 1, \tag{42}$$

then $(x^\sharp_q, x^*_\delta)$ *is an optimal solution of the problem* $\mathcal{P}$.

Proof. Let (x^*_q, x^*_δ) be an optimal solution of $\tilde{\mathcal{P}}$ and let $x^\sharp_q$ be defined by (41)–(42). Clearly, $(x^\sharp_q, x^*_\delta)$ is a feasible solution of $\tilde{\mathcal{P}}$. Define $q^*_0 = q^\sharp_0 = q_0$. Since x^*_q satisfies (39)–(40), we have $\max(q^*_k + \alpha_k \delta^*_k, b^{\mathrm{ls}}_k) \leqslant q^*_{k+1}$ for all k. Since $q^*_0 = q^\sharp_0$, this implies that $q^\sharp_1 \leqslant q^*_1$ and, by induction, also that $q^\sharp_k \leqslant q^*_k$ for $k = 2, 3, \ldots, N_\mathrm{p}$. As a consequence, we have $x^\sharp_q \leqslant x^*_q$ and thus also $J(x^\sharp_q, x^*_\delta) \leqslant J(x^*_q, x^*_\delta)$ since J is a monotonically nondecreasing function of x_q. Since $(x^\sharp_q, x^*_\delta)$ is a feasible solution of $\tilde{\mathcal{P}}$ and since (x^*_q, x^*_δ) is an optimal solution of $\tilde{\mathcal{P}}$, this implies that $(x^\sharp_q, x^*_\delta)$ is also an optimal solution of $\tilde{\mathcal{P}}$.

The set of feasible solutions of $\mathcal{P}$ is a subset of the set of feasible solutions of $\tilde{\mathcal{P}}$. Hence, the minimal value of J over the set of feasible solutions of $\tilde{\mathcal{P}}$ will be less than or equal to the minimal value of J over the set of feasible solutions of $\mathcal{P}$. Since $(x^\sharp_q, x^*_\delta)$ is a feasible solution of $\mathcal{P}$ and an optimal solution of $\tilde{\mathcal{P}}$, this implies that $(x^\sharp_q, x^*_\delta)$ is an optimal solution of $\mathcal{P}$. □

Recall that the objective functions J_1, J_2, J_3, J_4, and J_5 do not explicitly depend on x_q, since x_q can be computed from x_δ (and eliminated from the expressions for the objective functions before considering the relaxation of $\mathcal{P}$). So we have $J_l(\tilde{x}_q, x_\delta) = J_l(\hat{x}_q, x_\delta)$ for any $\tilde{x}_q, \hat{x}_q$ and for $l \in \{1, 2, 3, 4, 5\}$. This implies that J_1, J_2, J_3, J_4, and J_5 are monotonically nondecreasing functions of x_q. So we can use Proposition 4.1 to transform the optimal switching problem for the objective functions J_1 up to J_5 into an optimization problem with a convex feasible set. The resulting (global) solution of the relaxed problem can then be transformed into an optimal switching scheme using (41)–(42). Note, however, that although the feasible set of the relaxed problem is convex, the objective functions J_1 up to J_5 are not convex, so that the overall problem is still nonconvex (and thus in general not easily solvable). Therefore, we now introduce two subsequent approximations of the objective functions J_1 and J_4 that will lead to a linear programming problem, which can be solved very efficiently. The resulting solution can then be used as an initial starting point for the optimization of the relaxed problem.

4.2. A linear programming approximation.

The objective function J is a monotonically increasing function of x_q if for every x_δ and for every $\tilde{x}_q, \hat{x}_q$ with $\tilde{x}_q \leqslant \hat{x}_q$ and $\tilde{x}_q \neq \hat{x}_q$, we have $J(\tilde{x}_q, x_\delta) < J(\hat{x}_q, x_\delta)$. The optimal solution of $\tilde{\mathcal{P}}$ will in general not be a feasible solution of $\mathcal{P}$ unless J is a monotonically increasing function of x_q.

PROPOSITION 4.2. *If J is a monotonically increasing function of x_q, then any optimal solution of the relaxed problem $\tilde{\mathcal{P}}$ is also an optimal solution of the problem $\mathcal{P}$.*

Proof. Let (x^*_q, x^*_δ) be an optimal solution of $\tilde{\mathcal{P}}$ and construct $(x^\sharp_q, x^*_\delta)$ as in the proof of Proposition 4.1. So $x^\sharp_q \leqslant x^*_q$ and $(x^\sharp_q, x^*_\delta)$ is also a feasible solution of $\tilde{\mathcal{P}}$.

Now we show by contradiction that (x^*_q, x^*_δ) is also a feasible solution of $\mathcal{P}$, i.e., that it satisfies (34). Suppose that (x^*_q, x^*_δ) does not satisfy (34). So $x^\sharp_q \neq x^*_q$. Since $x^\sharp_q \leqslant x^*_q$, this implies that $J(x^\sharp_q, x^*_\delta) < J(x^*_q, x^*_\delta)$, which would mean that (x^*_q, x^*_δ) is not an optimal solution of $\tilde{\mathcal{P}}$. Since this is a contradiction, our initial assumption that (x^*_q, x^*_δ) does not satisfy (34) was wrong. Hence, $(x^*_q,\, x^*_\delta)$ also is a feasible solution of the problem $\mathcal{P}$. Since the set of feasible solutions of $\mathcal{P}$ is a subset of the set of feasible solutions of $\tilde{\mathcal{P}}$, this implies that $(x^*_q,\, x^*_\delta)$ is also an optimal solution of $\mathcal{P}$. □

Note that the objective functions J_1, J_2, J_3, J_4, and J_5 are not monotonically increasing functions of x_q. Now we introduce some approximations to the objective functions J_1 and J_4 that are strictly monotonically increasing functions of x_q and for which Proposition 4.2 can be used.[7] This will lead to suboptimal switching time sequences that can be computed very efficiently. We will consider only the approximations for J_1, but for J_4 a similar reasoning can be made.

For a given q_0 and t_0, we define the function $\tilde{q}_i(\cdot, x_q, x_\delta)$ as the piecewise-affine function with breakpoints $(t_k, q_{i,k})$ for $k = 0, 1, \ldots, N_\mathrm{p}$. The approximate objective function $\tilde{J}_1$ is also defined by (8) but with q_i replaced by $\tilde{q}_i$. The value of the objective functions J_1 and $\tilde{J}_1$ depends on the surface under the functions q_i and $\tilde{q}_i$, respectively.[8] If we are computing optimal traffic switching sequences, then the surface under the function $\tilde{q}_i$ will be a reasonable approximation of the surface under the function q_i and then the optimal value of $\tilde{J}_1$ will be a reasonably good approximation of the optimal value of J_1 (see also [6, 8]). Note that the values of J_1 and $\tilde{J}_1$ coincide if there is no saturation in the period under consideration. Since $\tilde{q}_i$ is a piecewise-affine with breakpoints $(t_k, q_{i,k})$ for $k = 0, 1, \ldots, N_\mathrm{p}$, we have [6]

$$\tilde{J}_1(x_q, x_\delta) = \sum_{i=1}^{M} \left(\frac{w_i}{2(\delta_0 + \delta_1 + \cdots + \delta_{N_\mathrm{p}-1})} \sum_{k=0}^{N_\mathrm{p}-1} \delta_k (q_{i,k} + q_{i,k+1}) \right), \tag{43}$$

where $\delta_{N_\mathrm{c}}, \ldots, \delta_{N_\mathrm{p}-c}$ can be replaced using (7). Since $\delta_k > 0$ for all k, $\tilde{J}_1$ is a monotonically increasing function of x_q, which implies that Proposition 4.2 can be applied.

Now we discuss a further approximation of $\tilde{J}_1$ that will lead to a linear programming problem, which can be solved very efficiently. Sometimes we already have a good idea about the relative lengths of the different phases (in a traffic signal situation we know, e.g., that the green phases will be much longer than the amber phases). If we assume that $\delta_k = \rho_k \bar{\delta}$ for all k and for some yet unknown $\bar{\delta}$, then (43) leads to

$$\tilde{J}_1(x_q, x_\delta) = \sum_{i=1}^{M} \frac{w_i}{2\rho_\mathrm{tot}} \left(\rho_0 q_{i,0} + \sum_{k=1}^{N_\mathrm{p}-1} (\rho_k + \rho_{k-1}) q_{i,k} + \rho_{N-1} q_{i,N_\mathrm{p}} \right) \overset{\mathrm{def}}{=} \hat{J}_1(x_q),$$

with $\rho_\mathrm{tot} = \rho_0 + \rho_1 + \cdots + \rho_{N_\mathrm{p}-1}$. Note that $\hat{J}_1$ is an affine function of x_q. Since $w_i > 0$ for all i and $\rho_k > 0$ for all k, $\hat{J}_1$ is a monotonically increasing function of x_q. Hence, by Proposition 4.2 any optimal solution of $\tilde{\mathcal{P}}$ with objective function $\hat{J}_1$ will also be an optimal solution of $\mathcal{P}$ (with objective function $\hat{J}_1$). So the optimal switching problem then reduces to a linear programming problem, which can be solved efficiently using a simplex or an interior point method.

Remark 4.3. The values of the ρ_k's are usually determined on the basis of an educated guess. Alternatively, if we have already performed an MPC step, then we can use the shifted values of the δ_k's of the previous MPC step to obtain an initial guess for the current ρ_k's. Furthermore, we could also use an iterative procedure in which we first select values for the ρ_k's, compute the optimal solution, use the resulting δ_k's to determine new values for the ρ_k's, after which we can again compute the optimal solution, and so on.

[7] This derivation is an extension of our work in [8] where we have considered a special subclass of first order linear hybrid systems with saturation at the lower level only. Although we did not yet use Proposition 4.1 there, we did use a proposition that is similar to Proposition 4.2.

[8] Recall that $q_i(t) \geqslant 0$ for all i, t since we have assumed that $b^{\mathrm{ls}}_{i,k} \geqslant 0$ for all i, k.

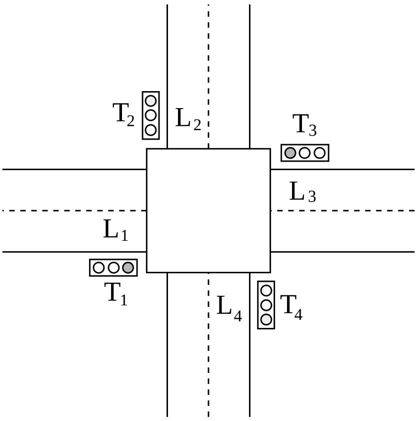


FIG. 1. *A traffic signal controlled intersection of two two-way streets.*

TABLE 1
The traffic signal switching scheme.

Period	T_1, T_3	T_2, T_4
t_0–t_1	red	green
t_1–t_2	red	amber
t_2–t_3	green	red
t_3–t_4	amber	red
t_4–t_5	red	green
t_5–t_6	red	amber
⋮	⋮	⋮

Also note that the assumption on the relative lengths ($\delta_k = \rho_k \bar{\delta}$ for all k) is only used to simplify the objective function; it will not be included explicitly in the linear programming problem. So the variables in this problem are still x_q and x_δ, but the objective depends only on x_q. As a consequence, the optimal δ_k's will in general not satisfy the assumption on the relative lengths (see, e.g., the example of section 5.2).

5. Application: Optimal traffic signal control.

5.1. Optimal traffic signal control. In order to illustrate the effectiveness of Proposition 4.1 we shall use the different approaches presented in this paper to design an optimal switching time sequence for a traffic signal controlled intersection and compare the results.

Consider an intersection of two two-way streets (see Figure 1) with lanes L_i and a traffic signal T_i on each corner of the intersection ($i = 1, 2, 3, 4$). The switching time sequence for the intersection is given in Table 1. Since queue lengths can never become negative and since all the cars can leave a queue provided that we make the length of the green phase large enough, we have $b_k^{\text{ls}} = 0$ for all k. We assume that there is no saturation at the upper level, either due to the fact that there is enough buffer space before the traffic signal in each lane or due to the fact that we impose additional maximal queue length conditions such that $q_{\max,k} \leqslant b_k^{\text{us}}$.

In order to obtain a model that is amenable to mathematical analysis, we shall make two extra assumptions (see also [8]):

(i) the queue lengths are continuous variables,

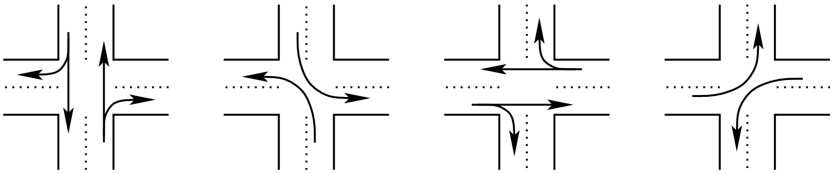


FIG. 2. *The four main phases of a more complex traffic signal switching scheme. The arrows indicate possible directions for the cars that receive a green signal.*

(ii) the average arrival and departure rates of the cars are constant or slowly time-varying.

These assumptions deserve a few remarks:

(i) Recall that the main purpose is to compute optimal traffic signal switching time sequences. Designing optimal switching time sequences is useful only if the arrival and departure rates of vehicles at the intersection are high since then the queue lengths will in general also be large and then approximating the queue lengths by continuous variables will introduce only small errors.

(ii) If we keep in mind that one of the main purposes of the model that we shall derive is the design of optimal traffic signal switching time sequences, then assuming that the average arrival and departure rates are constant is not a serious restriction provided that we use an MPC approach in which we can regularly update the estimates of the arrival and departure rates and of the state of the system.

Let α_i^{a} be the average arrival rate of cars in lane L_i, and let $\alpha_i^{\mathrm{d,green}}$ and $\alpha_i^{\mathrm{d,amber}}$ be the departure rates of cars in lane L_i when the traffic signal T_i is green, respectively, amber. If we define

$$\alpha_{i,k} = \begin{cases} \alpha_i^{\mathrm{a}} & \text{if } T_i \text{ is red in } (t_k, t_{k+1}), \\ \alpha_i^{\mathrm{a}} - \alpha_i^{\mathrm{d,green}} & \text{if } T_i \text{ is green in } (t_k, t_{k+1}), \\ \alpha_i^{\mathrm{a}} - \alpha_i^{\mathrm{d,amber}} & \text{if } T_i \text{ is amber in } (t_k, t_{k+1}) \end{cases}$$

for all i, k, then the relation between the switching time instants and the queue lengths is described by a system of equations of the form

$$\frac{dq_i(t)}{dt} = \begin{cases} \alpha_{i,k} & \text{if } b_{i,k}^{\mathrm{ls}} < q_i(t), \\ 0 & \text{otherwise.} \end{cases}$$

Thus the system can be considered as a first order linear hybrid system with lower saturation only. Hence, we can use the techniques presented in sections 3.2 and 4 to compute optimal and suboptimal traffic signal switching schemes.

In the simple traffic signal set-up discussed above we did not make a distinction between cars that turn left, right, or that go straight ahead. However, the approach presented in this paper can also be applied to more complex set-ups or more complex traffic signal switching schemes for single intersections such as, e.g., the one depicted in Figure 2 which consists of four main phases with amber phases in between, where in the first main phase cars on the north-south axis can go straight ahead or turn right, in the next main phase they can turn left, and in the next two main phases the same process is repeated for the traffic on the east-west axis.

5.2. Worked example. The following traffic signal control example illustrates that using Proposition 4.1 leads to efficient computation of optimal switching time sequences and that the approximations introduced in section 4.2 lead to reasonably good suboptimal solutions. Since we are mainly interested in the computation times, we will consider only one step of the MPC algorithm. All times will be expressed in seconds and all rates in vehicles per second. The numerical results will be given up to two decimal places.

Consider the intersection of Figure 1 with the switching scheme of Table 1 and with the following data: $N_\mathrm{p} = 14$, $N_\mathrm{c} = 8$, $\alpha_1^\mathrm{a} = 0.23$, $\alpha_2^\mathrm{a} = 0.12$, $\alpha_3^\mathrm{a} = 0.19$, $\alpha_4^\mathrm{a} = 0.11$, $\alpha_1^\mathrm{d,green} = 0.50$, $\alpha_2^\mathrm{d,green} = \alpha_4^\mathrm{d,green} = 0.35$, $\alpha_3^\mathrm{d,green} = 0.45$, $\alpha_1^\mathrm{d,amber} = \alpha_3^\mathrm{d,amber} = 0.03$, $\alpha_2^\mathrm{d,amber} = \alpha_4^\mathrm{d,amber} = 0.02$, $q_0 = [\,17 \;\; 12 \;\; 14 \;\; 8\,]^T$, and $q_{\max,k} = [\,20 \;\; 15 \;\; 20 \;\; 15\,]^T$ for all k. Since a green-amber-red cycle consists of four consecutive phases (see Table 1) we set $K_\mathrm{c} = 4$.

We want to compute a traffic signal switching sequence $t_0, t_1, \dots, t_{N_\mathrm{c}-1}$ that minimizes J_1 with $w = [\,2 \;\; 1 \;\; 2 \;\; 1\,]^T$. The minimum and maximum length of the green phases are, respectively, 9 and 90. Note that for the simple setup of this example and for the objective function J_1 it does not make sense to consider a varying amber duration since during the amber phases the average queue length always increases, which implies that the optimal duration of the amber phases in this case will always be equal to the given lower bound for the amber phase. Therefore, we fix the length of the amber phase by setting the minimal and the maximal length of the amber phases equal to 3.

We have computed an optimal switching interval vector $x^*_{\delta,\mathrm{elcp}}$ using the ELCP method, a suboptimal switching vector $x^*_{\delta,\mathrm{nlcon}}$ using constrained optimization with nonlinear constraints,[9] and a suboptimal solution $x^*_{\delta,\mathrm{penalty}}$ using constrained optimization[10] with a quadratic penalty function for queue lengths that exceed $q_{\max,k}$. Based on Propositions 4.1 and 4.2 we have computed a solution $x^*_{\delta,\mathrm{relaxed}}$ that minimizes J_1 for the relaxed problem $\tilde{\mathcal{P}}$ and a solution $x^*_{\delta,\mathrm{approx}}$ that minimizes the approximate objective function $\tilde{J}_1$ for the relaxed problem $\tilde{\mathcal{P}}$. Finally, we computed a switching interval vector $x^*_{\delta,\mathrm{lp}}$ that minimizes $\hat{J}_1$ for the relaxed problem $\tilde{\mathcal{P}}$ with the affine objective function obtained by assuming that for the east-west axis the length of the green phases is 1.5 times the length of the red phases and 10 times the length of the amber phases. (Note that this is just a rough guess.)

We have used the sequential quadratic programming function `e04ucc` of the NAG C Library for the nonlinear optimizations. To solve the linear programming problem we have used the function `e04mfc` of the NAG C library, which uses an active set method.

In Table 2 we have listed the value of the objective functions J_1, $\tilde{J}_1$, and $\hat{J}_1$ for the various switching interval vectors and the CPU time needed to compute the switching interval vectors on a Pentium II 300 MHz PC running Linux and with 64 MB RAM. The CPU time values listed in the table are average values over 10 experiments.[11] All the routines used in the computations either have been implemented in C or were

[9] We give the best solution over 20 runs with random initial points. Only 14 runs resulted in a feasible solution. For these 14 runs the mean of the objective values of the local minima returned by the minimization routine was 48.50 with a standard deviation of 3.38.

[10] We give the best solution over 10 runs with random initial points. For the 9 runs that resulted in a feasible solution, the mean of the objective values was 46.45 with a standard deviation of 0.08.

[11] For $x^*_{\delta,\mathrm{nlcon}}$ we have listed the CPU time needed for 20 runs with random initial points and for $x^*_{\delta,\mathrm{penalty}}$ we have listed the CPU time needed for 10 runs with random initial points (see also footnotes 9 and 10). Note, however, that even for a single run the average CPU time needed for these solutions is much higher that the CPU time needed for the $x^*_{\delta,\mathrm{relaxed}}$ solutions.

TABLE 2

The values of the objective functions J_1, $\tilde{J}_1$, and $\hat{J}_1$, and the CPU time needed to compute the (sub-)optimal switching vectors of the example of section 5.2.

x_δ^*	$J_1(x_\delta^*)$	$\tilde{J}_1(x_\delta^*)$	$\hat{J}_1(x_\delta^*)$	CPU time
$x_{\delta,\mathrm{elcp}}^*$	46.41	49.05	53.66	64 619.07
$x_{\delta,\mathrm{nlcon}}^*$	46.41	49.05	53.66	216.71
$x_{\delta,\mathrm{penalty}}^*$	46.41	49.05	53.66	29.71
$x_{\delta,\mathrm{relaxed}}^*$	46.41	49.05	53.66	0.36
$x_{\delta,\mathrm{approx}}^*$	46.41	49.05	53.66	3.56
$x_{\delta,\mathrm{lp}}^*$	46.63	49.12	53.62	0.16

compiled to object code. As a consequence, all the CPU times can be considered as a measure for the number of floating point operations that were needed to compute the various (sub-)optimal switching interval vectors.

Note that the optimal values of J_1 and $\tilde{J}_1$ differ by about 5 %, so that in this case the optimal value of $\tilde{J}_1$ is indeed a reasonably good approximation of the optimal value of J_1. While computing $x_{\delta,\mathrm{relaxed}}^*$ we have only N_c optimization variables (i.e., the δ_k's, since the q_k's do not appear in the objective function and since they can be eliminated from the constraints). However, for $x_{\delta,\mathrm{approx}}^*$ we have $N_\mathrm{c}+MN_\mathrm{p}$ optimization variables (i.e., the δ_k's and the components of the q_k's since in this case the q_k's appear in the objective function and thus cannot be eliminated). This is one of the reasons why the computation of $x_{\delta,\mathrm{relaxed}}^*$ requires less CPU time than the computation of $x_{\delta,\mathrm{approx}}^*$. Additional numerical experiments and simulations can be found in [6].

In this example the ELCP solution is only given as a reference since the CPU time needed to compute the optimal switching interval vector using the ELCP algorithm of [7] increases exponentially as M, N_p, or N_c increase (see also [6]). This implies that the ELCP approach should never be used in practice, but one of the other approaches should be used instead.

If we look at Table 2, then we see that the $x_{\delta,\mathrm{relaxed}}^*$ solution—which is based on Proposition 4.1—is clearly the most interesting. If we take the trade-off between optimality and efficiency into account, the $x_{\delta,\mathrm{relaxed}}^*$ solution outperforms the solutions obtained using the other approaches (see also [6]).

If we use an MPC approach, then the computation time required for the $x_{\delta,\mathrm{relaxed}}^*$ solution is less than the minimum lower bound for the phase lengths, which implies that we can first measure the queue lengths at t_0 and start the computation at time t_0. In that way we can use the exact initial state q_0. Note that using the exact initial state q_0 (or a good estimate) also introduces a kind of feedback in the control loop.

6. Conclusions and future research. We have considered the determination of optimal switching time sequences for a class of first order linear hybrid systems subject to saturation. First we have introduced the ELCP and indicated how it can be used to describe the set of feasible system trajectories for a first order linear hybrid system with saturation. Optimization over the solution set of the ELCP then yields the optimal switching time sequence. Since the ELCP is NP-hard, we have also discussed several other techniques to compute optimal and suboptimal switching time sequences for first order linear hybrid systems subject to saturation at the lower level only. We have shown that if the objective function is a monotonically nondecreasing function of the queue lengths, then the optimal switching problem can be transformed into an optimization problem with a convex feasible set and then the optimal switching time sequence can be computed more efficiently. By making some approximations, the optimal switching problem can even be transformed into a linear programming problem. We have illustrated these approaches by computing (sub-)optimal switching time sequences for a traffic signal controlled intersection. Since the time required for the computations using the most efficient approach is less than the minimum time between two consecutive switchings, our method can be used in a model predictive control framework in which the model of the system and the optimal switching sequence are re-estimated or recomputed after each switching.

In this paper we have derived methods to optimize quantitative performance measures such as average or worst case waiting times and queue lengths for a linear hybrid system with saturation. If we are more interested in qualitative properties such as, e.g., safety, we could use the techniques presented in [18].

An important topic for future research is the extension of the results obtained in this paper to networks of dependent queues, i.e., a situation where the outputs of some queues will be connected to the inputs of some other queues. If we use an MPC strategy in combination with a *decentralized control* solution, we can still apply the approach given in this paper: if we know or measure all routing rates[12] and all traveling times from one queue to another, we can use measurements from one queue to predict the arrival rates at the other queues. Other topics for further research include development of other efficient algorithms and/or approximations to compute optimal switching time sequences for first order linear hybrid systems with saturation, investigation of the use of the ELCP to model and to control other classes of hybrid systems, and extension of the results presented in this paper to more general classes of hybrid systems.

Acknowledgment. We would like to thank the anonymous referees for their useful suggestions and comments, which have certainly contributed to improving the presentation of the paper.

REFERENCES

[1] A. S. Morse, C. C. Pantelides, S. Sastry, and J. M. Schumacher, eds., *Special Issue on Hybrid Systems*, Automatica J. IFAC, 35 (1999).

[2] V. Blondel and J. Tsitsiklis, *Complexity of stability and controllability of elementary hybrid systems*, Automatica J. IFAC, 35 (1999), pp. 479–489.

[3] E. Camacho and C. Bordons, *Model Predictive Control in the Process Industry*, Springer-Verlag, Berlin, 1995.

[4] D. Clarke, C. Mohtadi, and P. Tuffs, *Generalized predictive control—Part* I. *The basic algorithm*, Automatica J. IFAC, 23 (1987), pp. 137–148.

[5] R. Cottle, J. Pang, and R. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[6] B. De Schutter, *Optimal Control of a Class of Linear Hybrid Systems with Saturation: Addendum*, Tech. Report bds:99-03a, Control Laboratory, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands, 1999.

[7] B. De Schutter and B. De Moor, *The extended linear complementarity problem*, Math. Programming, 71 (1995), pp. 289–325.

[8] B. De Schutter and B. De Moor, *Optimal traffic light control for a single intersection*, European J. Control, 4 (1998), pp. 260–276.

[9] C. García, D. Prett, and M. Morari, *Model predictive control: Theory and practice—A survey*, Automatica J. IFAC, 25 (1989), pp. 335–348.

[10] N. Gartner, J. Little, and H. Gabbay, *Simultaneous optimization of offsets, splits, and cycle time*, Transportation Research Record, (1976), pp. 6–15.

[12] The routing rates are the numbers of cars, the amount of fluid, ... that will be routed from the output of one queue to the input of another queue.

[11] T. Henzinger and S. Sastry, eds., *Hybrid Systems: Computation and Control*, Lecture Notes in Comput. Sci. 1386, Springer-Verlag, New York, 1998.

[12] P. J. Antsaklis and A. Nerode, eds., Special Issue on Hybrid Systems, IEEE Trans. Automat. Control, 43 (1998).

[13] P. Kachroo and K. Özbay, *Solution to the user equilibrium dynamic traffic routing problem using feedback linearization*, Transportation Research Part B, 32 (1998), pp. 343–360.

[14] A. May, *Traffic Flow Fundamentals*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[15] E. Park, J. Lim, I. Suh, and Z. Bien, *Hierarchical optimal control of urban traffic networks*, Internat. J. Control, 40 (1984), pp. 813–829.

[16] M. Singh and H. Tamura, *Modelling and hierarchical optimization for oversaturated urban road traffic networks*, Internat. J. Control, 20 (1974), pp. 913–934.

[17] F. Vaandrager and J. van Schuppen, eds., *Hybrid Systems: Computation and Control*, Lecture Notes in Comput. Sci. 1569, Springer, New York, 1999.

[18] H. Wong-Toi, *The synthesis of controllers for linear hybrid automata*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 4607–4612.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 852–871

A BAYES FORMULA FOR GAUSSIAN NOISE PROCESSES AND ITS APPLICATIONS*

PRANAB K. MANDAL† AND V. MANDREKAR‡

Abstract. An elementary approach is used to derive a Bayes-type formula, extending the Kallianpur–Striebel formula for the nonlinear filters associated with the Gaussian noise processes. In the particular cases of certain Gaussian processes, recent results of Kunita and of Le Breton on fractional Brownian motion are derived. We also use the classical approximation of the Brownian motion by the Ornstein–Uhlenbeck dispersion process to solve the "instrumentability" problem of Balakrishnan. We give precise conditions for the convergence of the filter based on the Ornstein–Uhlenbeck dispersion process to the filter based on the Brownian motion. It is also shown that the solution of the Zakai equation can be approximated by that of a (deterministic) partial differential equation.

Key words. filtering, Gaussian noise process, Bayes formula, Ornstein–Uhlenbeck dispersion process, Zakai equation, fractional Brownian motion

AMS subject classifications. 60G35, 60G15, 62M20, 93E11

PII. S0363012998343380

1. Introduction.

The general filtering problem can be described as follows. The *signal* or *system process* $(X_t, 0 \le t \le T)$ is unobservable. Information about (X_t) is obtained by observing another process Y which is a function of X corrupted by noise, i.e.,

$$Y_t = \beta_t + N_t, \quad 0 \le t \le T, \tag{1.1}$$

where β_t is measurable with respect to $\mathcal{F}_t^X$, the σ-field generated by the signal $\{X_u,\ 0 \le u \le t\}$ (augmented by the inclusion of zero probability sets), and (N_t) is some noise process. The observation σ-field $\mathcal{F}_t^Y = \sigma\{Y_u,\ 0 \le u \le t\}$ contains all the available information about X_t. The primary aim of filtering theory is to get an estimate of X_t based on the information $\mathcal{F}_t^Y$. This is given by the conditional distribution ν_t of X_t given $\mathcal{F}_t^Y$, or equivalently, the conditional expectation $E(f(X_t)|\mathcal{F}_t^Y)$ for a rich enough class of functions f. Since this estimate minimizes the squared error loss, ν is called the *optimal filter.*

In the classical case one considers the observation model

$$dY_t = h(t, X_t)\,dt + dW_t, \tag{1.2}$$

where W is the Wiener process independent of X and h satisfies the conditions for the Girsanov theorem (for details, see [10]). Kallianpur and Striebel [12] derived a Bayes-type formula for the conditional distribution ν_t of the form $\nu_t = \frac{\sigma_t}{\langle \sigma_t, 1\rangle}$, where σ_t is the so-called unnormalized conditional distribution. In the case when the signal process X_t is a Markov process, satisfying the SDE

$$dX_t = A(t, X_t)\,dt + B(t, X_t)\,d\tilde{W}_t,$$

[*] Received by the editors August 10, 1998; accepted for publication (in revised form) June 14, 2000; published electronically October 11, 2000.
http://www.siam.org/journals/sicon/39-3/34338.html

[†] EURANDOM / LG 1.21, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (mandal@eurandom.tue.nl).

[‡] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824 (mandrekar@stt.msu.edu).

where $\tilde{W}$ is another Wiener process independent of W, Zakai [20] showed that σ_t is the unique solution of a measure valued stochastic differential equation. It is also known that the filter $\mathcal{V}_t$ satisfies a stochastic differential equation widely known as the *Kushner* or *FKK* equation (see, e.g., [14] and [8]).

That the noise process (N_t) is a Wiener process plays an important part in deriving all of the above equations and formulas. However, in the real physical system, the noise process (N_t) may not be exactly a Wiener process. In this case no effective way of computing the filter is known. In a recent paper Kunita [13] considered the filtering problem with the observation process

$$Y_t = \int_0^t h(X_s)\, ds + N_t,$$

where N_t is a particular Gaussian process connected to W_t by a kernel. He derived a Bayes-type formula extending the one by Kallianpur and Striebel. We generalize this result to any Gaussian noise process N_t with β in the model (1.1) belonging almost surely (a.s.) to the reproducing kernel Hilbert space (RKHS) of the covariance of (N_t). It should be noted that this result with a modified Kallianpur–Striebel proof was first obtained by one of the authors [19]. However, the proof presented here is entirely new and is based on an extension of a one-dimensional result which makes (Y_t), under a change of measure, Gaussian with the same distribution as that of (N_t) and independent of (X_t). As an immediate consequence we get the result of Kunita and Kallianpur and Striebel with a simple proof.

In case (X_t) is a diffusion process, one can attempt to obtain a Zakai-type equation whose solution gives a recursive form of the filter. Unfortunately, in the full generality of the problem, it does not seem easy to even formulate such an equation. However, we have indicated how to obtain such an equation for the Ornstein–Uhlenbeck dispersion process. We have partial results in this direction for the case of Kunita and that of the fractional Brownian motion (fBm). The solution of these equations requires new methods. We shall present this work elsewhere once it becomes complete.

Recently, stochastic models appropriate for long-range dependent phenomena have been given a great deal of interest and numerous theoretical resets and successful applications have been already reported (see, e.g., Beran [4] and references therein). In this view we consider the filtering problem with the fBm noise process. We obtain a general form of the filter in this case. In particular, if $X_t = \eta$ for all t, then we obtain all the results in [6] under his assumptions.

We also discuss the issue raised by Balakrishnan [2] regarding "instrumenting" the filtering problem. An approach to this problem using finitely additive measures was given by Kallianpur and Karandikar in their well-known monograph [11]. They work on the Cameron–Martin space with a finitely additive measure and approximate the filter through an extension. Our method is to follow the classical approach of physics; namely, to approximate the Wiener noise process by the Ornstein–Uhlenbeck dispersion process (see, e.g., Nelson [16]). Using our Bayes formula we show that the usual filtering theory with the Wiener process can be obtained as a limit. The latter uses the ideas of Kunita [13] on stability. We give here the precise conditions for the validity of stability. It should be observed that the theory with the Ornstein–Uhlenbeck dispersion process can be instrumented. We approximate the dispersion process by neglecting a term of order σ^{-1} for σ large (cf. (6.15)) and for this process we obtain a Zakai equation which can be approximated by an ordinary partial differential equation.

The article is organized as follows. In section 2, we give a brief overview of RKHS and its connection with stochastic processes. The extension of the Kallianpur–Striebel formula is obtained in section 3. We discuss Kunita's result in section 4. Section 5 deals with the filtering problem with the fBm as the noise process. Finally, in section 6, the filtering problem corresponding to the Ornstein–Uhlenbeck dispersion noise process is considered along with its limit.

2. Reproducing kernel Hilbert space and stochastic processes. A Hilbert space H consisting of real-valued functions on some set $\mathbf{T}$ is said to be an RKHS if there exists a function K on $\mathbf{T}\times\mathbf{T}$ with the following two properties: for every t in $\mathbf{T}$ and g in H,

(i) $K(\cdot,t)\in H$,

(ii) $(g(\cdot),K(\cdot,t))=g(t)$ (the reproducing property).

K is called the *reproducing kernel* of H. The following basic properties can be found in Aronszajn [1].

(1^o) If a reproducing kernel exists, then it is unique.

(2^o) If K is the reproducing kernel of H, then $\{K(\cdot,t),t\in\mathbf{T}\}$ spans H.

(3^o) If K is the reproducing kernel of H, then it is nonnegative definite in the sense that for all $t_1,\dots,t_n$ in $\mathbf{T}$ and $a_1,\dots,a_n\in\mathbb{R}$

$$\sum_{i,j=1}^{n} K(t_i,t_j)a_ia_j \;\geq\; 0.$$

The converse of (3^o), stated in Theorem 2.1 below, is fundamental toward understanding the RKHS representation of Gaussian processes. A proof of the theorem can be found in Aronszajn [1].

THEOREM 2.1 (E. H. Moore). *A symmetric nonnegative definite function K on $\mathbf{T}\times\mathbf{T}$ generates a unique Hilbert space, which we denote by $H(K)$ or sometimes by $H(K,\mathbf{T})$, of which K is the reproducing kernel.*

Now suppose $K(s,t),\ s,t\in\mathbf{T}$, is a nonnegative definite function. Then, by Theorem 2.1, there is an RKHS, $H(K,\mathbf{T})$, with K as its reproducing kernel. If we restrict K to $\mathbf{T}'\times\mathbf{T}'$ where $\mathbf{T}'\subset\mathbf{T}$, then K is still a nonnegative definite function. Hence K restricted to $\mathbf{T}'\times\mathbf{T}'$ will also correspond to an RKHS $H(K,\mathbf{T}')$ of functions defined on $\mathbf{T}'$. The following result from Aronszajn [1, p. 351] explains the relationship between these two.

THEOREM 2.2. *Suppose $K_{\mathbf{T}}$, defined on $\mathbf{T}\times\mathbf{T}$, is the reproducing kernel of the Hilbert space $H(K_{\mathbf{T}})$ with the norm $\|\cdot\|$. Let $\mathbf{T}'\subset\mathbf{T}$, and $K_{\mathbf{T}'}$ be the restriction of $K_{\mathbf{T}}$ on $\mathbf{T}'\times\mathbf{T}'$. Then $H(K_{\mathbf{T}'})$ consists of all f in $H(K_{\mathbf{T}})$ restricted to $\mathbf{T}'$. Further, for such a restriction $f'\in H(K_{\mathbf{T}'})$ the norm $\|f'\|_{H(K_{\mathbf{T}'})}$ is the minimum of $\|f\|_{H(K_{\mathbf{T}})}$ for all $f\in H(K_{\mathbf{T}})$ whose restriction to $\mathbf{T}'$ is f'.*

If $K(s,t)$ is the covariance function for some zero mean process $Z_t,t\in\mathbf{T}$, then, by Theorem 2.1, there exists a unique RKHS, $H(K,\mathbf{T})$, for which K is the reproducing kernel. It is also easy to see (e.g., see Theorem 3D in [18]) that there exists a congruence (linear, one-to-one, inner product preserving map) between $H(K)$ and $\overline{\mathrm{sp}}^{L^2}\{Z_t,t\in\mathbf{T}\}$ which takes $K(\cdot,t)$ to Z_t. Let us denote by $\langle Z,h\rangle\in\overline{\mathrm{sp}}^{L^2}\{Z_t,t\in\mathbf{T}\}$ the image of $h\in H(K,\mathbf{T})$ under the congruence.

We conclude the section with an important special case.

2.1. A useful example. Suppose the stochastic process Z_t is a Gaussian process given by

$$Z_t = \int_0^t F(t,u) dW_u, \ 0 \le t \le T,$$

where $\int_0^t F^2(t,u)du < \infty$ for all $0 \le t \le T$. Then the covariance function

$$K(s,t) \equiv E(Z_s Z_t) = \int_0^{t \wedge s} F(t,u)F(s,u)du \tag{2.1}$$

and the corresponding RKHS is given by

$$H(K) = \left\{ g : g(t) = \int_0^t F(t,u)g^*(u)du, 0 \le t \le T \right\} \tag{2.2}$$

for some (necessarily unique) $g^* \in \overline{\text{sp}}^{L^2}\{F(t,\cdot)1_{[0,t]}(\cdot), 0 \le t \le T\}$ with the inner product

$$(g_1, g_2)_{H(K)} = \int_0^T g_1^*(u) g_2^*(u) du,$$

where

$$g_1(s) = \int_0^s F(s,u)g_1^*(u)du \quad \text{and} \quad g_2(s) = \int_0^s F(s,u)g_2^*(u)du.$$

For $0 \le t \le T$, by taking $K(\cdot,t)^*$ to be $F(t,\cdot)1_{[0,t]}(\cdot)$, we see, from (2.1) and (2.2), that $K(\cdot,t) \in H(K)$. To check the reproducing property suppose $h(t) = \int_0^t F(t,u)h^*(u)\,du \in H(K)$. Then

$$(h, K(\cdot,t))_{H(K)} = \int_0^T h^*(u)K(\cdot,t)^*\,du = \int_0^t h^*(u)F(t,u)\,du = h(t).$$

Also, in this case, it is very easy to check (cf. [17], Theorem 4D) that the congruence between $H(K)$ and $\overline{\text{sp}}^{L^2}\{Z_t, t \in \mathbf{T}\}$ is given by

$$\langle Z, g \rangle \ = \int_0^T g^*(u) dW_u. \tag{2.3}$$

3. Extension of the Kallianpur–Striebel formula.

Suppose $X_t, 0 \le t \le T$, is a real-valued signal process and the observation process is given by

$$Y_t = \beta(t,X) + N_t, \quad 0 \le t \le T, \tag{3.1}$$

where $\beta : [0,T] \times \mathbb{R}^{[0,T]} \to \mathbb{R}$ is a nonanticipative function and the noise process (N_t) is independent of the signal process (X_t). We are interested in finding the best estimate of $f(X_t)$ based on $\mathcal{F}_t^Y$, which is given by the conditional expectation $E(f(X_t)|\mathcal{F}_t^Y)$. First we consider the one-dimensional analogue of the problem which captures the main idea of obtaining a Bayes-type formula for $E(f(X_t)|\mathcal{F}_t^Y)$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Suppose Z is a standard normal random variable independent of X and $Y = X + Z$. Consider the problem of computing $E(X|Y)$. Suppose $P \ll Q$ and $\mathcal{G} \subset \mathcal{F}$ is a sub-σ-field. Then

$$E_P(X|\mathcal{G}) = \frac{E_Q\left(X\, \frac{dP}{dQ}\middle|\, \mathcal{G}\right)}{E_Q\left(\frac{dP}{dQ}\middle|\, \mathcal{G}\right)}.$$

If we define

$$dQ = \exp\left\{-XY + \frac{1}{2}X^2\right\} dP,$$

then Q is a probability measure. Also, considering the joint characteristic function, under Q, of X and Y it is easy to see that under Q, Y is a standard normal random variable independent of X, and X has the same probability distribution as under P.

We now give the analogue of the above-mentioned result for the general Gaussian processes. Suppose N_t is a Gaussian process with zero mean, i.e., $m_t \equiv E(N_t) = 0$ and with the covariance function $R(s,t) \equiv E(N_s N_t)$. Suppose that R is continuous on $[0,T] \times [0,T]$. Let $\{\xi_t, 0 \le t \le T\}$ be another process with values in a space $\mathcal{S}$ and independent of $\{N_t, 0 \le t \le T\}$. Suppose

$$Y_t = f(t, \xi) + N_t, \quad 0 \le t \le T, \tag{3.2}$$

where f is a measurable nonanticipative functional on $[0,T] \times \mathcal{S}^{[0,T]}$.

Let $H(R;t)$ denote the RKHS corresponding to $R|_{[0,t]\times[0,t]}$, with norm $\|\cdot\|_t$ and $H(R) = H(R;T)$. Also, let $\langle N, \cdot\rangle_t$ denote the congruence between $H(R;t)$ and $\overline{\text{sp}}^{L^2}\{N_s, 0 \le s \le t\}$ so that for $g, h \in H(R;t)$, the random variables $\langle N, g\rangle_t$ and $\langle N, h\rangle_t$ are normal random variables with mean zero and covariance $E(\langle N, g\rangle_t\, \langle N, h\rangle_t)$ $= (g,h)_{H(R;t)}$. Then we have the following.

THEOREM 3.1. *Suppose $f(\cdot) \equiv f(\cdot, \xi)$ in* (3.2) *is in $H(R)$ a.s. Define for each t, $(0 \le t \le T)$,*

$$dQ_t = e^{-\langle N, f\rangle_t - \frac{1}{2}\|f\|_t^2} dP. \tag{3.3}$$

Then Q_t is a probability measure, and under Q_t, we have that

(i) *$(Y_s)_{0 \le s \le t}$ is a Gaussian process with zero mean and covariance function R, and is independent of $(\xi_s)_{0 \le s \le T}$;*

(ii) *$(\xi_s)_{0 \le s \le T}$ has the same distribution as under P.*

Remark. It should be noted that, in case (N_t) is the Brownian motion, one can interpret (i) of the theorem as the analogue of the Girsanov theorem except that the functions f are from a smaller class than those considered by Girsanov. The property that Q_t is a probability measure is automatically satisfied in this case due to the independence of the processes (ξ_t) and (N_t). For this one uses the Cameron–Martin result for each fixed value of (ξ_t).

Proof of Theorem 3.1. Fix $0 \le t \le T$. First note that since $f(\cdot) \in H(R)$ a.s., by Theorem 2.2, $f|_{[0,t]} \in H(R;t)$ a.s. That Q_t is a probability measure follows from the fact that N and ξ are independent and for $g \in H(R;t)$, $\langle N, g\rangle_t$ is a zero mean normal random variable with variance $\|g\|_t^2$. Now suppose $0 \le s_1, \dots, s_m \le t$, $0 \le t_1, \dots, t_n \le T$, $g_1, \dots, g_n : \mathcal{S} \to \mathbb{R}$ are measurable, and $\alpha_1, \dots, \alpha_n, \gamma_1, \dots, \gamma_m$ are real numbers. Consider the joint characteristic function

$$\begin{aligned}
&E_{Q_t}\left[e^{i(\alpha_1 g_1(\xi_{t_1})+\cdots+\alpha_n g_n(\xi_{t_n}))+i(\gamma_1 Y_{s_1}+\cdots+\gamma_m Y_{s_m})}\right]\\
&= E_P\left[e^{i\sum_{k=1}^n \alpha_k g_k(\xi_{t_k})+i\sum_{j=1}^m \gamma_j Y_{s_j}}e^{-\langle N,f\rangle_t-\frac{1}{2}\|f\|_t^2}\right]\\
&= E_P\left[e^{i\sum_{k=1}^n \alpha_k g_k(\xi_{t_k})-\frac{1}{2}\|f\|_t^2+i\sum_{j=1}^m \gamma_j f(s_j)}e^{i\sum_{j=1}^m \gamma_j N_{s_j}-\langle N,f\rangle_t}\right]\\
&= E_P\left[e^{i\sum_{k=1}^n \alpha_k g_k(\xi_{t_k})-\frac{1}{2}\|f\|_t^2+i\sum_{j=1}^m \gamma_j f(s_j)}E_P\left(e^{i\sum_{j=1}^m \gamma_j N_{s_j}-\langle N,f\rangle_t}\middle|\mathcal{F}_T^\xi\right)\right]\\
&= E_P\left[e^{i\sum_{k=1}^n \alpha_k g_k(\xi_{t_k})-\frac{1}{2}\|f\|_t^2+i\sum_{j=1}^m \gamma_j f(s_j)}\right.\\
&\qquad\left.\times\, e^{-\sum_{j,l=1}^m \gamma_j\gamma_l R(s_j,s_l)-\frac{1}{2}2i\sum_{j=1}^m \gamma_j f(s_j)+\frac{1}{2}\|f\|_t^2}\right]\\
&= E_P\left[e^{i\sum_{k=1}^n \alpha_k g_k(\xi_{t_k})}\right]e^{-\sum_{j,l=1}^m \gamma_j\gamma_l R(s_j,s_l)}.
\end{aligned}$$

Hence the assertions (i) and (ii) follow. □

Let us now consider the observation process (Y_t) given by (3.1). Suppose that the noise process (N_t) is Gaussian with continuous covariance function R. It is easy to see, from (3.3) with $\mathcal{S}=\mathbb{R}$, $\xi = X$, and $f(\cdot,\xi)=\beta(\cdot,X)$, that

$$\frac{dP}{dQ_t} = \exp\left\{\langle Y,\beta(\cdot,X)\rangle_t - \frac{1}{2}\|\beta(\cdot,X)\|_t^2\right\} \quad \text{a.s. } [Q_t].$$

This is because if $\beta^n(\cdot)=\sum_{j=1}^{k_n} a_{nj}R(\cdot,t_j^n)\in H(R;t)$, $n=1,2,\ldots$, are such that $\beta^n\to\beta\equiv\beta(\cdot,X)$ in $H(R;t)$, then

$$\begin{aligned}
\langle Y,\beta\rangle_t &= \lim_{n\to\infty}\langle Y,\beta^n\rangle_t \quad (Q_t\text{-a.s. and hence } P\text{-a.s.})\\
&= \lim_{n\to\infty}\sum_{j=1}^{k_n} a_{nj}Y_{t_j^n} = \lim_{n\to\infty}\sum_{j=1}^{k_n} a_{nj}N_{t_j^n} + \lim_{n\to\infty}\sum_{j=1}^{k_n} a_{nj}\beta_{t_j^n}\\
&= \lim_{n\to\infty}\langle N,\beta^n\rangle_t + \lim_{n\to\infty}(\beta,\beta^n)_{H(R;t)} = \langle N,\beta\rangle_t + \|\beta\|_t^2 \quad P\text{-a.s.}
\end{aligned}\tag{3.4}$$

Then for any $\mathcal{F}_T^X$-measurable integrable function $g(T,X)$, we have

$$\begin{aligned}
E_P(g(T,X)|\mathcal{F}_t^Y) &= \frac{E_{Q_t}\left(g(T,X)\frac{dP}{dQ_t}\middle|\mathcal{F}_t^Y\right)}{E_{Q_t}\left(\frac{dP}{dQ_t}\middle|\mathcal{F}_t^Y\right)}\\
&= \frac{E_{Q_t}\left(g(T,X)e^{\langle Y,\beta(\cdot,X)\rangle_t-\frac{1}{2}\|\beta(\cdot,X)\|_t^2}\middle|\mathcal{F}_t^Y\right)}{E_{Q_t}\left(e^{\langle Y,\beta(\cdot,X)\rangle_t-\frac{1}{2}\|\beta(\cdot,X)\|_t^2}\middle|\mathcal{F}_t^Y\right)}.
\end{aligned}\tag{3.5}$$

From Theorem 3.1, $\{Y_s, 0\le s\le t\}$, under Q_t, is independent of $\{X_s, 0\le s\le T\}$ and the distribution of X, under Q_t, is the same as that under P. Hence the conditional expectations of the form $E_{Q_t}(\phi(X,Y)|\mathcal{F}_t^Y)$ can be evaluated as

$$E_{Q_t}(\phi(X,Y)|\mathcal{F}_t^Y)(\omega) = \int_\Omega \phi(X(\omega'),Y(\omega))Q_t(d\omega') = \int \phi(x,Y(\omega))dP_X(x),$$

where P_X is the probability distribution of X. Hence, from (3.5), we have the following.

THEOREM 3.2. *Suppose that the observation process* Y_t *is as in* (3.1) *and that* (N_t) *is Gaussian with continuous covariance kernel* R. *Let*

$$\beta(\cdot, X(\omega)) \in H(R) \quad \textit{for almost all } \omega. \tag{3.6}$$

Then for any $\mathcal{F}_T^X$*-measurable and integrable function* $g(T, X)$,

$$E\left(g(T,X)\,\middle|\mathcal{F}_t^Y\right) = \frac{\displaystyle\int g(T,x)e^{\langle Y,\beta(\cdot,x)\rangle_t - \frac{1}{2}\|\beta(\cdot,x)\|_t^2}\,dP_X(x)}{\displaystyle\int e^{\langle Y,\beta(\cdot,x)\rangle_t - \frac{1}{2}\|\beta(\cdot,x)\|_t^2}\,dP_X(x)}. \tag{3.7}$$

We next consider an important special case from which it can be easily shown that the formula (3.7) extends the Kallianpur–Striebel formula, as well as the one by Kunita.

3.1. An important special case.

Suppose the noise N_t is of the form

$$N_t = \int_0^t F(t,u)dW_u, \tag{3.8}$$

where $F(t,s)$ is continuous on $\{0 \le s \le t \le T\}$. It is easy to check that the covariance function of (N_t), $R(t,s) = \int_0^{t\wedge s} F(t,u)F(s,u)du$ is continuous on $[0,T]\times[0,T]$. Then from the example considered in section 2.1 we have

$$H(R;t) = \left\{\phi : \phi(s) = \int_0^s F(s,u)\phi^*(u)du, \phi^* \in \overline{sp}^{L^2}\{F(s,\cdot)1_{[0,s]}(\cdot), 0 \le s \le t\}\right\} \tag{3.9}$$

with the inner product

$$(\phi_1,\phi_2)_{H(R;t)} = \int_0^t \phi_1^*(u)\phi_2^*(u)du,$$

where

$$\phi_1(s) = \int_0^s F(s,u)\phi_1^*(u)du \;\text{ and }\; \phi_2(s) = \int_0^s F(s,u)\phi_2^*(u)du.$$

Suppose the observation process is given by

$$Y_t = \int_0^t F(t,u)\tilde{h}(u,X_u)du + N_t, \tag{3.10}$$

such that

$$\tilde{h}(\cdot, X_{(\cdot)}) \in \overline{sp}^{L^2}\{F(s,\cdot)1_{[0,s]}(\cdot), 0 \le s \le t\}.$$

Then, by (2.3) and by an argument similar to the one used in (3.4), we have for $\phi(\cdot) = \int_0^{(\cdot)} F(\cdot,u)\phi^*(u)du \in H(R)$,

$$\langle Y,\phi\rangle_t \;=\; \int_0^t \phi^*(u)\tilde{h}(u,X_u)du + \int_0^t \phi^*(u)dW_u \;=\; \int_0^t \phi^*(u)d\hat{Y}_u,$$

where

$$\hat{Y}_s = \int_0^s \tilde{h}(u, X_u)du + W_s, \quad 0 \le s \le T.$$

Hence the Bayes formula (3.7) becomes

$$E\left(g(T,X)\,\middle|\mathcal{F}_t^Y\right) = \frac{\displaystyle\int g(T,x)e^{\int_0^t \tilde{h}(u,x_u)d\hat{Y}_u - \frac{1}{2}\int_0^t |\tilde{h}(u,x_u)|^2 du}\, dP_X(x)}{\displaystyle\int e^{\int_0^t \tilde{h}(u,x_u)d\hat{Y}_u - \frac{1}{2}\int_0^t |\tilde{h}(u,x_u)|^2 du}\, dP_X(x)}. \tag{3.11}$$

Remark. It is now easy to see that the Bayes formula (3.7) is indeed an extension of the Kallianpur–Striebel formula. Take $F(t,u) \equiv 1$ in the model (3.8) and $\tilde{h}$ in the model (3.10) to be $h \in L^2[0,T] \equiv \overline{\mathrm{sp}}^{L^2}\{1_{[0,t]}(\cdot), 0 \le t \le T\}$, so that $N_t = W_t$ and the observation process satisfies the usual model

$$Y_t = \int_0^t h(u, X_u)du + W_t.$$

Note that, in this case, $\hat{Y}_t = Y_t$. Therefore the Bayes formula (3.7) reduces to the Kallianpur–Striebel formula

$$E\left(g(T,X)\,\middle|\mathcal{F}_t^Y\right) = \frac{\displaystyle\int g(T,x)e^{\int_0^t h(u,x_u)dY_u - \frac{1}{2}\int_0^t |h(u,x_u)|^2 du}\, dP_X(x)}{\displaystyle\int e^{\int_0^t h(u,x_u)dY_u - \frac{1}{2}\int_0^t |h(u,x_u)|^2 du}\, dP_X(x)}.$$

Our result also generalizes a similar result by Kunita. We show that in the next section.

4. Kunita's result. In this section we shall derive Kunita's result ([13], Theorem 2.1), when $d = 1$, as a corollary of our result. Suppose the signal process (X_t) is a continuous process taking values in a complete metric space S. Suppose the observation process is given by

$$Y_t = \int_0^t h(X_s)\, ds + N_t, \quad 0 \le t \le T, \tag{4.1}$$

where h is a continuous map from S into $\mathbb{R}$ and the noise process (N_t) is given by

$$N_t = m_t + \int_0^t \psi(t,s)\, dW_s, \quad 0 \le t \le T, \tag{4.2}$$

with $\psi(t,s)$ and m_t satisfying the following three conditions.

Condition 1. $\psi(t,s)$ is continuously differentiable in $(t,s) \in [0,T] \times [0,T]$.

Let $\mathbf{C}_0^r$ be the set of all r-times continuously differentiable functions from $[0,T]$ to $\mathbb{R}$ which vanish at zero. Define $\Psi : \mathbf{C}_0 \equiv \mathbf{C}_0^0 \to \mathbf{C}_0$ such that

$$(\Psi\phi)_t = \int_0^t \psi(t,s)\phi'(s)\, ds \tag{4.3}$$

for $\phi \in \mathbf{C}_0^1$. For general $\phi \in \mathbf{C}_0$, it is extended by integration by parts as

$$(\Psi\phi)_t = \psi(t,t)\phi(t) - \int_0^t \phi(s)\frac{\partial\psi}{\partial s}(t,s)\, ds. \tag{4.4}$$

Let $\mathcal{R}(\Psi) = \{\Psi\phi : \phi \in \mathbf{C}_0\}$. Note that for $f, g \in \mathbf{C}_0$ and $0 \le u \le t \le T$,

$$(\Psi f)_u - (\Psi g)_u = \psi(u,u)\,(f(u) - g(u)) - \int_0^u (f(s) - g(s))\,\frac{\partial\psi}{\partial s}(u,s)\,ds.$$

Hence Ψ is causal in the sense that

$$(\Psi f)_u = (\Psi g)_u \text{ holds for } u \le t \text{ if } f(s) = g(s) \text{ holds for } s \le t. \tag{4.5}$$

Condition 2. The transformation Ψ has a causal inverse transformation $K : \mathcal{R}(\Psi) \to \mathbf{C}_0$ such that $K\Psi\phi = \phi$ holds for all $\phi \in \mathbf{C}_0$. Further, Kg is differentiable whenever $g \in \mathbf{C}_0^1 \cap \mathcal{R}(\Psi)$ and the derivative is in $L^2[0,T]$.

Condition 3. m_t is continuously differentiable in t and it belongs to $\mathcal{R}(\Psi)$.

Set

$$\dot{m}_t = \frac{dm_t}{dt}, \tag{4.6}$$

$$(Lf)_t = \frac{d}{dt}(Kg)_t, \quad \text{where } g_t = \int_0^t f_s\,ds. \tag{4.7}$$

Since $R(s,t) = E(N_sN_t) = \int_0^{t\wedge s} \psi(t,u)\psi(s,u)\,du$ is as in the special case considered in section 2.1, from (3.9) we have

$$H(R) = \left\{ g : g(t) = \int_0^t g^*(u)\psi(t,u)\,du, g^* \in \overline{\text{sp}}^{L^2}\{\psi(t,\cdot)1_{[0,t]}(\cdot) : 0 \le t \le T\} \right\}. \tag{4.8}$$

With the help of Lemma 4.1 we can further simplify the form of $H(R)$.

Lemma 4.1. *If ψ satisfies Condition 1 and Condition 2, then*

$$\overline{sp}^{L^2}\{\psi(t,\cdot)1_{[0,t]}(\cdot) : 0 \le t \le T\} = L^2[0,T].$$

Proof. It suffices to show that if $f \in L^2[0,T]$ is such that $f \perp \psi(t,\cdot)1_{[0,t]}(\cdot)$ for all $t \in [0,T]$, then $f = 0$. So suppose $f \in L^2[0,T]$.

$$\int_0^t \psi(t,s)f(s)\,ds = 0 \quad \text{for all } t$$

$$\Rightarrow \Psi g = 0, \quad \text{where } g(t) = \int_0^t f(s)\,ds$$

$$\Rightarrow g = K\Psi g = 0 \;\Rightarrow \int_0^t f(s)\,ds = 0 \quad \text{for all } t \;\Rightarrow f = 0.$$

Hence the lemma is proved. □

Therefore, from Lemma 4.1 and from (4.8), we have

$$H(R) = \left\{ g : g(t) = \int_0^t g^*(u)\psi(t,u)\,du \text{ for some } g^* \in L^2[0,T] \right\}. \tag{4.9}$$

The following proposition describes a relationship between the spaces $\mathcal{R}(\Psi)$ and $H(R)$.

PROPOSITION 4.2. *Let $\mathcal{R}(\Psi)$ and $H(R)$ be as above. Then*

$$\boldsymbol{C}_0^1 \cap \mathcal{R}(\Psi) \subseteq H(R) \subseteq \mathcal{R}(\Psi).$$

Furthermore, for $g \in H(R)$, $(Kg)_t = \int_0^t g^*(u)\,du$ and if $f \in \boldsymbol{C}_0^1 \cap \mathcal{R}(\Psi)$, then $f^* = L(f')$, where L is given by (4.7).

Proof. Let $g \in H(R)$. From (4.9),

$$g(t) = \int_0^t \psi(t,s) g^*(s)\,ds. \tag{4.10}$$

Considering $\phi = \int_0^{(\cdot)} g^*(u)\,du$, we have $\phi \in \mathbf{C}_0$ and from (4.4),

$$\begin{aligned}
(\Psi\phi)_t &= \psi(t,t)\phi(t) - \int_0^t \frac{\partial\psi}{\partial s}(t,s)\phi(s)\,ds \\
&= \psi(t,t)\int_0^t g^*(u)\,du - \int_0^t \left\{\frac{\partial\psi}{\partial s}(t,s)\int_0^s g^*(u)\,du\right\} ds \\
&= \int_0^t \psi(t,s) g^*(s)\,ds \quad \text{(using integration by parts)} \\
&= g(t) \quad \text{(by (4.10))}.
\end{aligned}$$

Hence $H(R) \subset \mathcal{R}(\Psi)$ and for $g \in H(R)$, $(Kg)_t = \int_0^t g^*(u)\,du$.

On the other hand, for $f \in \mathbf{C}_0^1 \cap \mathcal{R}(\Psi)$, letting $\phi \in \mathbf{C}_0$ to be such that $\Psi\phi = f$, by Condition 2, we have that $\phi = K\Psi\phi = Kf$ is differentiable with $\phi' = L(f') \in L^2[0,T]$. Now

$$\begin{aligned}
f(t) = \Psi\phi(t) &= \psi(t,t)\phi(t) - \int_0^t \phi(s)\frac{\partial\psi}{\partial s}(t,s)\,ds \\
&= \int_0^t \psi(t,s)\phi'(s)\,ds \quad \text{using integration by parts.} \qquad (4.11)
\end{aligned}$$

Hence the proposition follows from (4.9). □

We are now ready to derive the result of Kunita ([13], Theorem 2.1) as a corollary of our result, Theorem 3.2.

THEOREM 4.3 (Kunita). *Suppose the noise process (N_t), given by* (4.2), *satisfies Conditions* $1-3$, *and the observation process (Y_t) is given by* (4.1). *Let P_X denote the probability distribution of X on $C[0,T]$. Assume further that $(\int_0^t h(X_s)\,ds)$ belongs to $\mathcal{R}(\Psi)$ a.s. Then for any measurable function g on S, the signal state space, such that $E(|g(X_t)|) < \infty$*

$$E(g(X_t)|\mathcal{F}_t^Y) = \frac{\int \alpha_t(x,Y) g(x(t))\,dP_X(x)}{\int \alpha_t(x,Y)\,dP_X(x)},$$

where

$$\alpha_t(x,Y) = \exp\left\{\int_0^t L(h(x)+\dot{m})_s\,d\hat{Y}_s - \frac{1}{2}\int_0^t |L(h(x)+\dot{m})_s|^2\,ds\right\}$$

and

$$\hat{Y}_t = \int_0^t Lh(x)_s\,ds + \int_0^t (L\dot{m})_s\,ds + W_t.$$

Remark. To check that $\alpha_t(x,Y)$ in the theorem is in fact $\mathcal{F}_t^Y$-measurable it has been shown in Kunita [13] that $\hat{Y}_t = (KY)_t$ and then the causality of K is used. In the proof given below we show that $\alpha_t(x,Y) = \langle Y, \beta(\cdot,x)\rangle_t$ which proves that it is indeed $\mathcal{F}_t^Y$-measurable.

Proof of Theorem 4.3. Let Ω_0 with $P(\Omega_0)=1$ be such that $\int_0^t h(X_s(\omega))\,ds \in \mathcal{R}(\Psi)$ for all $\omega \in \Omega_0$. Fix $\omega \in \Omega_0$. Since $h(X_s(\omega))$ is continuous in $s \in [0,T]$, $\int_0^{(\cdot)} h(X_s(\omega))\,ds \in \mathbf{C}_0^1 \cap \mathcal{R}(\Psi)$. So, by Proposition 4.2, $\int_0^{(\cdot)} h(X_s(\omega))\,ds$ belongs to $H(R)$. Hence $(\int_0^t h(X_s)\,ds)$ belongs to $H(R)$ a.s. with $(\int_0^{(\cdot)} h(X_s)\,ds)^*(t) = (Lh(x))_t$. Similarly, since by Condition 3, $m \in \mathbf{C}_0^1 \cap \mathcal{R}(\Psi)$, we have $m \in H(R)$ with $m^* = L\dot{m}$. Rewriting the observation model (4.1), we have

$$\begin{aligned} Y_t &= \int_0^t h(X_s)ds + m_t + \int_0^t \psi(t,s)dW_s \\ &= \int_0^t L(h(x)+\dot{m})_s \psi(t,s)ds + \int_0^t \psi(t,s)dW_s. \end{aligned}$$

The theorem then follows from the special case considered in section 3.1 with $F(t,s) = \psi(t,s)$ and $\tilde{h} = L(h(x)+\dot{m}) \in L^2[0,T]$. □

5. Fractional Brownian motion noise process. Suppose the observation process is given by

$$Y_t = \int_0^t h(X_u)du + B_H(t), \quad 0 \le t \le T, \tag{5.1}$$

where $B_H(t)$ is an fBm with Hurst parameter $H \in (\frac{1}{2},1)$ and is independent of the signal process (X_t). Here $R(s,t) = E[B_H(t)B_H(s)] = 1/2\{|t|^{2H} + |s|^{2H} - |t-s|^{2H}\}$. Assume that $h(u) \equiv h(X_u)$ is continuous a.s. To apply Theorem 3.2 we shall need the following lemma about the representation of functions in $H(R)$. It can be obtained from Theorem 4.4 of Barton and Poor [3], where a characterization of the functions in $H(R)$ is given. However, it takes some effort to relate it to our notation and concepts used in Theorem 3.2. We therefore give a self-contained short proof below.

LEMMA 5.1. *Let $(B_H(t), 0 \le t \le T)$ be an fBm with $H \in (\frac{1}{2},1)$ and the covariance function $R(s,t)$. For any continuous function $c(\cdot)$ on $[0,\tau]$ $(\tau > 0)$, suppose $g_c^\tau(\cdot)$ satisfies the equation (see Carleman* [7]*)*

$$\int_0^\tau g_c^\tau(u)H(2H-1)|v-u|^{2H-2}du = c(v), \quad 0 \le v \le \tau. \tag{5.2}$$

Suppose $a(\cdot)$ is continuous on $[0,T]$. Then $\int_0^{(\cdot)} a(u)du \in H(R)$ with

$$\left\langle \int_0^{(\cdot)} a(u)du, B_H \right\rangle_t = \int_0^t g_a^t(u)dB_H(u) \quad \textit{and} \quad \left\| \int_0^{(\cdot)} a(u)du \right\|_t^2 = \int_0^t g_a^t(u)a(u)du.$$

Proof. Recall that there exists a congruence between the RKHS, $H(R)$, and $\overline{\mathrm{sp}}^{L^2}\{B_H(s) : s \in [0,T]\}$ under which $R(\cdot,t) \mapsto B_H(t)$. Clearly, $\int_0^T g_a^T(u)dB_H(u) \in \overline{\mathrm{sp}}^{L^2}\{B_H(s) : 0 \le s \le T\}$. Hence there exists $\tilde{g} \in H(R)$ such that the image of $\tilde{g}$, under the congruence, is $\int_0^T g_a^T(u)dB_H(u)$. Then for $0 \le s \le T$, by (5.2),

$$\begin{aligned}\tilde{g}(s) = (R(\cdot, s), \tilde{g})_{H(R)} \ &= \ E\left(B_H(s)\int_0^T g_a^T(u)dB_H(u)\right)\\ &= \int_0^s\int_0^T g_a^T(u)H(2H-1)|v-u|^{2H-2}dudv \ = \ \int_0^s a(v)dv.\end{aligned}$$

This proves that $\int_0^{(\cdot)} a(u)du \in H(R)$ and following the notation of section 2 we have $\langle \int_0^{(\cdot)} a(u)du, B_H\rangle = \int_0^T g_a^T(u)dB_H(u)$. Exactly in the same way it follows that $\langle \int_0^{(\cdot)} a(u)du, B_H\rangle_t = \int_0^t g_a^t(u)dB_H(u)$. Finally,

$$\begin{aligned}\left\|\int_0^{(\cdot)} a(u)du\right\|_t^2 &= E\left(\int_0^t g_a^t(u)dB_H(u)\int_0^t g_a^t(u)dB_H(u)\right)\\ &= \int_0^t\int_0^t g_a^t(u)g_a^t(v)H(2H-1)|u-v|^{2H-2}dvdu\\ &= \int_0^t g_a^t(u)a(u)du \quad \text{from (5.2).} \qquad \Box\end{aligned}$$

Clearly, R is continuous on $[0,T]\times[0,T]$. Then from Theorem 3.1, under a suitable change of measure (Y_t) becomes an fBm. Therefore, from the Bayes formula (3.7) with $\beta(t,X) = \int_0^t h(u)du$ and $N_t = B_H(t)$, and from Lemma 5.1, we have

$$(5.3)\quad E\left[f(X_t)\,|\mathcal{F}_t^Y\right] = \frac{\displaystyle\int f(x_t)\exp\left\{\int_0^t g_h^t(u)dY(u) - \frac{1}{2}\int_0^t g_h^t(u)h(u)du\right\}dP_X(x)}{\displaystyle\int \exp\left\{\int_0^t g_h^t(u)dY(u) - \frac{1}{2}\int_0^t g_h^t(u)h(u)du\right\}dP_X(x)}.$$

When the signal process is actually a random variable η (independent of the noise process $B_H(t)$) such that $h(u) = \eta a(u)$, where a is a continuous (deterministic) function, then using the fact that for a constant k, $g_{ka}^t = kg_a^t$, from (5.3) we have

$$(5.4)\quad E\left[f(\eta)\,|\mathcal{F}_t^Y\right] = \frac{\displaystyle\int f(x)\exp\left\{x\int_0^t g_a^t(u)dY(u) - \frac{1}{2}x^2\int_0^t g_a^t(u)a(u)du\right\}dP_\eta(x)}{\displaystyle\int \exp\left\{x\int_0^t g_a^t(u)dY(u) - \frac{1}{2}x^2\int_0^t g_a^t(u)a(u)du\right\}dP_\eta(x)}.$$

If we further assume that η is a Gaussian random variable with mean η_0 and variance γ_0, then η being independent of $(B_H(t))$, we have (η, Y) jointly Gaussian. Hence the conditional distribution of η given $\mathcal{F}_t^Y$ is also Gaussian with mean $E(\eta|\mathcal{F}_t^Y) = \hat{\eta}_t$, say, and variance $E\left((\eta-\hat{\eta}_t)^2\,|\mathcal{F}_t^Y\right) = \hat{\gamma}_t$, say. Then

$$E\left(e^{\alpha\eta}\,|\mathcal{F}_t^Y\right) = \exp\left\{\alpha\hat{\eta} + \frac{1}{2}\alpha^2\hat{\gamma}_t\right\}.$$

Now from (5.4), taking $f(x) = e^{\alpha x}$, we have

$$(5.5)\quad \begin{aligned}&E\left[e^{\alpha\eta}\,|\mathcal{F}_t^Y\right]\\ &= \frac{\displaystyle\int e^{\alpha x}\exp\left\{x\int_0^t g_a^t(u)dY(u) - \frac{1}{2}x^2\int_0^t g_a^t(u)a(u)du\right\}\phi(x;\eta_0,\gamma_0)dx}{\displaystyle\int \exp\left\{x\int_0^t g_a^t(u)dY(u) - \frac{1}{2}x^2\int_0^t g_a^t(u)a(u)du\right\}\phi(x;\eta_0,\gamma_0)dx},\end{aligned}$$

where $\phi(x;\eta_0,\gamma_0)$ is the density of a Gaussian random variable with mean η_0 and variance γ_0.

Let us consider the numerator of the right-hand side of (5.5):

$$\begin{aligned}
&\int e^{x\left(\alpha+\int_0^t g_a^t(u)dY(u)\right)-\frac{1}{2}x^2\int_0^t g_a^t(u)a(u)du}\ \frac{1}{\sqrt{2\pi\gamma_0}}\ e^{-\frac{1}{2\gamma_0}(x-\eta_0)^2}dx\\
&\quad=\frac{1}{\sqrt{2\pi\gamma_0}}\int e^{-\frac{1}{2}\left[x^2\left(\gamma_0^{-1}+\int_0^t g_a^t(u)a(u)du\right)-2x\left(\alpha+\gamma_0^{-1}+\eta_0\int_0^t g_a^t(u)dY(u)\right)+\gamma_0^{-1}\eta_0^2\right]}dx\\
&\quad=\frac{1}{\sqrt{2\pi\gamma_0}}\int e^{-\frac{1}{2\gamma_t}\left[x^2-2x(\alpha+m_t)\gamma_t+(\alpha+m_t)^2\gamma_t^2\right]}e^{-\frac{1}{2}\gamma_0^{-1}\eta_0^2+\frac{1}{2}\gamma_t(\alpha+m_t)^2}dx,
\end{aligned}$$

$$\text{where } \gamma_t^{-1}=\gamma_0^{-1}+\int_0^t g_a^t(u)a(u)du \ \text{ and } \ m_t=\gamma_0^{-1}\eta_0+\int_0^t g_a^t(u)dY(u) \tag{5.6}$$

$$=\sqrt{\gamma_0^{-1}\gamma_t}\ e^{-\frac{1}{2}\gamma_0^{-1}\eta_0^2+\frac{1}{2}\gamma_t(\alpha+m_t)^2}. \tag{5.7}$$

Putting $\alpha=0$ in (5.7) we get the denominator of the right-hand side of (5.5):

$$\text{Denominator } =\sqrt{\gamma_0^{-1}\gamma_t}\ e^{-\frac{1}{2}\gamma_0^{-1}\eta_0^2+\frac{1}{2}\gamma_t m_t^2}.$$

Therefore, from (5.5), we have

$$E\left[e^{\alpha\eta}\left|\mathcal{F}_t^Y\right.\right]=e^{\frac{1}{2}\left[\gamma_t(\alpha+m_t)^2-\gamma_t m_t^2\right]}=e^{\frac{1}{2}\gamma_t\alpha(\alpha+2m_t)}.$$

Collecting the coefficients of α and α^2 and using (5.6), we get

$$\hat{\eta}_t=\gamma_t m_t=\gamma_t\left(\gamma_0^{-1}\eta_0+\int_0^t g_a^t(u)dY(u)\right),$$

$$\hat{\gamma}_t=\gamma_t=\left(\gamma_0^{-1}+\int_0^t g_a^t(u)a(u)du\right)^{-1}.$$

Note that these equations for the filter are exactly the same as those obtained by Le Breton [6].

Remark. Recently, Le Breton [5] considered the parametric estimation problem in a simple deterministic regression model setup with the fBm noise process. Our general Bayes formula can be used to study the parametric estimation problem in a more general setup with the fBm noise process, as done in Liptser and Shiryayev [15] in parameter estimation of the drift coefficient for diffusion-type processes with the Wiener noise. We leave that for a future note.

6. Ornstein–Uhlenbeck noise process. Although the use of the Wiener process as noise produces elegant, powerful mathematical techniques to calculate the optimal filter, one of the main criticisms against it (as expressed by Balakrishnan [2]) is from the practical point of view. Since the sample paths of a Wiener process are of unbounded variation with probability one, the actual data samples have zero probability of occurring and hence the results obtained cannot be instrumented. On the other hand, it has been argued by Nelson [16] that the Ornstein–Uhlenbeck (dispersion) process is natural to consider as an approximation to the Wiener process and the Ornstein–Uhlenbeck processes are realizable. In this section we consider the filtering problem corresponding to the Ornstein–Uhlenbeck noise process and show that it leads to the conventional theory with the Wiener noise process.

Suppose $v(t)$ is an Ornstein–Uhlenbeck velocity process satisfying the stochastic differential equation

$$dv(t) = -\beta v(t)dt + \sigma dW(t) \qquad (\beta > 0,\ \sigma > 0) \tag{6.1}$$

with the initial value $v(0) = 0$. Consider the Ornstein–Uhlenbeck (dispersion) process given by

$$\xi(t) = \int_0^t v(s)ds. \tag{6.2}$$

It is easy to see that if β and σ tend to infinity in such a way that $\sigma^2/\beta^2 \to 1$, then $\xi(t)$ converges in distribution to the standard Wiener process. See, for example, Theorem 9.5 of Nelson [16].

Now suppose the noise process (N_t) is given by an Ornstein–Uhlenbeck process so that, from (6.2) and (6.1), we have

$$N_t = \int_0^t \sigma \int_0^s \exp\{-\beta(s-r)\}dW_r ds = \int_0^t \frac{\sigma}{\beta}\left(1 - e^{-\beta(t-s)}\right) dW_s.$$

Also, suppose that the signal process X is independent of W and the observation process is given by

$$Y_t^{\beta,\sigma} = \int_0^t h(X_u)du + N_t, \tag{6.3}$$

where $h(u) \equiv h(X_u)$ is differentiable in $[0,T]$ and $h'(u) \in L^2[0,T]$.

Then, the covariance $R(t,s)$ of (N_t) is given by

$$R(s,t) = E(N_s N_t) = \int_0^{t\wedge s} F(t,u)F(s,u)du,$$

where

$$F(t,u) = \frac{\sigma}{\beta}\left(1 - e^{-\beta(t-u)}\right), \quad 0 \le u \le t \le T. \tag{6.4}$$

Also, it is easy to see that

$$\overline{sp}^{L^2}\{F(t,\cdot)1_{[0,t]}(\cdot), 0 \le t \le T\} = L^2[0,T].$$

This is because if $f \in L^2[0,T]$ such that $f \perp F(t,\cdot)1_{[0,t]}(\cdot)$ for all $0 \le t \le T$, then

$$\begin{aligned}
&\int_0^t f(u)F(t,u)du = 0 \quad \text{ for all } t\\
&\Rightarrow \int_0^t f(u)\frac{\sigma}{\beta}\left(1 - e^{-\beta(t-u)}\right) du = 0 \quad \text{ for all } t\\
&\Rightarrow \int_0^t f(u)du - e^{-\beta t}\int_0^t e^{\beta u} f(u)du = 0 \quad \text{ for all } t\\
&\Rightarrow f(t) + \beta e^{-\beta t}\int_0^t e^{\beta u} f(u)du - e^{-\beta t}e^{\beta t} f(t) = 0 \quad \text{ almost everywhere (a.e.) } [t]\\
&\Rightarrow \int_0^t e^{\beta u} f(u)du = 0 \quad \text{ a.e. } [t]\\
&\Rightarrow f(t) = 0 \quad \text{ a.e. } [t].
\end{aligned}$$

Hence from (3.9) we have

$$H(R) = \left\{ g : g(s) = \int_0^s F(s,u) g^*(u) du \text{ for some } g^* \in L^2[0,T] \right\}.$$

It is also easy to check (assuming, without loss of generality, $h(X_0) = 0$) that

$$\int_0^t h(X_u) du = \int_0^t F(t,u) \left[\frac{\beta}{\sigma} h(X_u) + \frac{1}{\sigma} h'(u) \right] du.$$

Hence the noise process and the observation process are as in the special case considered in section 3.1, that is, N_t is of the form (3.8) with $F(t,s)$ given by (6.4) and $Y_t^{\beta,\sigma}$ is of the form (3.10) with

$$\tilde{h}(u, X_u) = \frac{\beta}{\sigma} h(X_u) + \frac{1}{\sigma} h'(u).$$

In this case, therefore, from (3.11), we have

$$\nu_t^{\beta,\sigma}(f)(Y^{\beta,\sigma}) := E(f(X_t)|\mathcal{F}_t^{Y^{\beta,\sigma}}) = \frac{\int f(x_t) \alpha_t^{\beta,\sigma}(x, Y^{\beta,\sigma}) P_X(dx)}{\int \alpha_t^{\beta,\sigma}(x, Y^{\beta,\sigma}) P_X(dx)},$$

where

$$\begin{aligned} &\alpha_t^{\beta,\sigma}(x, Y^{\beta,\sigma}) \\ &= \exp\left\{ \int_0^t \left[\frac{\beta}{\sigma} h(x_u) + \frac{1}{\sigma} h'(u) \right] d\tilde{Y}_u^{\beta,\sigma} - \frac{1}{2} \int_0^t \left[\frac{\beta}{\sigma} h(x_u) + \frac{1}{\sigma} h'(u) \right]^2 du \right\} \end{aligned} \tag{6.5}$$

and

$$\tilde{Y}_t^{\beta,\sigma} = \int_0^t \left[\frac{\beta}{\sigma} h(x_u) + \frac{1}{\sigma} h'(u) \right] du + W_t. \tag{6.6}$$

Now suppose that ν_t is the classical filter based on the observation process

$$Y_t = \int_0^t h(X_s) ds + W_t.$$

Recall from the Kallianpur–Striebel formula that

$$\nu_t(f)(Y) := E(f(X_t)|\mathcal{F}_t^Y) = \frac{\int f(x_t) \alpha_t(x, Y) P_X(dx)}{\int \alpha_t(x, Y) P_X(dx)},$$

where

$$\alpha_t(x, Y) = \exp\left\{ \int_0^t h(x_u) dY_u - \frac{1}{2} \int_0^t h^2(x_u) du \right\}. \tag{6.7}$$

The following result shows that the conventional filter can be approximated by suitable filters corresponding to the Ornstein–Uhlenbeck noise process.

Theorem 6.1. *Suppose h satisfies the following condition*

$$E\left[\exp\left\{ 7 \int_0^T h^2(X_u) du + \int_0^T (h'(u))^2 du \right\} \right] < \infty. \tag{6.8}$$

Then for bounded function f, as $\beta,\ \sigma \to \infty$, with $\sigma^2/\beta^2 \to 1$,

$$\nu_t^{\beta,\sigma}(f)(Y^{\beta,\sigma}) \longrightarrow \nu_t(f)(Y) \qquad a.s. \tag{6.9}$$

through an appropriate subsequence.

Proof. Denote by $a_t(\beta,\sigma)$ and a_t the expressions in the curly brackets in (6.5) and (6.7), respectively. Then as $\beta\to\infty,\sigma\to\infty$ such that $\sigma^2/\beta^2\to 1$, we have

$$\begin{aligned} a_t(\beta,\sigma) &= \int_0^t \left[\frac{\beta}{\sigma}h(x_u)+\frac{1}{\sigma}h'(u)\right]dW_u+\frac{1}{2}\int_0^t\left[\frac{\beta}{\sigma}h(x_u)+\frac{1}{\sigma}h'(u)\right]^2du \\ &= \frac{\beta}{\sigma}\int_0^t h(x_u)dW_u+\frac{1}{\sigma}\int_0^t h'(u)dW_u+\frac{1}{2}\int_0^t\left[\frac{\beta}{\sigma}h(x_u)+\frac{1}{\sigma}h'(u)\right]^2du \\ &\longrightarrow \int_0^t h(x_u)dW_u+\frac{1}{2}\int_0^t |h(x_u)|^2du \ = \ a_t \ \text{ a.e. } x\ [P_X] \text{ and a.e. } [P]. \end{aligned} \tag{6.10}$$

Hence it is enough to show that

$$\int \alpha_t^{\beta,\sigma}(x,Y^{\beta,\sigma})P_X(dx)\longrightarrow\int\alpha_t(x,Y)P_X(dx) \quad \text{in } \ L^1, \tag{6.11}$$

for L^1-convergence will imply a.s. convergence through a subsequence and then the theorem will follow from Scheffe's theorem.

It is easy to check that for any numbers a and b,

$$|e^a-e^b|\le|a-b|\cdot\max\left(e^{|a|},e^{|b|}\right).$$

Then

$$\begin{aligned} &E\left(\left|\int\alpha_t^{\beta,\sigma}(x,Y^{\beta,\sigma})P_X(dx)-\int\alpha_t(x,Y)P_X(dx)\right|\right) \\ &\quad\le E\left(\int|\exp\{a_t(\beta,\sigma)\}-\exp\{a_t\}|P_X(dx)\right) \\ &\quad\le E\left(\int|a_t(\beta,\sigma)-a_t|\cdot\max\left(e^{|a_t(\beta,\sigma)|},e^{|a_t|}\right)P_X(dx)\right) \\ &\quad\le\left\{\int E\left(|a_t(\beta,\sigma)-a_t|^2\right)P_X(dx)\cdot\int E\left(e^{2|a_t(\beta,\sigma)|}+e^{2|a_t|}\right)P_X(dx)\right\}^{1/2} \\ &\quad\le\left(\int I_1P_X(dx)\right)^{1/2}\left(\int I_2P_X(dx)\right)^{1/2}, \ \text{ say.} \end{aligned} \tag{6.12}$$

Then

$$\begin{aligned} I_1 &= E\left(|a_t(\beta,\sigma)\}-a_t|^2\right) \\ &= E\Bigg(\bigg|\int_0^t\left\{\left(\frac{\beta}{\sigma}-1\right)h(x_u)+\frac{1}{\sigma}h'(u)\right\}dW_u \\ &\qquad\qquad +\frac{1}{2}\int_0^t\left\{\left(\frac{\beta}{\sigma}h(x_u)+\frac{1}{\sigma}h'(u)\right)^2-h^2(x_u)\right\}du\bigg|^2\Bigg) \\ &\le 2E\left(\left|\int_0^t\left\{\left(\frac{\beta}{\sigma}-1\right)h(x_u)+\frac{1}{\sigma}h'(u)\right\}dW_u\right|^2\right) \end{aligned}$$

$$
\begin{aligned}
&\qquad + 2\cdot\frac{1}{4}\left|\int_0^t \left(\frac{2\beta^2}{\sigma^2}-1\right) h^2(x_u)du + \int_0^t \frac{2}{\sigma^2}\left(h'(u)\right)^2 du\right|^2 \\
&\le 2\int_0^t \left\{\left(\frac{\beta}{\sigma}-1\right)h(x_u) + \frac{1}{\sigma}h'(u)\right\}^2 du \\
&\qquad + \frac{1}{2}\cdot 2\left\{\left(\frac{2\beta^2}{\sigma^2}-1\right)^2\left(\int_0^t h^2(x_u)du\right)^2 + \frac{4}{\sigma^4}\left(\int_0^t \left(h'(u)\right)^2 du\right)^2\right\} \\
&\le 4\left(\frac{\beta}{\sigma}-1\right)^2 \int_0^t h^2(x_u)du + \frac{4}{\sigma^2}\int_0^t \left(h'(u)\right)^2 du \\
&\qquad + \left(\frac{2\beta^2}{\sigma^2}-1\right)^2\left(\int_0^t h^2(x_u)du\right)^2 + \frac{4}{\sigma^4}\left(\int_0^t \left(h'(u)\right)^2 du\right)^2 .
\end{aligned}
$$

Hence, from (6.8), it follows that

$$
\int I_1 P_X(dx) \longrightarrow 0 \text{ as } \beta,\ \sigma\to\infty,\ \text{with } \frac{\sigma^2}{\beta^2}\to 1. \tag{6.13}
$$

Now, using the fact that for a normal random variable Z with zero mean and variance σ^2, $E(e^{|Z|}) \le 2e^{\sigma^2/2}$, we have

$$
\begin{aligned}
I_2 &= E\left(e^{2|a_t(\beta,\sigma)|} + e^{2|a_t|}\right) \\
&= E\left(\exp\left\{\left|2\int_0^t \left[\frac{\beta}{\sigma}h(x_u) + \frac{1}{\sigma}h'(u)\right]dW_u + \int_0^t \left[\frac{\beta}{\sigma}h(x_u)+\frac{1}{\sigma}h'(u)\right]^2 du\right|\right\}\right) \\
&\qquad + E\left(\exp\left\{\left|2\int_0^t h(x_u)dW_u + \int_0^t |h(x_u)|^2 du\right|\right\}\right) \\
&\le 2\exp\left\{3\int_0^t \left[\frac{\beta}{\sigma}h(x_u)+\frac{1}{\sigma}h'(u)\right]^2 du\right\} + 2\exp\left\{3\int_0^t |h(x_u)|^2 du\right\} \\
&\le 2\exp\left\{\frac{6\beta^2}{\sigma^2}\int_0^t h^2(x_u)du + \frac{6}{\sigma^2}\int_0^t \left(h'(u)\right)^2 du\right\} + 2\exp\left\{3\int_0^t h^2(x_u)du\right\}.
\end{aligned}
$$

Therefore, from (6.8), we have for large σ and β, $\int I_2 P_X(dx)$ is bounded and consequently, (6.11) follows from (6.12) and (6.13). □

Remark. Note that the condition (6.8) in Theorem 6.1 will hold if one assumes that the functions $h(\cdot)$ and $h'(\cdot)$ are bounded.

Next we address the issue of implementation of the results obtained by considering the Ornstein–Uhlenbeck dispersion process as the observation noise process. We would like to obtain a Zakai-type evolution equation for the so-called unnormalized conditional density of X_t given the observations up to time "t." So let us assume that the signal process X_t is a Markov process.

First, we shall prove the following properties of $\tilde{Y}_t \equiv \tilde{Y}_t^{\beta,\sigma}$ and its relationship with $Y_t \equiv Y_t^{\beta,\sigma}$.

LEMMA 6.2. *Suppose $\tilde{Y}_t$ is given by* (6.6). *Suppose Q is defined by*

$$
dP = \exp\left\{\int_0^T \left[\frac{\beta}{\sigma}h(X_u)+\frac{1}{\sigma}h'(u)\right]d\tilde{Y}_u - \frac{1}{2}\int_0^T \left[\frac{\beta}{\sigma}h(X_u)+\frac{1}{\sigma}h'(u)\right]^2 du\right\} dQ.
$$

Then

(i) *under Q, $\tilde{Y}_t$ is a Wiener process,*

(ii) $\overline{sp}^{L^2(Q)}\{Y_s, 0 \le s \le t\} = \overline{sp}^{L^2(Q)}\{\tilde{Y}_s, 0 \le s \le t\}$,

(iii) $\mathcal{F}_t^Y = \mathcal{F}_t^{\tilde{Y}}$.

Proof. Clearly (iii) follows from (ii) as, under Q, (Y_t) and $(\tilde{Y}_t)$ are Gaussian. (i) follows from Lemma 11.3.1 of [10] since (X_t) is independent of (W_t). For (ii) note that

$$\begin{aligned} Y_t &= \int_0^t h(X_u)du + \int_0^t F(t,u)dW_u \\ &= \int_0^t F(t,u)\left[\frac{\beta}{\sigma}h(X_u) + \frac{1}{\sigma}h'(u)\right]du + \int_0^t F(t,u)dW_u \\ &= \int_0^t F(t,u)d\tilde{Y}_u. \end{aligned} \tag{6.14}$$

Hence

$$\overline{\mathrm{sp}}^{L^2(Q)}\{Y_s, 0 \le s \le t\} \subset \overline{\mathrm{sp}}^{L^2(Q)}\{\tilde{Y}_s, 0 \le s \le t\}.$$

To show the reverse inclusion suppose $\xi \in \overline{\mathrm{sp}}^{L^2(Q)}\{\tilde{Y}_s, 0 \le s \le t\}$ and $E_Q(\xi Y_s) = 0$ for all $0 \le s \le t$. Since $\tilde{Y}_t$, under Q, is a Wiener process we can express ξ as an Ito integral, say, $\xi = \int_0^t \phi(u)d\tilde{Y}_u$. Then

$$\begin{aligned} &E_Q(\xi Y_s) = 0 \quad \text{for all } 0 \le s \le t \\ &\Rightarrow E_Q\left(\int_0^t \phi(u)d\tilde{Y}_u \int_0^s F(s,u)d\tilde{Y}_u\right) = 0 \quad \text{for all } 0 \le s \le t \\ &\Rightarrow \int_0^s \phi(u)F(s,u)du = 0 \quad \text{for all } 0 \le s \le t \\ &\Rightarrow \int_0^s \phi(u)\frac{\sigma}{\beta}\left(1 - e^{-\beta(s-u)}\right)du = 0 \quad \text{for all } 0 \le s \le t \\ &\Rightarrow \int_0^s \phi(u)du - e^{-\beta s}\int_0^s \phi(u)e^{\beta u}du = 0 \quad \text{for all } 0 \le s \le t \\ &\Rightarrow \text{(by differentiating) } \beta e^{-\beta s}\int_0^s \phi(u)e^{\beta u}du = 0 \quad \text{a.e. } s \in [0,t] \\ &\Rightarrow \phi(u) = 0 \quad \text{a.e. } u \in [0,t]. \end{aligned}$$

Hence

$$\overline{\mathrm{sp}}^{L^2(Q)}\{\tilde{Y}_s, 0 \le s \le t\} \subset \overline{\mathrm{sp}}^{L^2(Q)}\{Y_s, 0 \le s \le t\}.$$

This completes the proof of part (ii). □

Because of property (iii) of Lemma 6.2, the filter based on $\{Y_s, 0 \le s \le t\}$ will coincide with the filter based on $\{\tilde{Y}_s, 0 \le s \le t\}$, where

$$\tilde{Y}_t = \int_0^t \left[\frac{\beta}{\sigma}h(X_u) + \frac{1}{\sigma}h'(u)\right]du + W_t.$$

We shall, however, consider the observation process to be given by

$$\hat{Y}_t = \int_0^t \frac{\beta}{\sigma}h(X_u)du + W_t, \tag{6.15}$$

which, for large σ, will approximate $\tilde{Y}_t$. We can then use the classical theory with the Wiener noise process to obtain the following result.

Suppose A_t with domain $\mathcal{D}$ is the generator of the Markov signal process (X_t). Denote by $\Phi(u,t)$ the unnormalized conditional density of X_t given $\mathcal{F}_t^{\hat{Y}}$. Then

$$\Phi(u,t) = \Phi(u,0) + \int_0^t A_s^* \Phi(u,s) ds + \int_0^t \left[\frac{\beta}{\sigma} h(X_s) \right] \Phi(u,s) d\hat{Y}_s, \tag{6.16}$$

where A_s^* is the formal adjoint of A_s.

Now note that from (6.14) and the form (6.4) of F we have

$$\begin{aligned} Y_t = \frac{\sigma}{\beta} \int_0^t \left[1 - e^{-\beta(t-u)} \right] d\tilde{Y}_u &= \frac{\sigma}{\beta} \left[\tilde{Y}_t - e^{-\beta t} \int_0^t e^{\beta u} d\tilde{Y}_u \right] \\ = \frac{\sigma}{\beta} \left[\tilde{Y}_t - e^{-\beta t} \left\{ e^{\beta t} \tilde{Y}_t - \int_0^t \beta e^{\beta u} \tilde{Y}_u du \right\} \right] &= \sigma e^{-\beta t} \int_0^t e^{\beta u} \tilde{Y}_u du. \end{aligned}$$

Hence,

$$y_t := \frac{d}{dt} Y_t \ = \ \sigma e^{-\beta t} (-\beta) \int_0^t e^{\beta u} \tilde{Y}_u du + \sigma e^{-\beta t} e^{\beta t} \tilde{Y}_t \ = \ \sigma \tilde{Y}_t - \beta Y_t,$$

that is,

$$\hat{Y}_t \ = \ \tilde{Y}_t - \frac{1}{\sigma} \int_0^t h'(u) du \ = \ \frac{1}{\sigma} \left\{ y_t - \int_0^t h'(u) du \right\} + \frac{\beta}{\sigma} Y_t.$$

Therefore ignoring the first term in the expression for $\hat{Y}_t$ above, which is of the order of σ^{-1}, we see that the solution of the Zakai equation (6.16), for large σ, can be approximated by the solution of the following ordinary partial differential equation

$$\frac{d}{dt} \Phi(u,t) = A_t^* \Phi(u,t) + \left(\frac{\beta}{\sigma} \right)^2 h(X_t) \ \Phi(u,t) \ y_t.$$

Acknowledgment. We thank the referees for their insightful suggestions, which led to the improvement of our presentation.

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels,* Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] A. V. Balakrishnan, *Nonlinear white noise theory,* in Multivariate Analysis V, P. R. Krishnaiah, ed., North-Holland, Amsterdam, 1980, pp. 97–109.

[3] R. J. Barton and H. V. Poor, *Signal detection in fractional Gaussian noise,* IEEE Trans. Inform. Theory, 34 (1988), pp. 943–959.

[4] J. Beran, *Statistics for Long-Memory Processes,* Chapman & Hall, New York, 1994.

[5] A. Le Breton, *Filtering and parameter estimation in a simple linear system driven by a fractional Brownian motion,* Stat. Probab. Lett., 38 (1998), pp. 263–274.

[6] A. Le Breton, *A Girsanov-type approach to filtering in a simple linear system with fractional Brownian noise,* C. R. Acad. Sci. Paris Ser. I Math., 326 (1998), pp. 997–1002 (in French).

[7] T. Carleman, *Über die Abelsche Integralgleichung mit konstanten Integrationsgrenzen,* Math. Z., 15 (1922), pp. 111–120.

[8] M. Fujisaki, G. Kallianpur, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem,* Osaka J. Math., 9 (1972), pp. 19–40.

[9] M. Hitsuda, *Representation of Gaussian processes equivalent to Wiener process,* Osaka J. Math., 5 (1968), pp. 299–312.

[10] G. KALLIANPUR, *Stochastic Filtering Theory,* Springer-Verlag, New York, Heidelberg, Berlin, 1980.

[11] G. KALLIANPUR AND R. L. KARANDIKAR, *White Noise Theory of Prediction, Filtering and Smoothing,* Gordon Breach Science Publishers, New York, 1988.

[12] G. KALLIANPUR, AND C. STRIEBEL, *Estimations of stochastic processes: Arbitrary system process with additive white noise observation errors,* Ann. Math. Statist., 39 (1968), pp. 785–801.

[13] H. KUNITA, *Representation and stability of nonlinear filters associated with Gaussian noises,* in Stochastic Processes: A Festschrift in Honour of Gopinath Kallianpur, S. Cambanis et al., eds., Springer-Verlag, New York, 1993, pp. 201–210.

[14] H. KUSHNER, *Dynamical equations for optimal nonlinear filtering,* J. Differential Equations, 3 (1967), pp. 179–190.

[15] R. S. LIPTSER AND A. N. SHIRYAYEV, *Statistics of Random Processes,* Springer-Verlag, New York, 1978.

[16] E. NELSON, *Dynamical Theories of Brownian Motion,* Princeton University Press, Princeton, NJ, 1967.

[17] E. PARZEN, *Statistical inference on time series by Hilbert space methods,* I, in Time Series Analysis Papers, Holden-Day, San Francisco, 1967, pp. 251–382.

[18] E. PARZEN, *Regression analysis of continuous parameter time series,* in Proceedings of the 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I, Univ. California Press, Berkeley, CA, 1961, pp. 469–489.

[19] P. K. MANDAL, *Topics in Stochastic Nonlinear Filtering,* Ph.D. dissertation, University of North Carolina at Chapel Hill, Chapel Hill, NC, 1997.

[20] M. ZAKAI, *On the optimal filtering of diffusion processes,* Z. Wahrscheinlichkeitstheorie und Verw. Gebiete., 11 (1969), pp. 230–243.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 872–899

ON A PERTURBATION APPROACH FOR THE ANALYSIS OF STOCHASTIC TRACKING ALGORITHMS[*]

RAFIK AGUECH[†], ERIC MOULINES[‡], AND PIERRE PRIOURET[†]

Abstract. In this paper, a perturbation expansion technique is introduced to decompose the tracking error of a general adaptive tracking algorithm in a linear regression model. This method results in a tracking error bound and tight approximate expressions for the moments of the tracking error. These expressions allow the evaluation, both qualitatively and quantitatively, of the impact of several factors on the tracking error performance.

Key words. adaptive tracking algorithms, expansion techniques, performance analysis

AMS subject classifications. 60G35, 93E10, 93E12, 93E15, 93E23

PII. S0363012998333852

1. Introduction.

An important issue in system identification, signal processing, and automatic control is that of tracking the parameter variations of a linear dynamical system with observable stochastic inputs and noise-corrupted outputs:

$$y_t = \phi_t^T \theta_t + v_t, \quad t \geq 0, \tag{1}$$

where $\{y_t\}_{t\geq 0}$ and $\{v_t\}_{t\geq 0}$ are, respectively, the scalar observation and noise, and $\{\phi_t\}_{t\geq 0}$ and $\{\theta_t\}_{t\geq 0}$ are the d-dimensional stochastic regressors and the unknown time-varying parameter. This model encompasses many different applications, including channel equalization, time delay estimation, and echo cancellation (see [34] for other signal processing and automatic control applications; see also [18]). To track the variations of the parameter, it is customary to use a recursive algorithm for updating an estimate $\hat{\theta}_t$ of the parameter (see, for example, [21], [22], [23], [34], [13], [18], and the references therein). These algorithms may take many different forms depending on what one is willing to assume on the observation noise, the parameter variations, and the amount of computation that is acceptable; e.g., standard stochastic approximation with fixed step-size, recursive least-squares with forgetting factor, or adaptations of Kalman–Bucy filters. In this work, we focus on the fixed step-size stochastic approximation algorithm,

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \mu P_t(\mu)\phi_t(y_t - \phi_t^T \hat{\theta}_t), \tag{2}$$

where μ is referred to as the adaptation step-size and $\{P_t(\mu)\}_{t\geq 0}$ is a sequence of random matrices (possibly depending on the step-size μ) which can be chosen in a number of different ways.

Least mean square (LMS) algorithm. By far the most popular algorithm in that class is the least mean square algorithm (LMS), introduced in a landmark paper by [36]. In that case, $P_t(\mu) = I$. Such an algorithm is referred to as a gradient algorithm, because the increment of the algorithm is opposite to the (stochastic)

[*] Received by the editors February 6, 1998; accepted for publication (in revised form) October 26, 1999; published electronically October 11, 2000.
http://www.siam.org/journals/sicon/39-3/33385.html

[†] Université Pierre et Marie Curie, 4 Place Jussieu, 75006 Paris, France, (aguech@ccr.jussieu.fr, priouret@ccr.jussieu.fr).

[‡] Ecole Nationale Supérieure des Télécommunications, 46, rue Barrault, 75634 Paris, France (moulines@tsi.enst.fr).

gradient of the mean-square error, $e_t(\theta) = E(y_t - \phi_t^T\theta)^2$. A useful variant of the LMS is the normalized LMS, in which the gradient is normalized by its norm,

$$P_t = \frac{1}{1+\epsilon\|\phi_t\|^2}, \quad \epsilon > 0.$$

Recursive least square (RLS) algorithm. This algorithm is derived by minimizing $\sum_{i=0}^{t}(1-\mu)^{t-i}(y_i - \theta^T\phi_i)^2$ ("old" measurements are exponentially discounted, with a "forgetting factor" $(1-\mu)$, where μ controls the effective memory of the algorithm). The minimization of the least-squares criterion can be performed recursively using (2) with

$$P_t(\mu)^{-1} = (1-\mu)P_{t-1}^{-1}(\mu) + \mu\phi_t\phi_t^T, \quad t \geq 1, \tag{3}$$

$$P_t(\mu) = (1-\mu)^{-1}\left(P_{t-1}(\mu) - \mu\frac{P_{t-1}(\mu)\phi_t\phi_t^T P_{t-1}(\mu)}{(1-\mu)+\mu\phi_t^T P_{t-1}(\mu)\phi_t}\right), \tag{4}$$

where $P_0 > 0$.

Kalman filter. In this case,

$$P_t(\mu) = [K_t(\mu) - \mu Q]R^{-1}, \tag{5}$$

$$K_t(\mu) = [K_{t-1}(\mu)^{-1} + \mu R^{-1}\phi_t\phi_t^T]^{-1} + \mu Q, \tag{6}$$

where $P_0 \geq 0$, $R > 0$, and $Q > 0$ are deterministic and can be arbitrarily chosen. It is well known (see, e.g., Caines [5, Chapter 5]) that if ϕ_t is $\sigma(y_i, i \leq t)$ measurable and $(v_t, w_t) \triangleq (v_t, \theta_t - \theta_{t-1})$ is a Gaussian white noise process, then θ_t is the minimum variance estimate for θ_t, $\hat{\theta}_t = E(\theta_t|\mathcal{F}_{t-1})$ and $P_t(\mu) = E(\theta_t\theta_t^T|\mathcal{F}_{t-1})$, provided that $E\Delta_t\Delta_t^T = \mu Q$ and $R = Ev_t^2$, $\hat{\theta}_0 = E\theta_0$ and $P_0 = E((\hat{\theta}_0 - \theta_0)(\hat{\theta}_0 - \theta_0)^T) - P_0$.

There is a vast literature on the analysis of algorithms of type (2). In most contributions, the main goal is to obtain bounds on the tracking errors. Preliminary results in that direction have been obtained by [3], [8]; see [9] for a review of these early contributions.

As pointed out in [19], it is also of interest to have not only tracking bounds but "approximate" and "tractable" expressions for the moment of the tracking error, and in particular, for the covariance of the tracking error. Such expressions enable the evaluation of the impact, both qualitatively and quantitatively, of different factors on the tracking performance (e.g., the dependence structure of the regressor sequence, the observation noise, and the lag noise). The dependence on this quantity does not appear clearly in the expressions of the bounds.

The main purpose of this paper is to present a method enabling us to obtain "approximate" expansions of the tracking error covariance (as well as higher-order moments) of "arbitrary" accuracy. The proposed method, based on a technique initially proposed by [29], [30] (see also [31], [32]) to analyze the LMS algorithm, consists in approximating the process defined in (2) by a family of nested processes, with simpler structure than the original error process. This decomposition enables us to compute approximations for the moments of the tracking errors and other related quantities which are valid up to any arbitrary order. These approximations are obtained as solutions of linear equations with coefficients that can be readily deduced from the moments of the regression sequence, the regression noise, and the lag noise.

The paper is organized as follows. In section 2, the perturbation expansion method is presented. In section 3, the main results of this contribution are stated. An illustration of these results is presented in section 5.

2. Perturbation expansion: The method.

To analyze any adaptive algorithm, it is usually convenient to convert it to a so-called error form; indeed from (1) and (2), we can write

$$\tilde{\theta}_{t+1} = (I - \mu P_t(\mu)\phi_t\phi_t^T)\tilde{\theta}_t + \mu P_t(\mu)\phi_t v_t - w_{t+1}, \tag{7}$$

where $\tilde{\theta}_t \triangleq \hat{\theta}_t - \theta_t$ is the weight-error vector and $w_{t+1} \triangleq \theta_{t+1} - \theta_t$ is the lag noise. This is a time-varying nonhomogeneous difference equation. Since this equation is linear, $\tilde{\theta}_{t+1}$ can further be decomposed as

$$\tilde{\theta}_t = {}^{u}\tilde{\theta}_t + \mu\,{}^{v}\tilde{\theta}_t +{}^{w}\,\tilde{\theta}_t, \tag{8}$$

$${}^{u}\tilde{\theta}_{t+1} = (I - \mu P_t(\mu)\phi_t\phi_t^T)\,{}^{u}\tilde{\theta}_t, \qquad {}^{u}\tilde{\theta}_0 = \tilde{\theta}_0 = -\theta_0, \tag{9}$$

$${}^{v}\tilde{\theta}_{t+1} = (I - \mu P_t(\mu)\phi_t\phi_t^T)\,{}^{v}\tilde{\theta}_t + P_t(\mu)\phi_t v_t, \qquad {}^{v}\tilde{\theta}_0 = 0, \tag{10}$$

$${}^{w}\tilde{\theta}_{t+1} = (I - \mu P_t(\mu)\phi_t\phi_t^T)\,{}^{w}\tilde{\theta}_t - w_{t+1}, \qquad {}^{w}\tilde{\theta}_0 = 0. \tag{11}$$

$\{{}^{u}\tilde{\theta}_t\}$ is a transient term, reflecting the way the successive estimates of the regression coefficients forget the initial conditions. $\{{}^{v}\tilde{\theta}_t\}$ accounts for the errors introduced by the measurement noise, $\{v_t\}$; similarly, $\{{}^{w}\tilde{\theta}_{t+1}\}$ accounts for the errors associated with lag noise $\{w_t\}$. According to these definitions, ${}^{v}\tilde{\theta}_t$ and ${}^{w}\tilde{\theta}_t$ obey an inhomogeneous stochastic recurrence equation

$$\delta_{t+1} = (I - \mu F_t(\mu))\delta_t + \xi_t(\mu) = \sum_{s=0}^{t} \Phi(t, s; \mu)\xi_s, \quad \delta_0 = 0, \tag{12}$$

where $\{F_t(\mu)\}_{t\geq 0}$ is a $(d \times d)$ matrix valued random process, $\{\xi_t(\mu)\}_{t\geq 0}$ is a $(d \times 1)$ vector-valued random process, and $\Phi(t, s; \mu)$ is defined as

$$\Phi(t, s; \mu) = \begin{cases} (I - \mu F_t(\mu))(I - \mu F_{t-1}(\mu)) \cdots (I - \mu F_{s+1}(\mu)), & t > s, \\ I, & t = s, \\ 0 & \text{otherwise.} \end{cases}$$

Equations (10) and (11) may be rewritten as (12) with $F_t(\mu) = P_t(\mu)\phi_t\phi_t^T$ and

$$\xi_t(\mu) = P_t(\mu)\phi_t v_t \quad \text{measurement noise}, \quad \xi_t = -w_{t+1} \quad \text{lag noise}. \tag{13}$$

In what follows, we will for ease of notation omit the dependence in the step-size μ when this dependence can be readily inferred from the context.

In this section, we concentrate on the general recurrence equation (12). We apply the results to the lag-error term $\tilde{\theta}_t$ in the next section. Equations of the form (12) have received considerable attention in the literature.

The approach developed in this contribution relies upon a perturbation technique (see [24]). Applied to the recurrence equation (12), the whole procedure goes as follows. The basic idea consists in replacing the random matrix $F_t(\mu)$ by an appropriately chosen *deterministic* matrix $\bar{F}_t(\mu)$, and then decomposing the recurrence equations (12) into two separate recursions:

$$J_{t+1}^{(0)} = (I - \mu\bar{F}_t(\mu))J_t^{(0)} + \xi_t, \quad J_0^{(0)} = 0, \tag{14}$$

$$H_{t+1}^{(0)} = (I - \mu F_t(\mu))H_t^{(0)} + \mu Z_t J_t^{(0)}, \qquad H_0^{(0)} = 0, \tag{15}$$

$$\delta_t = J_t^{(0)} + H_t^{(0)}, \tag{16}$$

where $Z_t = \bar{F}_t(\mu) - F_t(\mu)$. In the original construction by [30], the deterministic matrix $\bar{F}_t(\mu)$ was chosen to be the expectation of $F_t(\mu)$. This choice is adequate for the LMS algorithm, but not for more general tracking algorithms (RLS/Kalman algorithms). Explicit constructions of "appropriate" deterministic sequences $\{\bar{F}_t(\mu)\}$ for such cases will be given later.

According to (14), $J_t^{(0)}$ satisfies a deterministic inhomogeneous first-order linear difference equation

$$J_{t+1}^{(0)} = \sum_{s=0}^{t} \psi(t,s)\xi_s, \tag{17}$$

where, as above,

$$\psi(t,s) = \begin{cases} (I-\mu\bar{F}_t)(I-\mu\bar{F}_{t-1})\cdots(I-\mu\bar{F}_{s+1}), & t > s, \\ I, & t = s, \\ 0 & \text{otherwise.} \end{cases}$$

Under appropriate assumptions on the matrix valued sequences $\{F_t\}_{t\geq 0}$ and on the excitation $\{\xi_t\}$, it will be shown that, for some $p > 0$, there exists a constant $C < \infty$ and $\mu_0 > 0$ such that $\forall\ 0 < \mu \leq \mu_0$,

$$\sup_{t\geq 0} \|J_t^{(0)}\|_p \leq C/\sqrt{\mu} \quad \text{and} \quad \sup_{t\geq 0} \|H_t^{(0)}\|_p \leq C, \tag{18}$$

where $C < \infty$ is a constant depending on $\{F_t\}$ and $\{\xi_t\}$ (see below). Thus, $J_t^{(0)}$ may be considered as the leading term in the expansion, while $H_t^{(0)}$ may be seen as a correction term. The same procedure can be iterated to obtain higher-order approximations. For that purpose, it suffices to iterate the decomposition up to the desired order $n > 1$. Using this technique, the weight-error vector δ_t may be decomposed as

$$\delta_t = J_t^{(0)} + J_t^{(1)} + \cdots + J_t^{(n)} + H_t^{(n)}, \tag{19}$$

where the processes $J_t^{(r)}$, $0 \leq r \leq n$, and $H_t^{(n)}$ are respectively defined as

$$J_{t+1}^{(0)} = (I-\mu\bar{F}_t)J_t^{(0)} + \xi_t, \qquad J_0^{(0)} = 0, \tag{20}$$

$$J_{t+1}^{(r)} = (I-\mu\bar{F}_t)J_t^{(r)} + \mu Z_t J_t^{(r-1)}, \qquad J_t^{(r)} = 0,\ 0 \leq t < r, \tag{21}$$

$$H_{t+1}^{(n)} = (I-\mu F_t)H_t^{(n)} + \mu Z_t J_t^{(n)}, \qquad H_t^{(n)} = 0,\ 0 \leq t < n. \tag{22}$$

The processes $J_t^{(r)}$ depend linearly on ξ_t and polynomially in the error $Z_t = \bar{F}_t - F_t$. It is thus feasible (examples are given below) to compute the joint moments of these processes and to obtain expressions for the moments of $\tilde{\delta}_t^{(n)} = J_t^{(0)} + \cdots + J_t^{(n)}$. The residual term $H_t^{(n)}$ is, under appropriate conditions, uniformly bounded, i.e., there exists some constant $C < \infty$ and $\mu_0 > 0$, such that, $\forall\ 0 < \mu \leq \mu_0$, we have

$$\sup_{t\geq 0} \|H_t^{(n)}\|_p \leq C\mu^{n/2}. \tag{23}$$

Upper bounds for the constant C (depending upon the regression sequence, the observation noise, and the lag noise) are given below. $\tilde{\delta}_t^{(n)}$ is thus the leading term of the expansion, whereas $H_t^{(n)}$ is a remainder, which is uniformly bounded in L_p. By computing the moments of $\tilde{\delta}_t^{(n)}$, one can obtain an approximation of the moments of the tracking error δ_t, the error between the exact and the approximate expressions being uniformly bounded by $\mu^{n/2}$.

Remark.

- First-order expansion has been used by many contributors to obtain an approximate expression of the tracking error covariance matrix; see, e.g., [9], [10], [20], [11], [12], [13], [14], [15] for applications of this method.
- Solo in [29], [30] (see also [34], [33]) was the first to propose a construction allowing the computation of higher-order approximations for the covariance matrix of the tracking error of the LMS algorithm (extensions of this construction to general tracking algorithms are presented in [34]). [30] and [34] used this construction to obtain explicitly a second-order expansion of the covariance of the tracking error of the LMS algorithm. The construction by [29], [30] is essentially equivalent to the perturbation expansion outlined above.

3. Main results.

3.1. Preliminaries.

Before stating the main assumptions and results, it is necessary to state some definitions. All the processes are assumed to be defined on the same probability space $(\Omega, \mathcal{A}, P)$. Let $X \overset{\Delta}{=} \{X_t\}_{t\in\mathbb{Z}}$ be a vector-valued random process. The σ-field generated by the random variables X_t, $a \le t \le b$, is denoted $\mathcal{M}_a^b(X)$.

Boundedness in L_p. For $p > 0$, and $\mathcal{B} \subset \mathcal{A}$, denote $L_p(\Omega, \mathcal{B}, P)$ as the space of $\mathcal{B}$-measurable random variables such that $\|X\|_p < \infty$; for brevity, we set $L_p(\Omega, \mathcal{A}, P) = L_p$. A random matrix sequence $\{X_t\}_{t\ge 0}$ is called L_p-bounded if $\sup_{t\ge 0} \|X_t\|_p < \infty$.

Exponential stability. After [12], define, for $p \ge 1$, $\mu^* > 0$, and $0 < \beta < 1/\mu^*$, $\mathcal{S}(p, \beta, \mu^*)$ as the set of *random* matrix-valued processes $\{A_k(\mu)\}_{k\in\mathbb{N}}$ satisfying

$$\mathcal{S}(p,\beta,\mu^*) = \left\{ A(\mu) = \{A_k(\mu)\}_{k\in\mathbb{N}} : \left\| \prod_{j=i+1}^{k} (I - \mu A_j(\mu)) \right\|_p \right.$$
$$\left. \le K_{\beta,\mu^*}(A)\,(1-\beta\mu)^{k-i} \qquad \forall\, \mu \in (0,\mu^*],\ \forall\, k \ge i \ge 0 \right\},$$

where the constant $K_{\beta,\mu^*}(A)$ does not depend on μ. $\mathcal{S}(p,\beta,\mu^*)$ is referred to as the L_p *exponentially stable* family. Likewise, define the *averaged exponentially stable* family as the set of deterministic matrix-valued processes $\{A_k(\mu)\}_{k\in\mathbb{N}}$:

$$\mathcal{S}(\beta,\mu^*) = \left\{ \bar{A}(\mu) = \{\bar{A}_k(\mu)\}_{k\in\mathbb{N}} : \left| \prod_{j=i+1}^{k} (I - \mu \bar{A}_j(\mu)) \right| \right.$$
$$\left. \le K_{\beta,\mu^*}(\bar{A})\,(1-\beta\mu)^{k-i} \qquad \forall\, \mu \in (0,\mu^*],\ \forall\, k \ge i \ge 0 \right\}.$$

Weak dependence. After [7], we define weak-dependence as follows. Let $q \geq 1$ and let $X = \{X_n\}_{n \geq 0}$ be an $(l \times 1)$ matrix-valued process. Let $\delta = (\delta(r))_{r \in \mathbb{N}}$ be a sequence of positive numbers decreasing to zero at infinity. The process $X = \{X_n\}_{n \geq 0}$ is said to be (δ, q)-weak dependent if there exist finite constants $C = \{C_1, \ldots, C_q\}$ such that for any $1 \leq m < s \leq q$, any $m$tuple $t_1, \ldots, t_m$, and any $(s-m)$tuple $t_{m+1}, \ldots, t_s$, with $t_1 \leq \cdots \leq t_m < t_m + r \leq t_{m+1} \leq \ldots t_s$, it holds

$$\sup_{1 \leq i_1, \cdots, i_s \leq l} \left| \operatorname{cov}\left(\tilde{X}_{t_1,i_1} \cdots \tilde{X}_{t_m,i_m}, \tilde{X}_{t_{m+1},i_{m+1}} \cdots \tilde{X}_{t_s,i_s} \right) \right| \leq C_s \delta(r),$$

where $\tilde{X}_{n,i}$ denotes the ith component of $X_n - E(X_n)$.

Rosenthal's class. Rosenthal's class is defined as the class of vector-valued stochastic processes $\epsilon \triangleq \{\epsilon_k\}_{k \geq 0}$ verifying Rosenthal's inequality (see, e.g., [27, Theorem 2.9]). More specifically,

$$(24) \quad \mathcal{N}(p) = \left\{ \epsilon : \left\| \sum_{k=s}^{t} D_k \epsilon_k \right\|_p \leq \rho_p(\epsilon) \left(\sum_{k=s}^{t} |D_k|^2 \right)^{1/2} \quad \forall\, 0 \leq s \leq t \text{ and } \right.$$
$$\left. \forall\, D = \{D_k\}_{k \in \mathbb{N}}\ (q \times l) \text{ deterministic matrices} \right\}.$$

Exponential stability. L_p exponential stability forms the basis of the stability analysis of linear state space systems (see, e.g., [10], [12], [13], [14], [34, sec. C6], [33]). It is a natural extension of the notion of exponential stability for deterministic linear systems. To our knowledge, the notion of L_p exponential stability was introduced by [8]; these authors provide restrictive conditions (m-independence and moment conditions) for a family of random matrices to be in $\mathcal{S}(p, \beta, \mu_0)$. As evidenced by many authors since then, L_p exponential stability is the crux of the analysis of stochastic difference equations of the form (12) and many practical methods and criteria to establish this property have been obtained. Without entering into too much detail, L_p exponential stability requires some strengthening of the classical "persistence of excitation" condition, which is the classical condition to study the stability of nonstationary deterministic linear systems. The type of strengthening precisely needed depends upon the algorithm under consideration, but generally involves moments and mixing conditions. Early contributions in this direction include the work of [10] on the Kalman algorithm and [1], [2], and [11] on the RLS algorithm. Some general conditions to check L_p exponential stability are presented in [12] (with applications to the LMS, the RLS, and the Kalman algorithm); some improvements of these results can be found in a series of papers coauthored by [13], [14], [15]. Necessary and sufficient conditions for the LMS algorithm are presented in [15]. Alternate conditions implying exponential stability have been obtained in the case where the matrix valued process $\{F_t\}_{t \geq 0}$ is Markovian symmetric [28].

Weak-dependence. Weak-dependence states a kind of "decorrelation" between the past and the future of the process. Weak-dependence is essential in the developments that follow and appear at several places. The notion of weak-dependence, introduced by [7], is a way to weaken more classical "strong mixing" assumptions. As shown in [7], weak-dependent processes encompass a large class of models and in particular strongly mixing processes, weak shift processes, or models with a Markovian representation. The following lemma shows that strong-mixing processes also are weak-mixing and give the relation between the strong-mixing coefficient $\alpha_X(\tau)$ and the sequence $\delta(\tau)$ appearing in the definition of weak-dependent processes. We have the following lemma.

LEMMA 1. *Assume that $X = \{X_t\}_{t\geq 0}$ is an $(l \times 1)$ strongly mixing process and that*

$$\sup_{t\geq 0} \sup_{1\leq i\leq l} \|X_{t,i}\|_{q'} < \infty$$

for some $q' > q$. Then $X = \{X_t\}_{t\geq 0}$ is (δ, q)-weak dependent, with, for $s \in \{1, \dots, q\}$,

$$\delta(r) = \alpha(r)^{(q'-q)/q'}, \quad r \geq 1,$$
$$C_s = 12 \left(\sup_{t\geq 0} \sup_{1\leq i\leq l} \|X_{t,i}\|_{s+q'-q} \right)^s .$$

This proof is directly adapted from [7] (see also [6], [35] for similar results and arguments).

Rosenthal's inequality. Rosenthal's class gathers all the processes that verify a Rosenthal's type inequality. By direct applications of the Rosenthal's inequality [27], $\mathcal{N}(p)$ includes sequences $\epsilon = \{\epsilon_k\}_{k\geq 0}$ of independent random variables with zero-mean and uniformly bounded pth order moment. By application of Burkholder's inequality for martingales, martingales with L_p stable increments also belong to this class. Of course, the class of processes belonging to this class is much larger than that, including many processes satisfying moment and mixing conditions. Some typical examples of processes belonging to $\mathcal{N}(p)$ are given in the proposition below.

PROPOSITION 2. *$\{\epsilon_t\}$ belongs to $\mathcal{N}(p)$ if*

(i) *$\{\epsilon_t\}_{t\geq 0}$ is an L_p-bounded martingale increment:* $\sup_{t\geq 0} \|\epsilon_t\|_p < \infty$. *In this case, $\rho_p(\epsilon)$ in* (24) *may be chosen to be*

$$\rho_p(\epsilon) = B_p \sup_{s\geq 0} \|\epsilon_s\|_p, \tag{25}$$

where B_p is a universal constant that depends only on p and l (the dimension of the $\{\epsilon_s\}$).

(ii) *$\{\epsilon_t\}_{t\geq 0}$ is zero-mean and (δ, q) weak-dependent with $q = 2[(p+1)/2]$ and $\sum_{r=1}^{\infty}(r+1)^{(q/2-1)}\delta(r) < \infty$ (the constant $\rho_p(\epsilon)$ can be found from the proof).*

As mentioned above, (i) is a consequence of Burkholder's inequality for martingales (see [16]). (ii) can be adapted from the results of [7]. A proof is in Appendix A (see also [17, Proposition 10, Corollary 11], where a Rosenthal's inequality is proved for α-mixing processes).

To conclude this section, it is worthwhile to note that $\mathcal{N}(p)$ is invariant under stable linear time-varying transformation. This means that, if $\{\epsilon_k\} \in \mathcal{N}(p)$, then the process $\{\epsilon'_k\}$ defined as

$$\epsilon'_k = \sum_{j=-\infty}^{\infty} A(k,j)\epsilon_{k-j} \qquad \sum_{j=-\infty}^{\infty} \sup_k |A(k,j)| < \infty,$$

where $A(k,j)$ are deterministic matrices, also belongs to $\mathcal{N}(p)$. Moreover, the constants are upper bounded by

$$\rho_p(\epsilon') \leq \rho_p(\epsilon) \left(\sum_{j=-\infty}^{\infty} \sup_k |A(k,j)| \right). \tag{26}$$

Thus, random processes generated from martingale differences or (δ, q) weak-dependent processes via an infinite order time-varying linear filter can all be included in $\mathcal{N}(p)$.

3.2. Assumptions on $\{F_k\}_{k\geq 0}$.

According to the structure of the recursion involved in the definition of the LMS, RLS, and Kalman algorithms, it is assumed in what follows that the sequence of matrices $\{F_k(\mu)\}_{k\geq 0}$ may be decomposed as a product

$$F_k(\mu) = P_k(\mu) G_k, \tag{27}$$

where $\{P_k(\mu)\}_{k\geq 0}$ is a sequence of random matrices that can be approximated (in a sense given below) by a sequence of deterministic matrices $\{\bar{P}_k(\mu)\}_{k\geq 0}$ and $\{G_k\}_{k\geq 0}$ is a sequence of random matrices (that do not depend on the step-size μ) and that verify some moment and weak-dependence conditions. For the three different tracking algorithms considered in this paper, we set $G_k = \phi_k \phi_k^T$, while $P_k(\mu)$ is a random weight whose specific form depends upon the tracking algorithm. The sequence of matrices $\bar{F}_k(\mu)$ needed to construct the approximations is defined as $\bar{F}_k(\mu) = \bar{P}_k(\mu) E(G_k)$. The conditions that are typically required for the sequence $F_k(\mu)$ and $\bar{F}_k(\mu)$ are that, for some $r, q \in \mathbb{N}$, $\mu_0 > 0$, and $0 < \beta < 1/\mu_0$,

(F1) (r, β, μ_0). $\{F_k(\mu)\}_{k\geq 0}$ is in $\mathcal{S}(r, \beta, \mu_0)$, i.e., $\{F_k(\mu)\}$ is L_r-exponentially stable,
(F2) (β, μ_0). $\{\bar{F}_k(\mu)\}_{k\geq 0}$ is in $\mathcal{S}(\beta, \mu_0)$, i.e., $\{\bar{F}_k(\mu)\}_{k\geq 0}$ is averaged exponentially stable,
(F3) (q, μ_0). $\sup_{t\in\mathbb{N}} \sup_{\mu\in(0,\mu_0]} \|P_t(\mu)\|_q < \infty$, $\sup_{t\in\mathbb{N}} \sup_{\mu\in(0,\mu_0]} |\bar{P}_t(\mu)| < \infty$,
(F4) (q, μ_0). $\sup_{t\in\mathbb{N}} \sup_{\mu\in(0,\mu_0]} \mu^{-1/2} \|P_t(\mu) - \bar{P}_t(\mu)\|_q < \infty$.

As mentioned above, assumptions (F1)–(F2) are classical in the study of random linear systems, and hold under "generalized persistence of excitation" conditions (see the discussion above on exponential stability). The same conditions typically imply (F3) (see, e.g., [20], [11], and [12]; see also [1], [2] for related results on the forgetting factor RLS alogorithm).

The last condition (F4), meaning that the random matrices $P_t(\mu)$ differ only by a (small) amount (controlled by the step-size μ) from a deterministic matrix $\bar{P}_t(\mu)$, is perhaps less classical, though it has often been used to study the forgetting factor RLS algorithm [1], [2] (see also [11], for the Kalman filter). For the LMS algorithm, $P_k(\mu) = I$ and we set $\bar{P}_k(\mu) = I$ and this assumption is trivially fulfilled. For the RLS algorithm and the Kalman algorithm, the situation is slightly more complicated. For completeness, we treat the matter in some detail for the RLS algorithm and only outline the necessary ingredients of the construction for the Kalman algorithm.

RLS algorithm. For the RLS algorithm, the deterministic sequence of matrices $\{\bar{P}_k(\mu)\}_{k\geq 0}$ is recursively defined as

$$\bar{P}_{k+1}^{-1}(\mu) = (1-\mu)\bar{P}_k^{-1}(\mu) + \mu E(\phi_{k+1}\phi_{k+1}^T), \qquad \bar{P}_0(\mu) = R_0, \tag{28}$$

where R_0 is the initial condition for $P_0(\mu)$, i.e., the same initial condition is used for $P_0(\mu)$ and $\bar{P}_0(\mu)$. Note that

$$P_k^{-1}(\mu) = \mu^k P_0 + \mu \sum_{j=0}^{k} (1-\mu)^{k-j} \phi_j \phi_j^T, \tag{29}$$

$$\bar{P}_k^{-1}(\mu) = \mu^k P_0 + \mu \sum_{j=0}^{k} (1-\mu)^{k-j} E(\phi_j \phi_j^T) \tag{30}$$

so that $P_k^{-1}(\mu) - \bar{P}_k^{-1}(\mu) = \mu \sum_{j=0}^k (1-\mu)^{k-j}(\phi_j \phi_j^T - E(\phi_j\phi_j^T))$. Provided that the sequence $\{\phi_k\phi_k^T - E(\phi_k\phi_k^T)\}_{k\geq 0}$ belongs to Rosenthal's class $\mathcal{N}(b)$ $(b > 1)$, it holds that

$$\|P_k^{-1}(\mu) - \bar{P}_k^{-1}(\mu)\|_b \leq \rho_b(\phi\phi^T)\mu \left(\sum_{j=0}^k (1-\mu)^{2(k-j)}\right)^{1/2} \leq C_b(\phi\phi^T)\sqrt{\mu} \tag{31}$$

for $0 < \mu < 1$. Since, by applying the Holder inequality

$$\|P_k(\mu) - \bar{P}_k(\mu)\|_p \leq \|P_k(\mu)\|_a \|P_k^{-1}(\mu) - \bar{P}_k^{-1}(\mu)\|_b \; |\bar{P}_k(\mu)|$$

for any $a > 0$, $b > 0$ such that $a^{-1} + b^{-1} = p^{-1}$, (31) shows that assumption (F4) holds as soon as (F3) (a, μ_0)) is satisfied.

Kalman algorithm. For the Kalman algorithm, the deterministic sequence of matrices $\bar{P}_k(\mu)$ is defined as

$$\bar{P}_k(\mu) = [\bar{K}_k(\mu) - \mu Q]R^{-1}, \tag{32}$$

$$\bar{K}_k(\mu) = [\bar{K}_{k-1}(\mu)^{-1} + \mu R^{-1} E(\phi_k\phi_k^T)]^{-1} + \mu Q, \tag{33}$$

where, by convention, we set $\bar{P}_0(\mu) = P_0(\mu) = P_0$. Assumption (F4) may be verified along the same lines as above (see, e.g., [13]).

3.3. Assumptions on the excitation sequence $\boldsymbol{\xi = \{\xi_k\}_{k\geq 0}}$. In addition to the above stated assumptions on the matrix process $F = \{F_k\}_{k\geq 0}$, we need to impose some conditions on the excitation sequence $\xi = \{\xi_k\}_{k\geq 0}$. To cover the analysis of stochastic tracking algorithms, it is convenient to assume that the excitation $\xi_t(\mu)$ may be decomposed as

$$\xi_t(\mu) = M_t(\mu)\epsilon_t, \tag{34}$$

where the process $M = \{M_t(\mu)\}_{t\geq 0}$ is a $(d \times l)$ matrix-valued process, $\epsilon = \{\epsilon_t\}_{t\geq 0}$ is an $(l \times 1)$ vector-valued process, and both processes verify the following assumptions.

(EXC1) $\{M_t(\mu)\}_{t\in\mathbb{Z}}$ is $\mathcal{M}_0^t(\phi)$-adapted and $\mathcal{M}_0^t(\epsilon)$ and $\mathcal{M}_0^t(\phi)$ are independent.

(EXC2) $(r, \mu 0)$, $(r > 0,\ \mu_0 > 0)$ $\sup_{\mu\in(0,\mu_0]} \ \sup_{t\geq 0} \ \|M_t(\mu)\|_r < \infty$.

(EXC3) (p, μ_0) $(p > 0,\ \mu_0 > 0)$ $\epsilon = \{\epsilon_t\}_{t\in\mathbb{N}}$ belongs to Rosenthal's class $\mathcal{N}(p)$.

Recall that, in the application to stochastic tracking algorithms, $M_t(\mu)$ is equal either to $P_t(\mu)\phi_t$ (for the measurement noise) or to the identity $M_t = I$ (for the lag noise). For general tracking algorithms, the sequence $P_t(\mu)$ is computed recursively from the sequence of regressors, so that $P_t(\mu)$ is adapted to the sequence of σ-subfields $\sigma(\phi_s, 0 \leq s \leq t)$. The noise ϵ_t is either equal to the measurement noise v_t or to the lag noise w_t. Assumption (EXC1) thus covers the case where the measurement noise and the lag noise are independent of the regressor sequence. It excludes some applications of interest, where the regressors ϕ_t depend on the past of θ_t (occurring in tracking in ARX models).

For the LMS algorithm, $M_t(\mu) = \phi_t$ (for the measurement) noise, and the boundedness of $M_t(\mu)$ (assumption (EXC2)) follows from the stability of the regressor sequence. For the RLS and Kalman algorithms, $M_t(\mu) = P_t(\mu)\phi_t$, and the boundedness of $M_t(\mu)$ follows from the boundedness of $P_t(\mu)$ and the stability of the regressor sequence. This assumption is a consequence of the generalized persistence of excitation condition, as mentioned above.

The third condition states that $\{v_t\}$ and $\{w_t\}$ belong to Rosenthal's class provided that $\{v_t\}$ and $\{w_t\}$ are either martingale increments or zero-mean weakly dependent sequences, with uniformly bounded moments. This assumption covers many cases of practical interest. When specialized to the study of the "tracking" component ($\xi_t = w_t$), it means that θ_t follows a generalized random walk, with nonstationary and dependent increment. The L_p moment of the parameter to track $\|\theta_t\|_p$ is assumed to grow like $\sqrt{t}$. The result can be adapted to the case where θ_t is zero-mean L_p stable (but the results are in that case slightly different). Tracking of deterministic trends is not covered.[1]

Because of the independence of the noise $\{\epsilon_k\}$ and of the regression sequence $\{\phi_k\}$, the following is easily shown.

PROPOSITION 3. *Let $q > p$ and $\mu_0 > 0$. Assume* (EXC1), (EXC2($pq/(q-p), \mu_0$), *and* (EXC3(p)). *Let $\{G_k(\mu)\}_{k\geq 0}$ be an $\mathcal{M}_0^t(\phi)$ adapted sequence, such that $\|G_k(\mu)\|_q < \infty$. Then,*

$$\left\| \sum_{k=s}^{t} G_k(\mu) M_k(\mu) \epsilon_k \right\|_p \leq \rho_p(\epsilon) \sup_{k\geq 0} \|M_k\|_{pq/(q-p)} \left(\sum_{k=s}^{t} \|G_k(\mu)\|_q^2 \right)^{1/2} . \tag{35}$$

3.4. The main results. We may now formulate the central results of our contribution. The first result gives condition upon which $J_s^{(r)}$ is uniformly bounded in L_p and provides an expression for that bound.

THEOREM 4. *Let $n \in \mathbb{N}$ and let $q \geq p \geq 2$. Assume* (EXC1), (EXC2($pq/(q-p), \mu_0$)), *and* (EXC3(p)). *For $a, b, \beta > 0$, $a^{-1} + b^{-1} = 1$, and some $\mu_0 > 0$, assume in addition* (F2(β, μ_0)), (F4(aqn, μ_0)) *and*

(i) $\{G_t\}_{t\geq 0}$ *is* $(\delta, (q+2)n)$ *weakly dependent, and* $\sum (r+1)^{((q+2)n/2)-1} \delta(r) < \infty$,

(ii) $\sup_{t\geq 0} \|G_t\|_{bqn} < \infty$.

Then, there exists a constant $K < \infty$ (depending on $\delta(k)$, $k \geq 0$, and on the numerical constants $p, q, n, a, b, \mu_0, \beta$ but not *otherwise on $\{\phi_t\}$, $\{\epsilon_t\}$ or on the step-size parameter μ), such that $\forall 0 < \mu \leq \mu_0$, $\forall\ 0 \leq r \leq n$,*

$$\sup_{s\geq 1} \|J_s^{(r)}\|_p \leq K\ \rho_p(\epsilon) \sup_{k\geq 0} \|M_k\|_{pq/(q-p)}\ \mu^{(r-1)/2}. \tag{36}$$

(The precise value of the constant K may be found from the proof.)

The proof is given in Appendix B. To complete our program, we need to bound the remainder term $H_s^{(n)}$. We will prove that, under appropriate conditions, the moments of $H_s^{(n)}$ are bounded by the moments of $J_s^{(n+1)}$. Since Theorem 4 states conditions upon which $\|J_s^{(n+1)}\|_p \leq K\mu^{n/2}$, this will in turn allow us to specify technical conditions upon which the remainder $H_s^{(n)}$ is bounded in L_p, i.e., $\sup_{s\geq 0} \|H_s^{(n)}\| \leq K\mu^{n/2}$.

THEOREM 5. *Let $p \geq 2$ and let $a, b, c > 0$ such that $1/a + 1/b + 1/c = 1/p$. Let $n \in \mathbb{N}$. Assume* (F1(a, β, μ_0)) *and*

(i) $\sup_{s\geq 0} \|Z_s\|_b < \infty$ *and*

(ii) $\sup_{s\geq 0} \|J_s^{(n+1)}\|_c < \infty$.

Then, there exists a constant $K' < \infty$ (depending on the numerical constants a, b, c, β, μ_0, n but not on the process $\{\epsilon_t\}$ or on the step-size parameter μ), such that

[1] It should be stressed, however, that the perturbation expansion technique can be adapted to the study of such tracking problems. This particular setting is not studied in detail due to space limitations.

$\forall\ 0 < \mu \leq \mu_0$,

$$\sup_{s \geq 0} \| H_s^{(n)} \|_p \leq K' \sup_{s \geq 0} \| J_s^{(n+1)} \|_c. \tag{37}$$

(The precise value of K' may be found from the proof.)

The proof of Theorem 5 is based on the decomposition $H_s^{(n)} = J_s^{(n+1)} + H_s^{(n+1)}$. It thus amounts to showing that $H_s^{(n+1)}$ is bounded in L_p. Since

$$H_t^{(n+1)} = \mu \sum_{s=0}^{t} \Phi(t,s) Z_t J_t^{(n+1)},$$

the proof stems directly from the exponential stability of $\Phi(t,s)$, the uniform boundedness of moments of $\{Z_s\}$ (which can be also translated in terms of boundedness of moments of $P_t(\mu)$ and of the driving terms G_t). Details are given in Appendix B.

4. Performance of adaptive tracking algorithms. A number of useful error bounds or expressions can be derived from the results above. We use the notations of section 1. According to (8), the tracking error may be expanded as $\tilde{\theta}_t =^u \tilde{\theta}_t + \mu^v \tilde{\theta}_t +^w$ $\tilde{\theta}_t$, where ${}^u\tilde{\theta}_t$, ${}^v\tilde{\theta}_t$, and ${}^w\tilde{\theta}_t$ are respectively defined in (9), (10), and (11). The terms ${}^v\tilde{\theta}_t$ and ${}^w\tilde{\theta}_t$ may further be decomposed as

$${}^v\tilde{\theta}_t = \sum_{k=0}^{r_v} {}^vJ_t^{(k)} + {}^vH_t^{(r_v)}, \tag{38}$$

$${}^w\tilde{\theta}_t = \sum_{k=0}^{r_w} {}^wJ_t^{(k)} + {}^wH_t^{(r_w)}, \tag{39}$$

where r_v and r_w are two integers (not necessarily equal) such that $0 \leq r_v \leq n-1$ and $0 \leq r_w \leq n-1$. To apply Theorems 4 and 5, we need only to check that the assumptions are fulfilled for the measurement noise term and the lag-noise term. Because the particular form of the random matrix sequence $F_t(\mu) = P_t(\mu)\phi_t\phi_t^T$ and of the excitation term $\xi_t(\mu) = P_t(\mu)\phi_t v_t$ (measurement noise) or $\xi_t = -w_t$ (lag noise) have been tailored for studying general tracking algorithms, the assumptions of these theorems easily translate (see the discussion above) into general conditions on the regression sequence, on the measurement and lag noise (in terms of existence and boundedness of moments, mixing coefficients, etc.), and to some "algorithm-dependent" conditions, mainly involving more or less stringent conditions on the regression sequence, in terms of moment and mixing conditions. By direct adaptations of the assumptions of Theorems 4 and 5, it may be shown that

$$\sup_t \| {}^vJ_t^{(k)} \|_p \leq K\rho_p(v) \sup_{k \geq 0} \| P_k(\mu)\phi_k \|_{pq/(q-p)} \mu^{(k-1)/2}, \quad k \in \{0, \ldots, r_v\}, \tag{40}$$

$$\sup_t \| {}^vH_t^{(r_v)} \|_p \leq K\rho_p(v) \sup_{k \geq 0} \| P_k(\mu)\phi_k \|_{pq/(q-p)} \mu^{r_v/2}, \tag{41}$$

$$\sup_t \| {}^wJ_t^{(k)} \|_p \leq K\rho_p(w) \mu^{(k-1)/2}, \quad k \in \{0, \ldots, r_w\}, \tag{42}$$

$$\sup_t \| {}^wH_t^{(r_w)} \| \leq K\rho_p(w) \mu^{r_w/2}. \tag{43}$$

Setting $r_v = r_w = 0$, one obtains the expansion of the tracking error bounds presented in [20], [11], [12] and later extended to more general algorithms by [13], [14], [15].

Higher-order expansions can be used to obtain refined approximation of the tracking error moments (see the discussion below).

Of particular interest is the covariance of the tracking error. If one assumes that the measurement noise and the lag noise are independent, the covariance of the tracking error may be expressed as

$$\Gamma(t) \triangleq E(\tilde{\theta}_t \tilde{\theta}_t^T) =^u \Gamma(t) + \mu^2 \, {}^v\Gamma(t) +^w \Gamma(t),$$

where ${}^u\Gamma(t) = E({}^u\tilde{\theta}(t) {}^u\tilde{\theta}(t)^T)$, ${}^v\Gamma(t) = E({}^v\tilde{\theta}_t^v \tilde{\theta}_t^T)$, and ${}^w\Gamma(t) = E({}^w\tilde{\theta}_t^w \tilde{\theta}_t^T)$. Equation (38) then implies that, e.g., ${}^v\Gamma(t)$ may be expanded as

$${}^v\Gamma(t) = \sum_{k+l=0}^{r_v} E({}^v J_t^{(k)^v} J_t^{(l)\,T}) + O(\mu^{(r_v-1)/2})$$

yielding to a valid approximation to order $(r_v - 1)/2$ (a similar result holds for the lag noise). Two illustrative examples are worked out in the next section.

Remark. The evaluation of the covariance $E({}^v J_t^{(k)^v} J_t^{(l)\,T})$ involves the computation of moments of the form

$$(44) \quad E(\psi(t, s_1) Z_{s_1} \psi(s_1 - 1, s_2) Z_{s_2} \cdots \\ \psi(s_k - 1, s) P_s(\mu) \phi_s v_s^2 \phi_s^T P_s(\mu)^T \psi^T(s, s_l' - 1) Z_{s_l'} \cdots Z_{s_1'} \psi^T(t, s_1')),$$

where $Z_s = \bar{F}_t - F_t = -P_t(\mu)\phi_t\phi_t^T + \bar{P}_t(\mu) E(\phi_t\phi_t^T)$. The evaluation of such moments is straightforward for the LMS algorithm (see below for a worked-out example). The direct evaluation of this moment is not possible when dealing with the RLS algorithm or the Kalman algorithm. When one is willing to obtain an approximation of the moment of the tracking error valid up to order n, it suffices to determine an order n approximation of the moment of the form (44). The way to derive such approximations can be obtained for the RLS algorithm, as shown below (a similar derivation can be done for the Kalman algorithm). The basic ingredients to derive such an expansion are (i) $P_t(\mu)$ differs by $\sqrt{\mu}$ for $\bar{P}_t(\mu)$ and (ii) the joint moments on $P_t^{-1}(\mu)$ can easily be evaluated, because $P_t^{-1}(\mu)$ depends linearly on $\{\phi_s\phi_s^T\}_{1 \le s \le t}$. We focus on the RLS algorithm; similar results can be derived for the Kalman algorithm. Note that, for any integer n, we may write

(45)

$$P_t(\mu) = (\bar{P}_t^{-1}(\mu) - (\bar{P}_t^{-1}(\mu) - P_t^{-1}(\mu)))^{-1} = (I - \bar{P}_t(\mu)(\bar{P}_t^{-1}(\mu) - P_t^{-1}(\mu)))^{-1}\bar{P}_t(\mu),$$

(46)

$$= \sum_{k=0}^{n} [\bar{P}_t(\mu)(\bar{P}_t^{-1}(\mu) - P_t^{-1}(\mu))]^k \bar{P}_t(\mu) + [(\bar{P}_t^{-1}(\mu) - P_t^{-1}(\mu))\bar{P}_t(\mu)]^{n+1} P_t(\mu).$$

Equation (31) shows that, when $\{\phi_s\phi_s^T\}_{s\in\mathbb{N}}$ belongs to Rosenthal's class $\mathcal{N}((n+1)p)$, it holds that

$$\sup_{t\in\mathbb{N}} \sup_{(0<\mu<1)} \mu^{-1/2} \|P_t^{-1}(\mu) - \bar{P}_t^{-1}(\mu)\|_{(n+1)p} < \infty.$$

In addition, under (F3(q, μ_0)), $\sup_{t\in\mathbb{N}} \sup_{\mu\in(0,\mu_0]} \|P_t(\mu)\|_q < \infty$ and $\sup_{t\in\mathbb{N}} \sup_{\mu\in(0,\mu_0]} |\bar{P}_t(\mu)| < \infty$. Hence, for $r, p, q > 0$ and $n \in \mathbb{N}$, $r^{-1} = p^{-1} + q^{-1}$, it holds that there exists some constant $K < \infty$ such that, $\forall\ \mu \in (0, \mu_0]$,

$$\sup_{t\in\mathbb{N}} \left\| P_t(\mu) - \sum_{k=0}^{n} [\bar{P}_t(\mu)(\bar{P}_t^{-1}(\mu) - P_t^{-1}(\mu))]^k \bar{P}_t(\mu) \right\|_r \leq K\mu^{(n+1)/2}$$

and the constant K can be chosen as

$$K = \sup_{t\in\mathbb{N}} \sup_{\mu\in(0,\mu_0]} \|P_t(\mu)\|_q |\bar{P}_t(\mu)|^{n+1} \sup_{t\in\mathbb{N}} \sup_{\mu\in(0,\mu_0]} \mu^{-1/2} \|P_t^{-1}(\mu) - \bar{P}_t^{-1}(\mu)\|_{(n+1)p} < \infty.$$

5. Some worked-out examples. In this section, approximate expressions for the tracking error covariance matrix for the LMS and the RLS algorithm are derived. To illustrate our findings, it is shown in this section that first-order approximation of the tracking error covariance may fail, in certain situations, to capture the behavior of the algorithm. It is argued that a second-order expansion leads to significantly better approximation, in many situations of practical interest; moreover, second-order approximation reveals the impact of certain factors which do not influence the first-order approximation, in particular the dependence between the successive regression vectors. To illustrate these effects without obscuring the presentation with cumbersome notations and details, a very simple situation is considered. Theoretical results are validated by means of a small-scale Monte Carlo experiment. More details will be given in a forthcoming paper.

(M1) The regressor $\{\phi_t\}_{t\geq 0}$ is a strict-sense stationary vector autoregressive process

$$\phi_{t+1} = \kappa\phi_t + u_{t+1},$$

where κ $(-1 < \kappa < 1)$ is a scalar, $\{u_t\}_{t\in\mathbb{Z}}$ is a sequence of independent and identically distributed Gaussian random vectors with zero-mean and covariance matrix $\sigma_u^2 I$.

(M2) The measurement noise process $\{v_t\}_{t\geq 0}$ and the lag-noise process $\{w_t\}_{t\geq 0}$ are two sequences of zero-mean independently and identically distributed (i.i.d) random variables (vectors), with bounded moments of order r, where $r > 2$. We denote $E(v_0^2) = \sigma_v^2$ and $E(w_0 w_0^T) = \gamma^2 I$.

(M3) $\mathcal{M}_0^\infty(v)$, $\mathcal{M}_0^\infty(\phi)$, and $\mathcal{M}(\theta)$ are independent.

Because our main concern in this section is the asymptotic regime, we set $\tilde{\theta}_0 = 0$. To apply the results in section 4, that one may apply Theorems 4 and 5 under (M1–3), we set $b = c = 2r$, $a = 2r/(r-2)$, and $d = r + \delta$, where $\delta > 0$ but is otherwise arbitrary. Note that $a^{-1} + b^{-1} + c^{-1} = p^{-1}$ with $p = 2$.

It follows from [25] that, under (M1), $\{\phi_t\}$ is geometrically completely regular (see also [4]), so that $F_t \triangleq \phi_t\phi_t^T$ is strongly mixing with exponentially decaying strong-mixing coefficient and thus, by Lemma 1 (see Appendix A), weak-dependent with exponentially decaying weak-dependent coefficient. It follows from [28, Theorem 1] that F_t is exponentially stable, i.e., for any $p \geq 1$ there exists $\mu_0 > 0$ and $0 < \beta < 1/\mu_0$ such that $\{F_t\} \in \mathcal{S}(\beta, \mu_0) \cap \mathcal{S}(p, \beta, \mu_0)$. Since $\{v_t\}$ and $\{w_t\}$ are i.i.d, $\{v_t\}$ and $\{w_t\}$ belong to $\mathcal{N}(r)$, with constants $\rho_r(v)$ and $\rho_r(w)$ defined as

$$\rho_r(v) = B_r\sigma_v\mu_r(v), \quad \rho_r(w) = B_r\gamma\mu_r(w),$$

where $\mu_r(v)$ and $\mu_r(w)$ are the standardized rth order moments of v and w, respectively, and B_r is a universal constant (see [27]). Under (M3), the processes $\{{}^v\tilde{\theta}_t\}_{t\geq 0}$ and $\{{}^w\tilde{\theta}_t\}$ are independent. Thus,

$$\Gamma = \lim_{t\to\infty} E(\tilde{\theta}_t\tilde{\theta}_t^T) = {}^v\Gamma\mu^2 + {}^w\Gamma,$$

where ${}^v\Gamma = \lim_{t\to\infty} E({}^v\tilde{\theta}_t^v\tilde{\theta}_t^T)$ and ${}^w\Gamma = \lim_{t\to\infty} E({}^w\tilde{\theta}_t^w\tilde{\theta}_t^T)$. We wish to obtain approximate expressions for ${}^v\Gamma$ and ${}^w\Gamma$, denoted ${}^v\bar{\Gamma}$ and ${}^w\bar{\Gamma}$, such that, $\forall\ \mu \in (0, \mu_0]$,

$$|{}^v\Gamma -{}^v \bar{\Gamma}| \leq K\mu^{1/2} \text{ and } |{}^w\Gamma -{}^w \bar{\Gamma}| \leq K\gamma^2\mu^{1/2},$$

where $K < \infty$ is some constant. To that purpose, we expand ${}^v\tilde{\theta}_t$ and ${}^w\tilde{\theta}_t$ to the second-order:

$$\begin{aligned}{}^v\tilde{\theta}_t &= {}^vJ_t^{(0)} + {}^vJ_t^{(1)} + {}^vJ_t^{(2)} + {}^vH_t^{(2)},\\ {}^w\tilde{\theta}_t &= {}^wJ_t^{(0)} + {}^wJ_t^{(1)} + {}^wJ_t^{(2)} + {}^wH_t^{(2)}.\end{aligned}$$

Under the stated assumptions, it follows from Theorems 4 and 5, that there exists some constant $C < \infty$, such that $\forall\ \mu \in (0, \mu_0]$, we have

$$\begin{aligned}&\sup_{t\geq 0}\left|E({}^vJ_t^{(1)}({}^vJ_t^{(2)} + {}^vH_t^{(2)})^T\right| \leq C\ \ \|\phi_0\|_{r(r+\delta)/\delta}\ \rho_r^2(v)\mu^{1/2}\quad \sup_{t\geq 0}\left|E({}^vJ_t^{(0)v}H_t^{(2)})\right|\\ &\leq C\rho_r(v)\|\phi_0\|_{r(r+\delta)/\delta}\mu^{1/2}\sup_{t\geq 0}\left|E({}^wJ_t^{(1)}({}^wJ_t^{(2)} +^w H_t^{(2)})^T\right|\\ &\leq C\gamma^2\mu_r^2(w)\mu^{1/2}\sup_{t\geq 0}\left|E({}^wJ_t^{(0)w}H_t^{(2)})\right| \leq C\gamma^2\mu_r^2(w)\mu^{1/2}.\end{aligned}$$

It remains to evaluate $\lim_{t\to\infty} E({}^vJ_t^{(0)v}J_t^{(i)})$, $\lim_{t\to\infty} E({}^wJ_t^{(0)w}J_t^{(i)})$, $i = 0, 1, 2$, and $E({}^vJ_t^{(1)v}J_t^{(1)})$ and $E({}^wJ_t^{(1)w}J_t^{(1)})$. Denote $\alpha = \sigma_u^2/(1-\kappa^2)$. Tedious but straightforward calculations show that

$$\begin{aligned}\lim_{t\to\infty} E({}^vJ_t^{(0)v}J_t^{(0)^T}) &= \frac{\sigma_v^2}{\mu(2-\mu\alpha)}I,\\ \lim_{t\to\infty} E({}^vJ_t^{(0)v}J_t^{(1)^T}) &= -\frac{\kappa^2\sigma_v^2(d+1)\alpha}{2(1-\kappa^2)}I + O(\mu),\\ \lim_{t\to\infty} E({}^vJ_t^{(0)v}J_t^{(2)^T}) &= \frac{\kappa^2\sigma_v^2\alpha(d+1)\alpha}{4(1-\kappa^2)}I + O(\mu),\\ \lim_{t\to\infty} E({}^vJ_t^{(1)v}J_t^{(1)^T}) &= \frac{(1+\kappa^2)\sigma_v^2\alpha(d+1)}{4(1-\kappa^2)}I + O(\mu),\end{aligned}$$

yielding the following expression for ${}^v\bar{\Gamma}$:

$$ {}^v\bar{\Gamma} = \frac{\sigma_v^2}{2\mu}I + \alpha\sigma_v^2\frac{d+2}{4}I + O(\mu). \tag{47}$$

It is worthwhile to note that the first-order correction does not depend upon the autoregressive coefficient κ, i.e., the dependence among the successive regressors does not influence the covariance ${}^v\bar{\Gamma}$ up to the second order in the step-size μ. It may be shown that this result holds under much weaker assumptions on the regression sequence $\{\phi_t\}$ (see [26] for a more general discussion), as long as $\{v_t\}$ is a martingale increment and is independent from $\{\phi_t\}$. Similarly, it can be shown that

$$\begin{aligned}
&\lim_{t\to\infty} E({}^wJ_t^{(0)}\,{}^wJ_t^{(0)^T}) = \frac{\gamma^2}{\mu\alpha(2-\mu\alpha)}I,\\
&\lim_{t\to\infty} E({}^wJ_t^{(0)}\,{}^wJ_t^{(1)^T}) = 0,\\
&\lim_{t\to\infty} E({}^wJ_t^{(0)}\,{}^wJ_t^{(2)^T}) = \frac{\gamma^2\kappa^2(d+1)}{4(1-\kappa^2)}I + O(\gamma^2\mu),\\
&\lim_{t\to\infty} E({}^wJ_t^{(1)}\,{}^wJ_t^{(1)^T}) = \frac{\gamma^2(1+\kappa^2)(d+1)}{4(1-\kappa^2)}I + O(\gamma^2\mu),
\end{aligned}$$

yielding the following approximate expression for ${}^w\Gamma$:

$$(48)\qquad {}^w\bar{\Gamma} = \frac{\gamma^2}{2\mu\alpha}I + \frac{\gamma^2}{4}\left(1 + (d+1)\frac{1+2\kappa^2}{1-\kappa^2}\right)I + O(\gamma^2\mu).$$

It is interesting to note that the first-order correction depends upon the autoregressive coefficient κ: when κ is close to 1, the correction term becomes large. This behavior is illustrated in Figures 1 and 2, where the asymptotic tracking error variance $\lim_{t\to\infty}\|\tilde{\theta}_t\|^2$ is displayed as a function of the step-size μ. In both cases, we set $\gamma = 0.05$, $d = 10$, $\sigma_v^2 = 3$ and, for every value of κ, $\sigma_u^2 = 1-\kappa^2$ (so that $\alpha = 1$). Two values of κ are considered: $\kappa = 0$ (Figure 1) and $\kappa = 0.9$ (Figure 2). On the figures, the value of the asymptotic tracking error variance obtained by Monte Carlo simulations (solid line) are compared with the approximate expressions obtained by (i) retaining only the first term in (47) and (48) (dashed line) or (ii) including the two terms in (47) and (48) (dashed-dotted line). As seen in these figures, the autoregressive coefficient strongly affects the asymptotic tracking error variance, as predicted by the second-order approximation (whereas the first-order approximation does not predict any effect with respect to the variation of the autoregressive parameter). It is, however, interesting to note that the optimal value for the step-size (the value which minimizes the asymptotic tracking error variance) obtained by minimizing the second-order approximation does not vary with κ and is reasonably close to the one obtained by minimizing the first-order approximation.

The study of the tracking behavior of the RLS algorithm is more involved than for the LMS algorithm. To illustrate our results, we only sketch how the results presented in this paper can be obtained, and we postpone a complete discussion on the potential advantages of the RLS algorithm with respect to the LMS algorithm in a forthcoming paper. For simplicity, we study only the contribution of the lag noise to the total weight-error covariance matrix. The contribution of the measurement noise can be studied along the same lines. The low order terms in the decomposition of the contribution of the lag noise to the weight error may be expressed as

$${}^wJ_{n+1}^{(0)} = \gamma\sum_{k=0}^{n}(1-\mu)^{n-k}w_{k+1},\quad \bar{P}_0 = \bar{P}_s = E(\phi_0\phi_0^T)^{-1} = \alpha^{-1}I,\qquad \forall\, s>0,$$

$$\begin{aligned}
{}^wJ_{n+1}^{(1)} &= \mu\gamma\sum_{s=1}^{n}(1-\mu)^{n-s}Z_s\sum_{k=0}^{s-1}(1-\mu)^{s-k}\,w_{k+1}\\
&= \mu\gamma\sum_{k=0}^{n-1}(1-\mu)^{n-k}\left(\sum_{s=k+1}^{n}Z_s\right)w_{k+1},
\end{aligned}$$

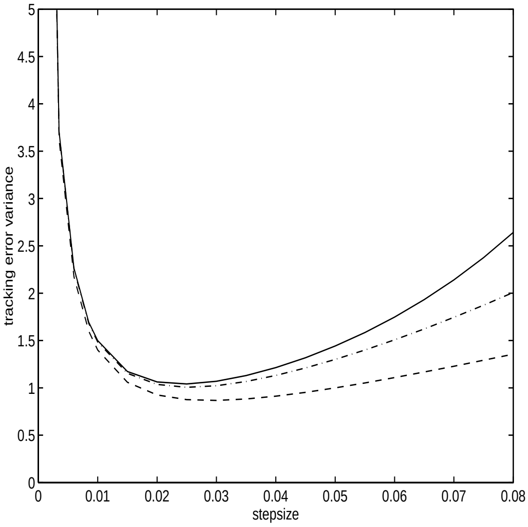


FIG. 1. $\kappa = 0$. *Solid line: Monte Carlo simulation. Dashed line: first-order approximation. Dashed-dotted line: second-order approximation.*

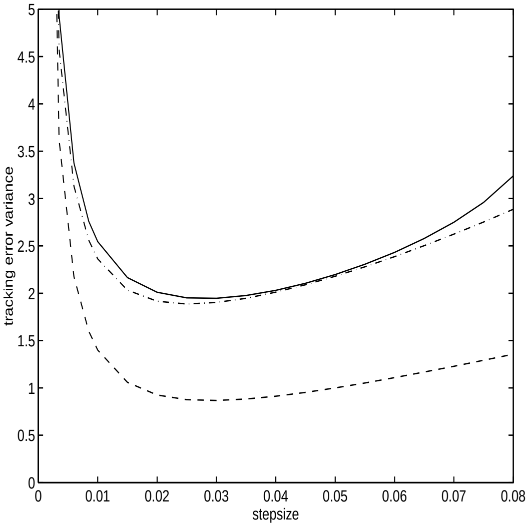


FIG. 2. $\kappa = 0.9$. *Solid line: Monte Carlo simulation. Dashed line: first-order approximation. Dashed-dotted line: second-order approximation.*

where the error term is given by

$$Z_s = \bar{P}_s \, E\{\phi_s \phi_s^T\} - P_s \, \phi_s \phi_s^T .$$

It is easily seen that

$$\lim_{t\to\infty} E({}^w J_t^{(0)} {}^w J_t^{(0)^T}) = \gamma^2/\mu(2-\mu)I.$$

Since $\{\phi_s\}$ is independent of $\{w_s\}$, we may write

$$E[{}^wJ^{(0)}_{n+1}\,{}^wJ^{(1)}_{n+1}{}^T] = \mu\gamma^2 \sum_{k=0}^{n-1}(1-\mu)^{2(n-k)} \sum_{s=k+1}^{n} E[Z_s].$$

Note that this cross-term was equal to zero for the LMS algorithm but does not vanish for the RLS algorithm, because $E(Z_s) \neq 0$. Before going further, we need to compute $E\{Z_s\}$. We proceed as outlined in section 4. We have

$$\begin{aligned} E[Z_s] &= E[(\bar{P}_s - P_s)\phi_s\phi_s^T] \\ &= E[\bar{P}_s(P_s^{-1} - \bar{P}_s^{-1})\bar{P}_s\phi_s\phi_s^T] - E[\bar{P}_s(P_s^{-1} - \bar{P}_s^{-1})^2\bar{P}_s\phi_s\phi_s^T] + O(\mu^{3/2}). \end{aligned}$$

Using tedious but straightforward calculations it may be shown that

$$\begin{aligned} E\{(P_s^{-1} - \bar{P}_s^{-1})\phi_s\phi_s^T\} &= \frac{\mu(d+1)}{1-\kappa^2(1-\mu)}(1-(\kappa(1-\mu))^s), \\ E[(P_s^{-1} - \bar{P}_s^{-1})^2\phi_s\phi_s^T] &= \mu^2[A_1(s) + 2A_2(s)], \\ A_1(s) &= \frac{(d^2+d+2)}{1-\kappa^2(1-\mu)^2}[1-(\kappa(1-\mu))^{2s}] + \frac{d}{\mu}(1-(1-\mu)^{2s}), \\ A_2(s) &= \frac{(d^2+3d+4)}{\mu}\left[\frac{(1-(\kappa^2(1-\mu))^s)}{1-\kappa^2(1-\mu)} - \frac{(1-(\kappa(1-\mu))^{2s})}{1-\kappa^2(1-\mu)^2}\right] \\ &\quad + \frac{d+1}{1-\kappa^2(1-\mu)}\left[\frac{1-(1-\mu)^s}{\mu} - \frac{1-(\kappa(1-\mu))^{2s}}{1-\kappa^2(1-\mu)^2}\right]. \end{aligned}$$

Thus,

$$\begin{aligned} \lim_n \mu^3\gamma^2\sum_{k=0}^{n-1}(1-\mu)^{2(n-k)}\sum_{s=k+1}^{n} A_1(s) &= \mu\gamma^2\frac{d^2+d+2}{4(1-\kappa^2(1-\mu)^2)}I + O(\gamma^2), \\ \lim_n \mu^3\gamma^2\sum_{k=0}^{n-1}(1-\mu)^{2(n-k)}\sum_{s=k+1}^{n} A_2(s) &= \frac{\gamma^2}{2}(d^2+3d+4) \\ &\quad \left[\frac{1}{1-\kappa^2(1-\mu)} - \frac{1}{1-\kappa^2(1-\mu)^2}\right] \\ &\quad + \frac{\gamma^2}{2}\frac{d+1}{(1-\kappa^2(1-\mu))}\left[1 - \frac{\mu}{1-\kappa^2(1-\mu)^2}\right] \\ &\quad + O(\mu+\gamma^2), \end{aligned}$$

yielding the following expression for the asymptotic weight-error covariance matrix $\bar{\Gamma}^w$:

$$\begin{aligned} \bar{\Gamma}^w = \frac{\gamma^2}{2\mu}I &+ \frac{\gamma^2(d+1)}{(1-\kappa^2(1-\mu))}I - \gamma^2(d^2+3d+4)\left[\frac{1}{1-\kappa^2(1-\mu)} - \frac{1}{1-\kappa^2(1-\mu)^2}\right]I \\ &- \frac{\gamma^2(d+1)}{1-\kappa^2(1-\mu)}\left[1/2 - \frac{\mu}{1-\kappa^2(1-\mu)^2}\right]I + O(\mu+\gamma^2). \end{aligned}$$

As before, this correction term depends upon the autoregressive coefficient: when κ is close to 1 (e.g., when the regressors are strongly positively correlated), the corrections terms can become very large. Also, it is interesting to note that the magnitude of the correction term is proportional to the square of the size of the regressors. This also shows that the correction can become very significant when d is large.

Appendix A. Proof of Proposition 2. In this section, we will show some useful extensions of Rosenthal's inequality for weakly dependent processes.

PROPOSITION 6. *Let $G = \{G_t\}_{t\geq 0}$ be a $(d \times d)$ zero-mean matrix-valued process. Let q be an* even *integer and $j \in \mathbb{N}$. Assume that G is (δ, qj)-weak dependent. Assume in addition that*

$$\sum_{r=0}^{\infty}(r+1)^{qj/2-1}\delta(r) < \infty. \tag{49}$$

Then, there exists a finite constant $\bar{D}_{q,j}(G)$, such that

$$\left\| \sum_{s\leq i_1<\cdots<i_j\leq t} G_{i_1}\cdots G_{i_j} \right\|_q \leq \bar{D}_{q,j}(G)\ (t-s)^{j/2} \tag{50}$$

$\forall\ 0 \leq s \leq t$.

Proof of Proposition 6. The proof is a direct application of the following result.

LEMMA 7. *Let $q \geq 2$ and let $X = \{X_t\}_{t\geq 0}$ be an $(l \times 1)$ zero-mean vector-valued (δ, q)-weak dependent process. Assume in addition that $\sum_0^\infty (r+1)^{q/2-1}\delta(r) < \infty$. Then, there exist finite constants $\gamma = \{\gamma_2, \dots, \gamma_q\}$ such that, $\forall\ 1 \leq s \leq q$ and $\forall\ n \geq 1$,*

$$\sup_{(i_1,\cdots,i_s)} \left(\sum_{1\leq t_1,\cdots,t_s\leq n} |E(X_{t_1,i_1}\cdots X_{t_s,i_s})| \right)\ \leq\ \gamma_s\ s!\ n^{s/2}. \tag{51}$$

Remark. The constants $\gamma_2, \dots, \gamma_q$ can be evaluated recursively as follows. Let $\sigma_s = \sum_{r=0}^\infty (r+1)^{s/2-1}\delta(r)$ for $1 < s \leq q$. Set $\gamma_1 = 0$, $\gamma_2 = C_2\sigma_2$ and evaluate, for $2 < s \leq q$,

$$\gamma_s = \sum_{m=1}^{s-1} \gamma_m\gamma_{s-m} + (s-1)C_s\sigma_s. \tag{52}$$

Proof of Lemma 7. The proof is adapted from [7]. Define, for $1 < s \leq q$,

$$I(n,s) = \{(t_1, t_2, \dots, t_s) :\ 1 \leq t_1 \leq \cdots \leq t_s \leq n\}, \tag{53}$$

$$A(n,s) = \sup_{(i_1,\dots,i_s)} \sum_{I(n,s)} |E(X_{t_1,i_1}\dots X_{t_s,i_s})|. \tag{54}$$

Note that

$$\sup_{(i_1,\dots,i_s)} \sum_{1\leq t_1,\dots,t_s\leq n} |E(X_{t_1,i_1}\dots X_{t_s,i_s})| \leq s!A(n,s).$$

Define for $1 \leq m \leq s-1$ and $0 \leq r \leq n-1$ the sets

$$I(n,s,m,r) = \{(t_1, \dots, t_s) \in I(n,s) :\ t_{m+1} - t_m = r = \max(t_{i+1} - t_i)\},$$

$$I(n,s,m) = \bigcup_{r=0}^{n-1} I(n,s,m,r).$$

It is easily seen that

$$I(n,s) = \bigcup_{m=1}^{s-1} I(n,s,m) = \bigcup_{m=1}^{s-1} \bigcup_{r=0}^{n-1} I(n,s,m,r)$$

and the cardinal of set $I(n,s,m,r)$ is bounded by $n(r+1)^{s-2}$. Let $1 \leq m \leq s-1$. Note that

$$\begin{aligned} E(X_{t_1,i_1} \dots X_{t_s,i_s}) &= E(X_{t_1,i_1} \dots X_{t_m,i_m}) \\ &\quad \times E(X_{t_{m+1},i_{m+1}} \dots X_{t_s,i_s}) + \operatorname{cov}(X_{t_1,i_1} \dots X_{t_m,i_m}, X_{t_{m+1},i_{m+1}} \dots X_{t_s,i_s}). \end{aligned}$$

This implies, under the weak-dependence condition, that

$$\sup_{(i_1,\dots,i_s)} \sum_{I(n,s,m)} |E(X_{t_1,i_1} \dots X_{t_s,i_s})| \leq A(n,m)A(n,s-m) + C_s \sum_{r=0}^{n-1} \delta(r) n (r+1)^{s-2}.$$

For $0 \leq r \leq n-1$, $n(r+1)^{s-2} \leq n^{s/2}(r+1)^{s/2-1}$. Thus

$$A(n,s) \leq \sum_{m=1}^{s-1} A(n,m)A(n,s-m) + (s-1)C_s n^{s/2} \sum_{r=0}^{n-1} (r+1)^{s/2-1}\delta(r). \tag{55}$$

The proof and the expression of the constant are obtained by a straightforward induction. □

COROLLARY 8. *Let $p \geq 1$ and $n \in \mathbb{N}$. Let $G = \{G_t\}_{t \geq 0}$ be a $(d \times d)$ matrix-valued process. Assume that $\{G_t\}_{t\geq 0}$ is $(\delta, (p+2)n)$-weak dependent and that*

$$\sum (r+1)^{(p+2)n/2-1} \delta(r) < \infty.$$

Then, there exists a finite constant $D_{p,n}(G)$, such that $\forall\ j \in \{1,\dots,n\}$ and $\forall\ 0 \leq s \leq t < \infty$, we have

$$\left\| \sum_{s \leq i_1 < \cdots < i_j \leq t} (G_{i_1} - E(G_{i_1})) \dots (G_{i_j} - E(G_{i_j})) \right\|_{pn/j} \leq D_{p,n}(G)(t-s)^{j/2}. \tag{56}$$

Proof of Corollary 8. For $j \in \{1,\dots,n\}$, denote $q(n,j)$ the smallest even integer verifying $pn/j \leq q(n,j)$. It is easily seen that $pn/j \leq q(n,j) \leq pn/j+2$, which implies that $\max_{j \in \{1,\dots,n\}} jq(n,j) \leq (p+2)n$. Under the stated assumption, Proposition 6 implies that, $\forall\ 0 \leq s \leq t$,

$$\begin{aligned} \left\| \sum_{s \leq i_1 < \cdots < i_j \leq t} G_{i_1} \dots G_{i_j} \right\|_{pn/j} &\leq \left\| \sum_{s \leq i_1 < \cdots < i_j \leq t} G_{i_1} \dots G_{i_j} \right\|_{q(n,j)}, \\ &\leq \bar{D}_{q(n,j),j}(G)(t-s)^{j/2} \end{aligned}$$

which concludes the proof. □

In what follows, we will have to work with sums of products of the form

$$\sum_{s \leq i_1 < i_2 < \cdots < i_j \leq t} (P_{i_1} G_{i_1} - \bar{P}_{i_1} E(G_{i_1})) \dots (P_{i_j} G_{i_j} - \bar{P}_{i_j} E(G_{i_j})),$$

where $\{P_i\}_{i\in\mathbb{N}}$ is a matrix-valued random process (indexed by μ), and $\{\bar{P}_i\}_{i\in\mathbb{N}}$ is a set of deterministic matrices such that $\sup_{i\in\mathbb{N}}\|P_i-\bar{P}_i\|_c = O(\sqrt{\mu})$ as $\mu\to 0^+$. It is easily seen that $E(P_iG_i)\neq \bar{P}_iE(G_i)$ and that $\{P_iG_i\}_{i\geq 0}$ is not weak-dependent and the proposition above cannot be directly applied. An extension (based on the fact that $\{P_i\}$ is well "approximated" by a set of deterministic matrices) is given below.

PROPOSITION 9. *Let $p\geq 1$ and $n\in\mathbb{N}$. Let $G=\{G_t\}_{t\geq 0}$ be a $(d\times d)$ matrix-valued process. Assume that $\{G_t\}_{t\geq 0}$ is $(\delta,(p+2)n)$-weak dependent and*

$$\sum (r+1)^{(p+2)n/2-1}\delta(r)<\infty.$$

Assume in addition that (i) *there exist $\mu_0>0$ and a constant $C_{p,n}(\mu_0,P)<\infty$ such that* $\sup_{t\geq 0}\|P_t-\bar{P}_t\|_{\alpha pn}<C_{p,n}(\mu_0,P)\sqrt{\mu}$ *and* (ii) $\sup_{t\geq 0}\|G_t\|_{\beta pn}<\infty$, *where $\alpha,\beta>0$ and $\alpha^{-1}+\beta^{-1}=1$. Then, there exists a finite constant $D_{p,n}(G)$ such that $\forall\ j\in\{1,\dots,n\}$ and $\forall\ 0\leq s\leq t<\infty$, we have*

$$\left\|\sum_{s\leq i_1<\cdots<i_j\leq t}(P_{i_1}G_{i_1}-\bar{P}_{i_1}E(G_{i_1}))\cdots(P_{i_j}G_{i_j}-\bar{P}_{i_j}E(G_{i_j}))\right\|_{pn/j} \leq D_{p,n}(G)(t-s)^{j/2}\sum_{l=0}^{j}\mu^{l/2}(t-s)^{l/2}. \tag{57}$$

Proof of Proposition 9. Denote $\bar{G}_i=E(G_i)$, and, for $t\geq s$,

$$D_1(t,s)=\sum_{u=s}^{t}(P_uG_u-\bar{P}_u\bar{G}_u),\quad \tilde{D}_1(t,s)=\sum_{u=s}^{t}\bar{P}_u(G_u-\bar{G}_u), \tag{58}$$

$$D_k(t,s)=\sum_{u=s}^{t}D_{k-1}(t,u+1)(P_uG_u-\bar{P}_u\bar{G}_u), \tag{59}$$

$$\tilde{D}_k(t,s)=\sum_{u=s}^{t}\tilde{D}_{k-1}(t,u+1)\bar{P}_u(G_u-\bar{G}_u).$$

By convention, we set $D_k(t,s)=\tilde{D}_k(t,s)=0$ for $s<t+k$ and $D_0(t,s)=\tilde{D}_0(t,s)=0\ \forall\ t,s$. Since $\sup_{0\leq\mu\leq\mu_0}\sup_{u\geq 0}|\bar{P}_u|<\infty$, and $\{G_u-\bar{G}_u\}_{u\geq 0}$ is $(\delta,p(n+2))$-weak dependent, then $\{\bar{P}_u(G_u-\bar{G}_u)\}_{u\geq 0}$ also is $(\delta,p(n+2))$-weak dependent, so that, by application of Proposition 2, there exists a constant $D_{p,n}(\bar{P}G)$ (*independent* from μ), such that, for $j\in\{1,\dots,n\}$ and all $0\leq s\leq t<\infty$, it holds that

$$\|\tilde{D}_j(t,s)\|_{pn/j}\leq D_{p,n}(\bar{P}G)(t-s)^{j/2}. \tag{60}$$

The proof is by induction. We have

$$\sum_{s\leq u\leq t}(P_uG_u-\bar{P}_u\bar{G}_u)=\sum_{s\leq u\leq t}(P_u-\bar{P}_u)G_u+\sum_{s\leq u\leq t}\bar{P}_u(G_u-\bar{G}_u).$$

Thus,

$$\left\|\sum_{u=s}^{t}(P_uG_u-\bar{P}_u\bar{G}_u)\right\|_{pn}\leq (t-s)\sup_{u\geq 0}\|P_u-\bar{P}_u\|_{\alpha qn}\sup_{u\geq 0}\|G_u\|_{\beta qn}+D_{p,n}(\bar{P}G)(t-s)1/2$$

showing that there exists a constant $C_1 < \infty$ such that $\|D_1(t,s)\|_{pn} \le C_1(t-s)^{1/2}(1+\sqrt{\mu}(t-s)^{1/2})$ $\forall$ $0 \le s \le t$ and $\mu \in (0,\mu_0]$. Let $j \in \{1,\dots,n-1\}$. Assume that for $i \in \{1,\dots,j-1\}$, there exists $C_i < \infty$ such that, $\forall$ $0 \le s \le t$ and $\mu \in (0,\mu_0]$, we have $\|D_i(t,s)\|_{pn/i} \le C_i(t-s)^{i/2}\sum_{l=0}^{i}\mu^{l/2}(t-s)^{l/2}$. It holds that

$$D_j(t,s) = \tilde{D}_j(t,s) + \sum_{r=1}^{j}\sum_{i_r=s}^{t} D_{r-1}(i_r,s)(P_{i_r} - \bar{P}_{i_r})G_{i_r}\tilde{D}_{j-r}(t,i_{j-1}+1),$$

$$\begin{aligned}\|D_j(t,s)\|_{pn/j} \le\ & \|\tilde{D}_j(t,s)\|_{pn/j} \\ & + \sum_{r=1}^{j}\sum_{i_r=s}^{t} \|D_{r-1}(i_r,s)\|_{pn/(r-1)}\|\tilde{D}_{j-r}(t,i_r+1)\|_{pn/(j-r)}\|P_{i_r} \\ & - \bar{P}_{i_r}\|_{\alpha pn}\|G_{i_r}\|_{\beta pn}, \\ & \le D_{p,n}(\bar{P}G)\left((t-s)^{1/2} + \sum_{r=1}^{j} C_r(t-s)^{j/2}\sum_{l=0}^{r}\mu^{l/2}(t-s)^{l/2}\right),\end{aligned}$$

which concludes the proof. □

Proof of Proposition 2. It is assumed (without any loss of generality) that $\{D_k\}_{k\ge0}$ and $\{\epsilon_k\}_{k\ge0}$ are scalar-valued. (i) Assume first that $\{\epsilon_k\}_{k\ge0}$ is an L_p stable martingale increment. According to Burkholder's inequality, there exists a universal constant $B_p < \infty$ (independent of the sequence of scalar weights $\{D_k\}_{k\ge0}$ and on the process $\{\epsilon_k\}_{k\ge0}$) such that

$$\begin{aligned}E\left|\sum_{k=s}^{t} D_k\epsilon_k\right|^p & \le B_p\left\|\sum_{k=s}^{t} D_k^2\epsilon_k^2\right\|_{p/2}^{p/2} \\ & \le B_p \sup_k \|\epsilon_k\|_p^p \left(\sum_{k=s}^{t} D_k^2\right)^{p/2},\end{aligned}$$

which concludes the proof for martingale increments. (ii) Assume that q is an even integer. For $n \ge 0$, we have

$$\begin{aligned}E\left|\sum_{k=1}^{n} D_k\epsilon_k\right|^q & = \sum_{1\le i_1,\dots,i_q\le n} E(D_{i_1}\dots D_{i_p}\epsilon_{i_1}\dots\epsilon_{i_p}), \\ & \le \sum_{1\le i_1,\dots,i_q\le n} |D_{i_1}\dots D_{i_q}||E(\epsilon_{i_1}\dots\epsilon_{i_q})|, \\ & \le q!\ A_q(n),\end{aligned}$$

where $A_q(n) \triangleq \sum_{1\le i_1\le\dots\le i_q\le n}|D_{i_1}\dots D_iq||E(\epsilon_{i_1}\dots\epsilon_{i_q})|$. We will conclude the proof by showing that $A_q(n) \le C\sum_{i=1}^{n}(|D_i|^2)^{q/2}$, uniformly in n, where $C < \infty$ is a constant which is related to the weak-dependence constants of the process ϵ (see the definition above), i.e., $C_1,\dots,C_q$ and $\{\delta(k)\}_{k\ge0}$. The proof is by induction. Let

$1 \leq s \leq q$. We have

$$A_s(n) = \sum_{m=1}^{s-1} \sum_{r=0}^{n-1} \sum_{I(n,s,m,r)} |D_{i_1} \dots D_{i_s}| \, |E(\epsilon_{i_1} \dots \epsilon_{i_m})| \, |E(\epsilon_{i_{m+1}} \dots \epsilon_{i_s})| \tag{61}$$

$$+ V_s(n) \tag{62}$$

$$\leq \sum_{m=1}^{s-1} A_m(n) A_{s-m}(n) + V_s(n), \tag{63}$$

$$V_s(n) = \sum_{m=1}^{s-1} \sum_{r=0}^{n-1} V_s(n, r, m), \tag{64}$$

$$V_s(n, r, m) = \sum_{I(n,s,m,r)} |D_{i_1} \dots D_{i_s}| |\mathrm{cov}(\epsilon_{i_1} \dots \epsilon_{i_m}, \epsilon_{i_{m+1}} \dots \epsilon_{i_s})|. \tag{65}$$

Denote $b_r = \max_{1 \leq t \leq n} \sum_{i=t+1}^{t+r} |D_i|$. We have

$$\begin{aligned} V_s(n, r, m) \leq C_s \sum_{i_m=1}^{n-r} |D_{i_m}||D_{i_m+r}| \delta(r) \sum_{i_{m-1}=\max(i_m-r,0)}^{i_m} |D_{i_{m-1}}| \dots \sum_{i_1=\max(i_2-r,0)}^{i_2} |D_{i_1}| \\ \times \sum_{i_{m+2}=i_{m+1}}^{\min(i_{m+1}+r,n)} |D_{i_{m+2}}| \dots \sum_{i_s=i_{s-1}}^{\min(i_{s-1}+r,n} |D_{i_s}|, \\ \leq C_s b_r^{s-2} \delta(r) \sum_{i=1}^{n} |D_i|^2. \end{aligned}$$

Note that the bound for $V_s(n, r, m)$ does not depend on m. Hence,

$$V_s(n) \leq s \; C_s \sum_{r=0}^{n-1} \delta(r) b_r^{q-2} \sum_{i=1}^{n} |D_i|^2. \tag{66}$$

We have

$$b_r \leq \sqrt{r} \max_{1 \leq t \leq n} \left(\sum_{i=1}^{t+r} |D_i|^2 \right)^{1/2} \leq \left(\sum_{i=1}^{n} |D_i|^2 \right)^{1/2} \sqrt{r}.$$

Plugging this latter relation into (66) yields

$$V_s(n) \leq \left(\sum_{i=1}^{n} |D_i|^2 \right)^{s/2} \sum_{r=0}^{n-1} (r+1)^{s/2-1} \delta(r).$$

The proof is now concluded by an easy induction.

Appendix B. Proof of Theorems 4 and 5.

Proof of Theorem 4. Before going further, we need some additional notation. Define $S_0(t, s) = \psi(t, s)$ for $t \geq s$ and $S_0(t, s) = 0$ for $s > t$ (the dependence in the step-size μ is implicit). For $k \geq 1$, define $S_k(t, s)$ recursively as

$$S_k(t, s) = \sum_{u=s}^{t} \psi(t, u) Z_u S_{k-1}(u-1, s) = \sum_{u=s}^{t} S_{k-1}(t, u) Z_u \psi(u-1, s). \tag{67}$$

Denote, respectively,

$$D_1(t,s) = \sum_{u=s}^{t} Z_u, \quad t \geq s, \quad D_1(t,s) = 0, \quad s > t, \tag{68}$$

$$D_k(t,s) = \sum_{u=s}^{t} D_{k-1}(t,u+1) Z_u. \tag{69}$$

Note that, by construction, for $k \geq 0$,

$$S_k(t,s) = 0, \quad s > t-k, \quad \text{and } D_k(t,s) = 0, \quad s > t-k+1.$$

From (20), (21), and (67), it is easily seen that $J_{t+1}^{(r)}$ may be decomposed as

$$J_{t+1}^{(r)} = \mu^r \sum_{s=0}^{t} S_r(t,s) \xi_s. \tag{70}$$

Let r be an integer and c, μ_0, β some real numbers such that $c > 0$, $\mu_0 > 0$, and $0 < \beta < 1/\mu_0$. Consider $W(\mu) \triangleq \{W(u,v;\mu)\}_{(u,v)\in\mathbb{N}\times\mathbb{N}}$ a family of processes on $\mathbb{N}\times\mathbb{N}$ indexed by $\mu \in \mathbb{R}^+$. We say that $W(\mu)$ belongs to the set $\mathcal{L}(\beta,\mu_0,c,r)$ if there exist a *finite* constant $C(W)$ and a *finite* integer $q(W)$, such that, $\forall\ 0 \leq u \leq v$ and $0 < \mu \leq \mu_0$

$$\|W(u,v;\mu)\|_c \leq C(W)(1-\beta\mu)^{v-u}(v-u)^{r/2} \sum_{k=0}^{q(W)} \mu^{k/2}(v-u)^{k/2}. \tag{71}$$

It is easily seen that the subspaces $\mathcal{L}(\beta,\mu_0,c,r)$ are nested in the sense that for $c' \geq c$ and $r' \geq r$, we have $\mathcal{L}(\beta,\mu_0,c,r) \subset \mathcal{L}(\beta,\mu_0,c',r')$. Similarly, we say that $W(\mu)$ belongs to the $\mathcal{M}(\mu_0,c,r)$ if there exist a finite constant $C_{\mathcal{M}}(W)$ and a finite integer $q(W)$, such that, $\forall\ 0 \leq u \leq v$ and $0 < \mu \leq \mu_0$,

$$\|W(u,v;\mu)\|_c \leq C_{\mathcal{M}}(W)(v-u)^{r/2} \sum_{j=0}^{q(W)} \mu^{j/2}(v-u)^{j/2}. \tag{72}$$

LEMMA 10. *Assume that $S_r \triangleq \{S_r(t,s)\} \in \mathcal{L}(\beta,\mu_0,q,r)$ and $\xi \triangleq \{\xi_k\} \in \mathcal{N}(p)$. Then, there exists a constant $C(S_r) < \infty$ (depending upon $\beta,\mu_0,p,q,r,$ and $\{F_k\}$ but not on $\{\xi_k\}$ or on the step-size μ) such that*

$$\sup_{t\geq 0} \|J_t^{(r)}\|_p \leq C(S_r)\ \rho_p(\xi)\mu^{(r-1)/2} \qquad \forall\ \mu\ \in\ (0,\mu_0].$$

Proof of Lemma 10. Since $\{\xi_t\} \in \mathcal{N}(p)$, there exists a constant $\rho_{p,q}(\xi)$ such that

$$\|J_{t+1}^{(r)}\|_p \leq \rho_{p,q}(\xi)\mu^r \left(\sum_{s=0}^{t} \|S_r(t,s)\|_q^2\right)^{1/2}.$$

Under the stated assumptions, $S_r(t,s) \in \mathcal{L}(\beta,\mu_0,q,r)$: there exist a constant $C(S_r) < \infty$ and $q(S_r) \in \mathbb{N}$ such that, $\forall\ t \geq s \geq 0$,

$$\|S_r(t,s)\|_q \leq C(S_r)\ (1-\beta\mu)^{t-s}(t-s)^{r/2} \sum_{k=0}^{q(S_r)} \mu^{k/2}(t-s)^{k/2} \qquad \forall\ \mu\ \in\ (0,\mu_0].$$

This latter relation implies that

$$\sum_{s=0}^{t} \|S_r(t,s)\|_q^2 \leq q(S_r)\, C^2(S_r) \sum_{k=0}^{q(S_r)} \mu^k \sum_{u=0}^{t} (1-\beta\mu)^{2u} u^{k+r} \qquad \forall\, \mu \,\in\, (0,\mu_0].$$

For all $\alpha \geq 0$, there exists a constant $D_\alpha < \infty$, such that, $\forall\ 0 < \mu \leq 1/\beta$,

$$\sum_{u=0}^{\infty} (1-\beta\mu)^{2u} u^\alpha \leq D_\alpha/(\beta\mu)^{1+\alpha}$$

which concludes the proof. □

The previous lemma will allow us to conclude the proof of Theorem 4 if we are able to prove that

$$\{S_r(t,s)\} \ \in \mathcal{L}(\beta,\mu_0,q,r), \text{ for } r \in \{1,\dots,n\}.$$

In fact, we will prove a slightly stronger property: $S_r(t,s) \in \mathcal{L}(\beta,\mu_0,qn/r,r)$, for $r \in \{1,\dots,n\}$. This part is more intricate and requires some preparatory lemmas.

LEMMA 11. *Let $j \in \{1,\dots,r-1\}$. Denote $\Delta_j(v,u) = D_j(v,u) - D_j(v,u+1)$. We have, for $t \geq s$ and for $j \in \{1,\dots,r-1\}$,*

$$\begin{aligned}&\sum_{u=s}^{t} \psi(t,u)\Delta_j(t,u)S_{r-j}(u-1,s) - \sum_{u=s}^{t} \psi(t,u)\Delta_{j+1}(t,u)S_{r-j-1}(u-1,s)\\ &= \mu\sum_{u=s}^{t} \psi(t,u)\bar{F}_{u-1}D_j(t,u)S_{r-j}(u-2,s) - \mu\sum_{u=s}^{t} \psi(t,u)D_j(t,u)\bar{F}_{u-1}S_{r-j}(u-2,s).\end{aligned}$$

In addition,

$$\begin{aligned}&\sum_{u=s}^{t} \psi(t,u)\Delta_r(t,u)\psi(u-1,s) - \psi(t,s+1)D_r(t,s+1)\\ &\qquad = \mu\sum_{u=s}^{t} \psi(t,u)\bar{F}_{u-1}D_r(t,u)\psi(u-2,s) - \mu\sum_{u=s}^{t} \psi(t,u)D_r(t,u)\bar{F}_{u-1}\psi(u-2,s).\end{aligned}$$

Proof of Lemma 11. The proof basically amounts to applying the Abel transform recursively. Hence, it involves only simple algebraic manipulations. First note that

$$\begin{aligned}\sum_{u=s}^{t} \psi(t,u)\Delta_j(t,u)S_{r-j}(u-1,s) &= \sum_{u=s}^{t} \left(\psi(t,u) - \psi(t,u-1)\right) D_j(t,u)S_{r-j}(u-2,s)\\ &\quad + \sum_{u=s}^{t} \psi(t,u)D_j(t,u)\left(S_{r-j}(u-1,s) - S_{r-j}(u-2,s)\right).\end{aligned}$$

For $t \geq s$ we have

$$\begin{aligned}&\psi(t,s) - \psi(t,s-1) = \mu\psi(t,s)\bar{F}_{s-1},\\ &\psi(t,s) - \psi(t-1,s) = \begin{cases} -\mu\bar{F}_t\psi(t-1,s), & t > s,\\ I, & t = s,\end{cases}\\ &S_k(t,s) - S_k(t-1,s) = Z_tS_{k-1}(t-1,s) - \mu\bar{F}_tS_k(t-1,s).\end{aligned}$$

The proof of this lemma is concluded by noting that

$$D_j(t,u)Z_{u-1} = D_{j+1}(t,u-1) - D_{j+1}(t,u). \qquad \square$$

LEMMA 12. *Let $r \in \mathbb{N}$ and $j \in \{1,\dots,r-1\}$ and let $V(\mu) \triangleq \{V(u,v;\mu)\}_{(u,v)\in\mathbb{N}\times Nset}$ and $W(\mu) \triangleq \{W(u,v;\mu)\}_{(u,v)\in\mathbb{N}\times\mathbb{N}}$ be such that $V(\mu) \in \mathcal{M}(\mu_0, cr/j, j)$ and $W \in \mathcal{L}(\beta,\mu_0,cr/(r-j), r-j)$, for some $c,\mu_0 > 0$, $0 < \beta < 1/\mu_0$. Then, the process $U(\mu) \triangleq \{U(t,s;\mu)\}_{(u,v)\in\mathbb{N}\times\mathbb{N}}$ defined by*

$$U(t,s;\mu) = \mu \sum_{u=s}^{t-1} \psi(t,u) V(t,u;\mu) W(u,s;\mu)$$

belongs to the set $\mathcal{L}(\beta,\mu_0,c,r)$.

Proof of Lemma 12. The proof is elementary. Let $t \ge s \ge 0$. Since $\{F_t\} \in \mathcal{S}(\beta,\mu_0)$, $|\psi(t,s)| \le K_{\beta,\mu_0}(F)(1-\beta\mu)^{t-s}$ $\forall$ $\mu \in (0,\mu_0]$ and $0 \le s \le t$. This implies

$$\|U(t,s;\mu)\|_c \le K_{\beta,\mu_0}(F)\mu \sum_{u=s}^{t} (1-\beta\mu)^{t-u} \|V(t,u;\mu)\|_{cr/j} \|W(u,s;\mu)\|_{cr/(r-j)}$$
$$\le K_{\beta,\mu_0}(F)\mu \sup_{s\le u\le t} \left(\|V(t,u;\mu)\|_{cr/j}\right) \sup_{s\le u\le t} \left(\|W(u,s;\mu)\|_{cr/(r-j)}/(1-\beta\mu)^{u-s}\right)$$
$$(t-s)(1-\beta\mu)^{t-s}.$$

Under the stated assumptions, there exist constants $C(V) < \infty$, $q(V) \in \mathbb{N}$, $C(W) < \infty$, and $q(W) \in \mathbb{N}$ such that, $\forall$ $0 \le s \le t$ and $\forall$ $0 < \mu \le \mu_0$,

$$\sup_{s\le u\le t} \|V(t,u;\mu)\|_{cr/j} \le C(V)(t-s)^{j/2} \sum_{l=0}^{q(V)} \mu^{l/2}(t-s)^{l/2},$$

$$\sup_{s\le u\le t} \left(\|W(u,s;\mu)\|_{cr/(r-j)}/(1-\beta\mu)^{u-s}\right) \le C(W)(t-s)^{(r-j)/2} \left(\sum_{k=0}^{q(W)} \mu^{k/2}(t-s)^{k/2}\right),$$

so that

$$\|U(t,s;\mu)\|_c \le C(U)(1-\beta\mu)^{t-s}(t-s)^{r/2} \sum_{l=0}^{q(W)+q(V)+2} \mu^{l/2}(t-s)^{l/2}$$

for some finite constant $C(U)$. $\square$

By combining the two preceding lemmas, we obtain the following useful criterion.

LEMMA 13. *Let $c > 0$ and $r \in \mathbb{N}$. Assume in addition that $\{D_j(t,s)\} \in \mathcal{M}(cr/j, j)$ for $1 \le j \le r$ and $\{S_j(t,s)\} \in \mathcal{L}(\beta,\mu_0,cr/j,j)$ for $1 \le j < r$. Then, $\{S_r(t,s)\} \in \mathcal{L}(\beta,\mu_0,c,r)$.*

Proof of Lemma 13. By iterated application of Lemma 11, we have for $t \ge s \ge 0$

$$S_r(t,s) = \sum_{u=s}^{t} \psi(t,u)\Delta_r(t,u)\psi(u-1,s) + \sum_{j=1}^{r-1} G_j(t,s), \tag{73}$$

where for $1 \le j \le r-1$, the processes $G_j(t,s)$ are defined as

$$G_j(t,s) = \mu \sum_{u=s}^{t} \psi(t,u)\bar{F}_{u-1} D_j(t,u) S_{r-j}(u-2,s)$$
$$- \mu \sum_{u=s}^{t} \psi(t,u) D_j(t,u) \bar{F}_{u-1} S_{r-j}(u-2,s).$$

Under the stated assumptions, Lemma 12 shows that $G_j(t,s) \in \mathcal{L}(\beta,\mu_0,c,r)$ for $1 \leq j \leq r-1$. The first term on the RHS of (73) may be decomposed (Lemma 11) as

$$\sum_{u=s}^{t} \psi(t,u)\Delta_r(t,u)\psi(u-1,s) = \psi(t,s+1)D_r(t,s+1)$$
$$+ \mu \left(\sum_{u=s}^{t} \psi(t,u)\bar{F}_{u-1}D_r(t,u)\psi(u-2,s) - \sum_{u=s}^{t} \psi(t,u)D_r(t,u)\bar{F}_{u-1}\psi(u-2,s) \right).$$

Since $D_r(t,s) \in \mathcal{M}(\mu_0,c,r)$, $\|D_r(t,s)\|_c \leq C(D_r)(t-s)^{r/2}\sum_{l=0}^{q(D_r)}\mu^{l/2}(t-s)^{l/2}$ for some $C(D_r) < \infty$, $q(D_r) \in \mathbb{N}$, $\forall\ t \geq s \geq 0$ and all $0 < \mu \leq \mu_0$. Thus,

$$\left\| \sum_{u=s}^{t} \psi(t,u)\Delta_r(t,u)\psi(u-1,s) \right\|_c$$
$$\leq K_{\beta,\mu_0}(F)C(D_r)(1-\beta\mu)^{t-s}(t-s)^{r/2}\left(\sum_{l=0}^{q(D_r)}\mu^{l/2}(t-s)^{l/2}\right)\left(1+\mu\sup_{s\geq 0}|\bar{F}_s|(t-s)\right)$$

$\forall\ t \geq s \geq 0$ and all $0 < \mu \leq \mu_0$, which concludes the proof. □

Under assumption (ii) of Theorem 4, an application of Proposition 2 shows that $\{D_r(t,s)\} \in \mathcal{M}(qn/r,r)$, $r \in \{1,\ldots,n\}$; Lemma 13 leads us to a condition upon which $S_r(t,s)$ belongs to $\mathcal{L}(\beta,\mu_0,qn/r,r)$, $r \in \{1,\ldots,n\}$.

LEMMA 14. *Under the assumptions of Theorem* 4, *it holds that*

$$\textit{for}\ \ r \in \{1,\ldots,n\},\quad \{S_r(t,s)\} \in \mathcal{L}(\beta,\mu_0,qn/r,r).$$

Proof of Lemma 14. The proof is by induction on r. By application of Proposition 2, we have $D_1(t,s) \in \mathcal{M}(qn,1)$. This implies by Lemma 13 that $S_1(t,s) \in \mathcal{L}(\beta,\mu_0,qn,1)$. Assume now that the property is verified up to order $r-1$ with $1 < r \leq n$. Set $c = qn/r$. By application of Proposition 2, we have $D_j(t,s) \in \mathcal{M}(qn/j,j) = \mathcal{M}(cr/j,j)$. The induction hypothesis implies that $S_j(t,s) \in \mathcal{L}(\beta,\mu_0,qn/j,j) = \mathcal{L}(\beta,\mu_0,cr/j,j)$ for $1 \leq j < r$. We have $S_r(t,s) \in \mathcal{L}(\beta,\mu_0,c,r) = \mathcal{L}(\beta,\mu_0,qn/r,r)$ by Lemma 13, which concludes the proof. □

The proof of Theorem 4 is concluded by applying Lemma 10 and Lemma 14.

Proof of Theorem 5. Solving recursively the difference equation (22), we may express $H_{t+1}^{(n+1)}$ as a linear combination of $J_s^{(n+1)}$, with random matrix-valued weights $\Phi(t,s)$

$$H_{t+1}^{(n+1)} = \mu\sum_{s=0}^{t}\Phi(t,s)Z_sJ_s^{(n+1)}. \tag{74}$$

Since $\{F_t\} \in \mathcal{S}(a,\beta,\mu_0)$ we have, $\forall\ 0 < \mu \leq \mu_0$ and all $0 \leq s \leq t$, $\|\Phi(t,s)\|_a \leq K'_{a,\beta,\mu_0}(F)(1-\beta\mu)^{t-s}$, which implies

$$\|H_{t+1}^{(n+1)}\|_p \leq K'_{a,\beta,\mu_0}(F)\beta^{-1}\sup_{s\geq 0}\|Z_s\|_b\sup_{s\geq 0}\|J_s^{(n+1)}\|_c.$$

By construction, $H_s^{(n)}$ may be decomposed as $H_s^{(n)} = J_s^{(n+1)} + H_s^{(n+1)}$; the Minkowski inequality implies

$$\|H_s^{(n)}\|_p \leq \|J_s^{(n+1)}\|_p + \|H_s^{(n+1}\|_p \leq \left(1 + K'_{a,\beta,\mu_0}(F)\beta^{-1}\sup_{s\geq 0}\|Z_s\|_b\right)\sup_{s\geq 0}\|J_s^{(n+1)}\|_c,$$

which concludes the proof. □

REFERENCES

[1] S. Bittanti and M. Campi, *Adaptive RLS algorithms under stochastic excitation—L^2 convergence analysis*, IEEE Trans. Automat. Control, 36 (1991), pp. 963–967.

[2] S. Bittanti and M. Campi, *Mean square convergence of adaptive RLS algorithm with stochastic excitation*, in Advances in Adaptive Control, K. S. Narendra, R. Ortega, and P. Dorato, eds., IEEE Press, New York, 1991.

[3] R. R. Bitmead, B. Anderson, and T. Ng, *Convergence rate determination for gradient based adaptive estimators*, Automatica J. IFAC, 25 (1980), pp. 185–191.

[4] A. Davydov, *Mixing conditions for Markov chains*, Theory Probab. Appl., 18 (1973), pp. 312–328.

[5] P. E. Caines, *Linear Stochastic Systems*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1988.

[6] C. Deo, *A note on empirical processes of strong mixing sequences*, Ann. Probab., 1 (1973), pp. 870–875.

[7] P. Doukhan and S. Louhichi, *Weak dependence and moment inequalities*, Preprint Université de Paris-Sud, 1997.

[8] E. Eweda and O. Macchi, *Tracking error bounds of adaptive nonstationary filtering*, Automatica J. IFAC, 2 (1985), pp. 293–302.

[9] S. Gunnarson and L. Ljung, *Frequency domain tracking characteristics of adaptive algorithms*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 1072–1089.

[10] L. Guo, L. Xia, and J. B. Moore, *Tracking randomly varying parameters*: *Analysis of a standard algorithm*, Math. Control Signals Systems, 1991, pp. 1–16.

[11] L. Guo, L. Ljung, and P. Priouret, *Performance analysis of the forgetting factor RLS algorithm*, Internat. J. Adapt. Control Signal Process., 7, (1993), pp. 225–237.

[12] L. Guo, *Stability of recursive stochastic tracking algorithms*, SIAM J. Control Optim., 32 (1994), pp. 1195–1225.

[13] L. Guo and L. Ljung, *Exponential stability of general tracking algorithms*, IEEE Trans. Automat. Control, 40 (1995), pp. 1376–1387.

[14] L. Guo and L. Ljung, *Performance stability of general tracking algorithms*, IEEE Trans. Automat. Control, 40 (1995), pp. 1388–1398.

[15] L. Guo, L. Ljung, and G.-J. Wang, *Necessary and sufficient conditions for stability of LMS*, IEEE Trans. Automat. Control, 42 (1997), pp. 761–769.

[16] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Applications*, Academic Press, New York, 1980.

[17] M. Kouritzin, *Inductive methods and rates of r-th mean convergence in adaptive filtering*, Stochastics Stochastics Rep., 51 (1994), pp. 241–266.

[18] H. J. Kushner and G. Yin, *Stochastic approximation algorithms and applications*, in Applications of Mathematics, Stochastic Modelling and Applied Probability, Springer-Verlag, New York, 1997.

[19] L. Ljung and S. Gunnarson, *Adaptation and tracking in system identification: A survey*, Automatica J. IFAC, 26 (1990), pp. 7–21.

[20] L. Ljung and P. Priouret, *A result on the mean square error obtained using general tracking algorithms*, Internat. J. Adapt. Control Signal Process., 5 (1991), pp. 231–250.

[21] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.

[22] O. Macchi, *Optimisation of adaptive identification for time-varying filters*, IEEE Trans. Automat. Control, 31 (1986), pp. 283–287.

[23] O. Macchi, *LMS Adaptive Processing with Applications in Transmission*, Wiley, New York, 1994.

[24] A. Mazroui, *Poursuite et controle pour certains modèles de régressions*, Thèse Université Pierre et Marie Curie, Paris, France, 1993.

[25] A. Mokkadem, *Mixing properties of ARMA processes*, Theory Probab. Appl., 29 (1988), pp. 309–315.

[26] A. Perrier, B. Delyon, and E. Moulines, *Performance analysis of stochastic gradient identification algorithms using a perturbation expansion approach*, Technical report, Ecole Nationale Superieure de Télécommunications, Paris, France, 1998.

[27] V. Petrov, *Limit Theorems of Probability Theory*, Oxford University Press, Cambridge, 1995.

[28] P. Priouret and A. Veretennikov, *A remark on the stability of the LMS tracking algorithm*, Stochastic Anal. Appl., 16 (1998), pp. 119–129.

[29] V. Solo, *The limiting behavior of LMS*, IEEE Trans. Acoust. Speech and Signal Process., 37 (1989) pp. 1909–1922.

[30] V. Solo, *The error variance of LMS with time-varying weights*, IEEE Trans. Signal Process., 40 (1992), pp. 803–813.

[31] V. Solo, *Averaging analysis of the LMS algorithm*, Control and Dynamic Systems, Advances in Theory Appl., 65 (1994), pp. 379–390.

[32] V. Solo, *Deterministic adaptive control with slowly varying parameters: An averaging analysis*, Internat. J. Control, 64 (1996), pp. 99–125.

[33] V. Solo, *The Stability of LMS*, IEEE Trans. Signal Process., 45 (1997), pp. 3017–3026.

[34] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*: *Stability and Performance*, Prentice Hall, Upper Saddle River, NJ, 1995.

[35] R. Yokoyama, *Moment bounds for Stationary Mixing Sequences*, Z. Wahrsch. Verw. Gbiete, 52 (1980), pp. 45–57.

[36] B. Widrow and M. Hopf, *Adaptive switching circuits*, IRE Wescon Convention Record Part IV, 1960, pp. 96–104.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 900–916

BROCKETT'S PROBLEM OF CLASSIFICATION OF FINITE-DIMENSIONAL ESTIMATION ALGEBRAS FOR NONLINEAR FILTERING SYSTEMS[*]

SHANJIAN TANG[†]

Abstract. In his talk at the International Congress of Mathematicians in 1982, Brockett proposed to classify all finite-dimensional estimation algebras arising in nonlinear filters. Recently, Chen, Yau, and Leung [*SIAM J. Control Optim.*, 35 (1997), pp. 1132–1141] claimed that they had classified all finite-dimensional estimation algebras of maximal rank when the dimension of the state-space is less than or equal to four. In this paper, we introduce a series of completely new computations about the estimation algebra, and we find two sets of new equations about the Ω-matrix. As a consequence, we can prove, without any dimension assumption on the state-space, that the Ω-matrix is a constant matrix, and thus we succeed in classifying all finite-dimensional estimation algebras of maximal rank with the state-space dimension being arbitrary.

Key words. finite-dimensional filter, estimation algebra of maximal rank, nonlinear drift, arbitrary state-space dimension

AMS subject classifications. 17B30, 35J15, 60G35, 93E11

PII. S036301299833464X

1. Introduction.

Brockett and Clark [3], Brockett [2], and Mitter [14] initially began to use estimation algebras to construct finite-dimensional nonlinear filters. The study of estimation algebras then became important. Ocone [15] observed the following property of finite-dimensional estimation algebras: a function in a finite-dimensional estimation algebra must be a polynomial of degree less than or equal to two. This observation turns out to be a fundamental tool in later studies of finite-dimensional estimation algebras.

In his talk at the International Congress of Mathematicians in 1982, Brockett [1] proposed to classify all finite-dimensional estimation algebras. Since then, a number of progresses have been made on Brockett's classification problem. Tam, Wong, and Yau [17] classified all finite-dimensional exact estimation algebras of maximal rank with arbitrary state-space dimension. Chiou and Yau [8] introduced the concept of maximal rank estimation algebras and classified all finite-dimensional estimation algebras of maximal rank with the state-space dimension less than or equal to two. Later, Chen, Yau, and Leung [4, 7] improved Chiou and Yau's result in that the dimension of the state-space is assumed to be less than or equal to three and four, respectively. But their proofs strongly depend on the dimension assumptions that $n \leq 3$ and that $n \leq 4$, and they are difficult to be generalized to the case of arbitrary dimension n.

In this paper, we classify all finite-dimensional estimation algebras of maximal rank with arbitrary state-space dimension. Our result is stated as follows.

[*] Received by the editors February 27, 1998; accepted for publication (in revised form) June 7, 2000; published electronically October 11, 2000. This research was supported by the National Natural Science Foundation of China (project 79790130). The result of this paper has been presented as a short communication by Tang at the International Congress of Mathematicians of 1998 in Berlin.
http://www.siam.org/journals/sicon/39-3/33464.html

[†] Department of Mathematics and the Laboratory of Mathematics for Nonlinear Sciences at Fudan University, Fudan University, Shanghai 200433, China. Current address: Department of Mathematics and Statistics, University of Konstanz, D-78457, Konstanz, Germany (tang@fmi.uni-konstanz.de).

$$\begin{tikzcd}
& U_j \arrow[d, equal] & V_j \arrow[d, equal] & W_j \arrow[d, equal] \\
D_j \arrow[r, "L_0"] & {[L_0, D_j]} \arrow[r, "L_0"] \arrow[d, "D_l"] & {[L_0, U_j]} \arrow[r, "L_0"] \arrow[d, "D_j"] & {[L_0, V_j]} \arrow[d, "D_j"] \\
& {[U_j, D_l]} & {[V_j, D_j]} \arrow[d, "D_l"] & {[W_j, D_j]} \arrow[d, "D_j"] \\
& & {[[V_j, D_j], D_l]} & {[[W_j, D_j], D_j]} \arrow[d, "D_j"] \\
& & & {[[[W_j, D_j], D_j], D_j]}
\end{tikzcd}$$

FIG. 1.1. *The computation chart.*

THEOREM 1.1. *Let $\mathcal{E}$ be a finite-dimensional estimation algebra of* (2.1) *of maximal rank, and let $\mathcal{E}_0$ be the real vector space of dimension $2n+2$ with basis given by $1, x_1, \dots, x_n, D_1, \dots, D_n$ and L_0. Then,*

1. *the drift term f must be a linear vector field (i.e., each component is a polynomial of degree less than or equal to one) plus a gradient vector field;*
2. *$\mathcal{E} = \mathcal{E}_0$;*
3. *η is a polynomial of degree less than or equal to two.*

Theorem 1.1 improves both results of Tam, Wong, and Yau [17] and Chen, Yau, and Leung [7] in that it neither assumes that the finite-dimensional estimation algebra under consideration is exact nor assumes that the state-space dimension ≤ 4.

The difficulty of proving Theorem 1.1 is to prove that ω_{ij} is constant for $k+1 \leq i \leq n$ and $k+1 \leq j \leq n$ (where k denotes the quadratic rank of the finite-dimensional estimation algebra under consideration), which is stated as Lemma 3.5. This difficulty can be known from Chen and Yau [5, 6], Yau [22], and Yau's recent work coauthored with Guoqing Hu.

In this paper, in order to overcome the above difficulty, we first establish Lemma 3.4, which plays a key role in the proof of Lemma 3.5. After Lemma 3.5 has been proved, the rest of the proof of Theorem 1.1 is straightforward.

In order to establish Lemma 3.4, we develop a series of computations about the estimation algebra $\mathcal{E}$, which is inspired by Ocone's series of computations in his paper [15]. Figure 1.1 is our computation chart and is helpful for the reader to trace our proof.

As shown in Fig. 1.1, we construct the two elements $[[V_j, D_j], D_l]$ and $[[[W_j, D_j], D_j], D_j]$ in $\mathcal{E}$. These two elements can be computed by using the last six formulas (10)–(15) of Lemma 2.2, and they turn out to have the same form:

"a polynomial of degree two" + "an element of $\mathcal{E}_0$."

In this way, we obtain two polynomials in $\mathcal{E}$ of degree two, and then by analyzing the coefficients of x_j^2 $(j > k)$ in these two polynomials and by applying Theorem 2.3, we establish two sets of new equations about the ω_{ij}, which are formulated as Lemmas 3.2 and 3.3, respectively. Lemmas 3.2 and 3.3 together immediately imply the elegant equation

$$A_j^4(j, j) = 0, \quad j = k+1, \dots, n, \tag{1.1}$$

which leads to Lemma 3.4.

In the proof of Lemma 3.4, it is important to think of proving (1.1), while in our arguments, the term $A_j^4(j,j)$ appears automatically when we compute the coefficients of x_j^2 $(j > k)$ in $[[[W_j, D_j], D_j], D_j]$.

There are many other striking works which are related to this paper. Among these are Cohen De Lara [9, 10], Davis and Marcus [11], Dong, Tam, Wong, and Yau [12], Duncan [13], Marcus [16], Wong [19, 20, 21], and Yau [22]. The recent perspective paper [13] by Professor Tyrone E. Duncan provides a good account of the past and the present of the filtering theory.

The rest of this paper is organized as follows. Section 2 contains some basic concepts and previous results from Ocone [15], Chiou and Yau [8], Yau [22], and Chen and Yau [5]. Section 3 contains the proof of Theorem 1.1.

2. Preliminaries.

Consider the signal observation model:

$$\begin{cases} dx(t) = f(x(t))dt + g(x(t))dv(t), & x(0) = x_0, \\ dy(t) = h(x(t))dt + dw(t), & y(0) = 0, \end{cases} \tag{2.1}$$

in which x, v, y, and w are, respectively, $\mathbb{R}^n$-, $\mathbb{R}^n$-, $\mathbb{R}^m$-, and $\mathbb{R}^m$-valued processes and v and w are independent, standard Brownian processes. Suppose that the vector functions f and h are C^∞ smooth and that $g(x)$ is an orthogonal matrix for each $x \in \mathbb{R}^n$. We shall refer to $x(t)$ as the state of the system at time t and $y(t)$ as the observation at time t. $\rho(t,x)$, the conditional probability density of the state, $x(t)$, given the observation $\{y(s) : 0 \le s \le t\}$ is determined by the Duncan–Mortensen–Zakai equation, which in the unnormalized form is given by

$$d\rho(t,x) = L_0\rho(t,x)dt + \sum_{i=1}^m L_i\rho(t,x)dy_i(t), \quad \rho(0,x) = \rho_0(x), \tag{2.2}$$

where

$$L_0 = \frac{1}{2}\sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^n f_i\frac{\partial}{\partial x_i} - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} - \frac{1}{2}\sum_{i=1}^m h_i^2 \tag{2.3}$$

and for $i = 1, \dots, m, L_i$ is the zero-degree differential operator of multiplication by h_i. (If a is a vector, we use the notation a_i to denote the ith component of a.) ρ_0 is the probability density of the initial point x_0.

Define

$$D_i = \frac{\partial}{\partial x_i} - f_i \tag{2.4}$$

and

$$\eta = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^n f_i^2 + \sum_{i=1}^m h_i^2. \tag{2.5}$$

Then

$$L_0 = \frac{1}{2}\left(\sum_{i=1}^n D_i^2 - \eta\right). \tag{2.6}$$

Throughout this paper, we denote by $\mathcal{P}_i$ the totality of all the polynomials of degree less than or equal to i. Whenever necessary, we use $\mathcal{P}_i(x_1, \dots, x_r)$ to specify the dependent variables of $\mathcal{P}_i$. The default dependent variables are $x_1, \dots, x_n$.

DEFINITION 2.1. *The estimation algebra $\mathcal{E}$ of the filtering system* (2.1) *is defined to be the Lie algebra generated by $\{L_0, L_1, \dots, L_m\}$.*

Ocone [15] observed the following basic property for an estimation algebra $\mathcal{E}$ to be finite-dimensional.

THEOREM 2.1. *Let $\mathcal{E}$ be a finite-dimensional estimation algebra. If φ is a function in $\mathcal{E}$, then φ is a polynomial of degree less than or equal to two, i.e., $\varphi \in \mathcal{P}_2$.*

DEFINITION 2.2. *An estimation algebra $\mathcal{E}$ is said to be of maximal rank if for every $1 \le i \le n$ there exists a constant c_i such that $x_i + c_i \in \mathcal{E}$.*

Let $\mathcal{E}_0$ be the real vector space spanned by $1, x_1, \dots, x_n, D_1, \dots, D_n$ and L_0. The following lemma is valid for an estimation algebra of maximal rank with arbitrary dimension, and it can be found in Chen, Yau, and Leung [7].

LEMMA 2.1. *Let $\mathcal{E}$ be an estimation algebra of maximal rank associated with the filtering system* (2.1)*. Then $\mathcal{E} \supset \mathcal{E}_0$.*

If $\mathcal{E}$ is an estimation algebra of maximal rank, we have by Lemma 2.1 that

$$\mathcal{E} \ni [D_j, D_i] = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} =: \omega_{ij}, \quad 1 \le i,\ j \le n.$$

If, further, $\mathcal{E}$ is finite-dimensional, then $\omega_{ij} \in \mathcal{P}_2$ for $1 \le i,\ j \le n$ by Theorem 2.1.

Note that f_i as a tensor field is the differential 1-form $\sum_{i=1}^n f_i \, dx_i$ and ω_{ij} as a tensor field is the differential 2-form $\sum_{1 \le i < j \le n} \omega_{ij} \, dx_i \wedge dx_j$. The exterior derivative of the former is just the latter, i.e., they have the following relation:

$$d\left(\sum_{i=1}^n f_i \, dx_i\right) = \sum_{1 \le i < j \le n} \omega_{ij} \, dx_i \wedge dx_j.$$

On one hand, since $d^2 = 0$, we deduce that ω_{ij}'s satisfy the cyclic relation

$$\frac{\partial \omega_{ij}}{\partial x_l} + \frac{\partial \omega_{li}}{\partial x_j} + \frac{\partial \omega_{jl}}{\partial x_i} = 0, \quad 1 \le i,\ j,\ l \le n. \tag{2.7}$$

On the other hand, the Poincaré lemma means that every d-closed differential form in $\mathbb{R}^n$ is d-exact. Then, $\omega_{ij} \equiv 0$ for $1 \le i,\ j \le n$ means that f_i is d-closed and thus is d-exact, i.e., f_i is a gradient vector field.

Theorem 2.2 summarizes some results of Yau [22].

THEOREM 2.2. *Let $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$ be constant functions. Then Theorem* 1.1 *holds.*

The proof of Theorem 1.1 is then reduced to showing that $\omega_{ij}, 1 \le i,\ j \le n$ are all constants. In the following, we recall the existing results about the constant structure of the Ω-matrix.

Denote by $\mathcal{H}_2$ the space of quadratic forms in n variables, i.e., the real vector space spanned by $x_i x_j$, with $1 \le i \le j \le n$. Let $X = (x_1, \dots, x_n)'$.

DEFINITION 2.3. *For any quadratic form $p \in \mathcal{H}_2$, there exists a symmetric $n \times n$ matrix A such that $p(x) = X'AX$. The rank of the quadratic form p denoted by* $\mathrm{r}(p)$ *is defined to be the rank of the matrix A.*

DEFINITION 2.4. *A fundamental quadratic form of the estimation algebra $\mathcal{E}$ is an element $p_0 \in \mathcal{E} \cap \mathcal{H}_2$ with the biggest positive rank, i.e.,* $\mathrm{r}(p_0) \ge \mathrm{r}(p)$ *for any*

$p \in \mathcal{E} \cap \mathcal{H}_2$. The quadratic rank of the estimation algebra $\mathcal{E}$ is defined to be the rank of a fundamental quadratic form of $\mathcal{E}$.

Let p_0 be a fundamental quadratic form of $\mathcal{E}$ and $k := r(p_0)$. After an orthogonal transformation on x, p_0 can be written as

$$p_0 = \sum_{i=1}^{k} c_i x_i^2, \quad c_i \neq 0.$$

From p_0, we can construct a sequence of quadratic forms in $\mathcal{E} \cap \mathcal{H}_2$ as follows:

$$\begin{aligned} q_0 :=& p_0, \\ q_j :=& [[L_0, q_{j-1}], q_0] = \sum_{i=1}^{k} 4^j c_i^{j+1} x_i^2, \quad j = 1, 2, \ldots . \end{aligned}$$

In view of the invertibility of the Vandermonde matrix, we can assume that

$$p_0 = \sum_{i=1}^{k} x_i^2. \tag{2.8}$$

Chen and Yau [5, 6] (see also a recent work by Yau and Hu) introduced the above concepts and proved Theorems 2.3 and 2.4.

THEOREM 2.3. *Let $\mathcal{E}$ be a finite-dimensional estimation algebra of maximal rank. Let k be the quadratic rank of $\mathcal{E}$, and let p_0 (defined by (2.8)) be a fundamental quadratic form of $\mathcal{E}$. Then $p \in \mathcal{E} \cap \mathcal{P}_2$ implies that the coefficients of $x_i x_j$ in p are equal to zero for $i = k+1, \ldots, n$ and $j = 1, \ldots, n$.*

THEOREM 2.4. *Let $\mathcal{E}$ be a finite-dimensional estimation algebra of maximal rank. Let k be the quadratic rank of $\mathcal{E}$. Then*

(1) *the observation terms $h_i \in \mathcal{P}_1$ for $1 \leq i \leq m$;*

(2) *ω_{ij} are constants for $1 \leq i \leq k$ or $1 \leq j \leq k$; and $\omega_{ij} \in \mathcal{P}_1(x_{k+1}, \ldots, x_n)$ for $k+1 \leq i,\ j \leq n$.*

For the sake of convenience, we also provide the following lemma without proof, which can be checked out directly.

LEMMA 2.2. *Assume that g and h are C^∞ functions defined on $\mathbb{R}^n$. Then, we have the following.*

(1) $[XY, Z] = X[Y, Z] + [X, Z]Y$, *where X, Y, and Z are differential operators.*

(2) $[gD_i, h] = g\frac{\partial h}{\partial x_i}$.

(3) $[gD_i, hD_j] = gh\omega_{ji} + g\frac{\partial h}{\partial x_i}D_j - h\frac{\partial g}{\partial x_j}D_i$.

(4) $[gD_i^2, h] = 2g\frac{\partial h}{\partial x_i}D_i + g\frac{\partial^2 h}{\partial x_i^2}$.

(5) $[D_i^2, hD_j] = 2\frac{\partial h}{\partial x_i}D_iD_j + 2h\omega_{ji}D_i + \frac{\partial^2 h}{\partial x_i^2}D_j + h\frac{\partial \omega_{ji}}{\partial x_i}$.

(6) $[D_i^2, D_j^2] = 4\omega_{ji}D_jD_i + 2\frac{\partial \omega_{ji}}{\partial x_j}D_i + 2\frac{\partial \omega_{ji}}{\partial x_i}D_j + \frac{\partial^2 \omega_{ji}}{\partial x_i \partial x_j} + 2\omega_{ji}^2$.

(7) $[D_r^2, hD_iD_j] = 2\frac{\partial h}{\partial x_r}D_rD_iD_j + 2h\omega_{jr}D_iD_r + 2h\omega_{ir}D_rD_j + \frac{\partial^2 h}{\partial x_r^2}D_iD_j + 2h\frac{\partial \omega_{jr}}{\partial x_i}D_r$
$+ h\frac{\partial \omega_{jr}}{\partial x_r}D_i + h\frac{\partial \omega_{ir}}{\partial x_r}D_j + h\frac{\partial^2 \omega_{jr}}{\partial x_i \partial x_r}$.

(8) $[gD_iD_j, hD_r] = g\frac{\partial h}{\partial x_j}D_iD_r + g\frac{\partial h}{\partial x_i}D_jD_r + gh\omega_{rj}D_i + gh\omega_{ri}D_j + g\frac{\partial^2 h}{\partial x_i \partial x_j}D_r$
$+ gh\frac{\partial \omega_{rj}}{\partial x_i} - h\frac{\partial g}{\partial x_r}D_iD_j$.

(9) $[D_iD_j, h] = \frac{\partial h}{\partial x_i}D_j + \frac{\partial h}{\partial x_j}D_i + \frac{\partial^2 h}{\partial x_i \partial x_j}$.

(10) $[[g, D_j], D_l] = \frac{\partial^2 g}{\partial x_l \partial x_j}$.

(11) $[[[g, D_i], D_j], D_l] = -\frac{\partial^3 g}{\partial x_l \partial x_j \partial x_i}$.

(12) $[[gD_i, D_j], D_l] = \frac{\partial^2 g}{\partial x_l \partial x_j} D_i - \frac{\partial (g\omega_{ji})}{\partial x_l} - \frac{\partial g}{\partial x_j}\omega_{li}$.

(13) $[[[gD_i, D_j], D_j], D_j] = -\frac{\partial^3 g}{\partial x_j^3} D_i + 3\frac{\partial^2 g}{\partial x_j^2}\omega_{ji} + 3\frac{\partial g}{\partial x_j}\frac{\partial \omega_{ji}}{\partial x_j} + g\frac{\partial^2 \omega_{ji}}{\partial x_j^2}$.

(14) $[[D_i D_r, D_j], D_l] = -\frac{\partial \omega_{jr}}{\partial x_l} D_i - \frac{\partial \omega_{ji}}{\partial x_l} D_r + \omega_{ji}\omega_{lr} + \omega_{jr}\omega_{li} - \frac{\partial^2 \omega_{jr}}{\partial x_l \partial x_i}$.

(15) *Assume that* $g \in \mathcal{P}_1$ *and* $\omega_{ij} \in \mathcal{P}_1$. *Then, we have*

$$\begin{aligned}[[[gD_iD_r, D_j], D_j], D_j] =& 3\frac{\partial g}{\partial x_j}\left(\frac{\partial \omega_{jr}}{\partial x_j} D_i + \frac{\partial \omega_{ji}}{\partial x_j} D_r\right) \\ & - 3\left(\frac{\partial g}{\partial x_j}\omega_{jr}\omega_{ji} + \frac{\partial (g\omega_{jr}\omega_{ji})}{\partial x_j}\right).\end{aligned}$$

3. Proof of Theorem 1.1. By Theorem 2.4, we have $\omega_{ij} \in \mathcal{P}_1 \ \forall\, i, j = 1, \dots, n$. In the following, denote by $A_r(i,j)$ the coefficient of x_r in ω_{ij}, and denote by A_r the matrix whose (i,j)-component is $A_r(i,j)$. That is,

$$A_r(i,j) = \frac{\partial \omega_{ij}}{\partial x_r}, \quad A_r := (A_r(i,j))_{1\le i,j\le n}. \tag{3.1}$$

Note that A_r is a *constant skew-symmetric* matrix.

For convenience of notation, we also write

$$C(j,l) := \sum_{r=1}^{n} \omega_{jr}\omega_{lr} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_l \partial x_j} \qquad (\in \mathcal{P}_2;\ \text{see (3.7) below}). \tag{3.2}$$

LEMMA 3.1. *For* $i = k+1, \dots, n$, *we have*

$$\begin{aligned}\sum_{r=1}^{n} A_i(j,r)A_i(r,j) &= \sum_{r=1}^{n} A_j(i,r)A_j(r,i) \\ =& \frac{1}{2}\sum_{r=1}^{n}[A_i(i,r)A_j(r,j) + A_j(i,r)A_i(r,j)], \quad j = k+1, \dots, n\end{aligned} \tag{3.3}$$

and

$$\sum_{r=1}^{n} A_i(i,r)A_i(r,j) = \sum_{r=1}^{n} A_i(i,r)A_j(r,i), \quad j = 1, \dots, n. \tag{3.4}$$

Lemma 3.1 has been derived by Chen, Yau, and Leung [7, pp. 1137–1138], and the following proof refines some of their arguments.

Proof of Lemma 3.1. Introduce U_j and compute $[[L_0, D_j], D_l]$. By (2), (3), and (5) of Lemma 2.2, we have

$$\begin{aligned}U_j :=& [L_0, D_j] = \frac{1}{2}\sum_{r=1}^{n}[D_r^2, D_j] - \frac{1}{2}[\eta, D_j] \\ &= \sum_{r=1}^{n}\left(\omega_{jr}D_r + \frac{1}{2}\frac{\partial \omega_{jr}}{\partial x_r}\right) + \frac{1}{2}\frac{\partial \eta}{\partial x_j} \in \mathcal{E}\end{aligned} \tag{3.5}$$

and

$$
\begin{aligned}
[U_j, D_l] &= \sum_{r=1}^{n} [\omega_{jr} D_r, D_l] - \frac{1}{2}\left[D_l, \frac{\partial \eta}{\partial x_j}\right] \\
&= \sum_{r=1}^{n} \left(\omega_{jr}\omega_{lr} - \frac{\partial \omega_{jr}}{\partial x_l} D_r\right) - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_l \partial x_j} \\
&= C(j,l) - \sum_{r=1}^{n} A_l(j,r) D_r \in \mathcal{E}.
\end{aligned}
\tag{3.6}
$$

Since the last sum of (3.6) is an element of $\mathcal{E}$ by Lemma 2.1 and by the fact that A_r is a constant matrix, we deduce (noting Theorem 2.1) that

$$C(j,l) \in \mathcal{E} \cap \mathcal{P}_2, \quad 1 \le l,\ j \le n. \tag{3.7}$$

Then by (3.2), we conclude that $\eta \in \mathcal{P}_4$ and that $\forall i > k$ or $\forall q > k$,

$$C_{iq}(j,l) := \frac{\partial^2 (C(j,l))}{\partial x_i \partial x_q} = 0 \qquad \text{(by Theorem 2.3)}, \tag{3.8}$$

which implies (by (3.2)) that

$$\frac{1}{2}\frac{\partial^4 \eta}{\partial x_i \partial x_q \partial x_l \partial x_j} = \sum_{r=1}^{n} (A_i(j,r)A_q(l,r) + A_q(j,r)A_i(l,r)). \tag{3.9}$$

Note that (i,q,l,j) is permutable in the left-hand side of (3.9), and this property immediately gives (3.3) and (3.4), using the fact that A_r is skew-symmetric. □

In the following, please keep in mind the following three facts: (1) $\omega_{ij} \in \mathcal{P}_1$, (2) $\eta \in \mathcal{P}_4$, and (3) A_r is skew-symmetric.

Next we analyze the two elements $[[V_j, D_j], D_l]$ and $[[[W_j, D_j], D_j], D_j]$ in $\mathcal{E}$. As a consequence, we establish two sets of new equations among ω_{ij}. They are formulated as Lemmas 3.2 and 3.3, respectively, below.

LEMMA 3.2. *For $j = k+1, \dots, n$ and $l = 1, \dots, n$, we have*

$$\sum_{i,r=1}^{n} A_j(j,r) A_r(j,i) A_j(i,l) = 0. \tag{3.10}$$

Proof of Lemma 3.2. Recall that U_j is given by (3.5), and note that it can be rewritten as

$$U_j = \sum_{i=1}^{n} \left(\omega_{ji} D_i + \frac{1}{2}\frac{\partial \omega_{ji}}{\partial x_i}\right) + \frac{1}{2}\frac{\partial \eta}{\partial x_j} \in \mathcal{E}.$$

Introduce V_j and compute $[[V_j, D_j], D_l]$.

(3.11)
$$\begin{aligned}
V_j :=&[L_0, U_j]\\
=&\frac{1}{2}\sum_{i,r=1}^{n}[D_r^2, \omega_{ji}D_i] + \frac{1}{4}\sum_{r=1}^{n}\left[D_r^2, \frac{\partial\eta}{\partial x_j}\right] + \frac{1}{2}\sum_{i=1}^{n}[\omega_{ji}D_i, \eta]\\
=&\sum_{i,r=1}^{n}\frac{\partial\omega_{ji}}{\partial x_r}D_rD_i + \sum_{i,r=1}^{n}\omega_{ji}\omega_{ir}D_r + \frac{1}{2}\sum_{i,r=1}^{n}\omega_{ji}\frac{\partial\omega_{ir}}{\partial x_r} \qquad \text{by (5) of Lemma 2.2}\\
&+\frac{1}{4}\sum_{r=1}^{n}\left(2\frac{\partial^2\eta}{\partial x_r\partial x_j}D_r + \frac{\partial^3\eta}{\partial x_r^2\partial x_j}\right) + \frac{1}{2}\sum_{i=1}^{n}\omega_{ji}\frac{\partial\eta}{\partial x_i} \qquad \text{by (4) and (2)}\\
=&\sum_{i,r=1}^{n}\frac{\partial\omega_{ji}}{\partial x_r}D_rD_i - \sum_{r=1}^{n}C(j,r)D_r + p_4 + p_1 \in \mathcal{E} \qquad \text{by (3.2),}
\end{aligned}$$

where the notation p_4 and p_1 are defined, respectively, as

$$\begin{aligned}
p_4 :=&\frac{1}{2}\sum_{i=1}^{n}\omega_{ji}\frac{\partial\eta}{\partial x_i} \in \mathcal{P}_4,\\
p_1 :=&\frac{1}{2}\sum_{i,r=1}^{n}\omega_{ji}\frac{\partial\omega_{ir}}{\partial x_r} + \frac{1}{4}\sum_{r=1}^{n}\frac{\partial^3\eta}{\partial x_r^2\partial x_j} \in \mathcal{P}_1.
\end{aligned}$$

We use (10), (12), and (14) of Lemma 2.2 to compute $[[V_j, D_j], D_l]$ as follows. Since $[[p_1, D_j], D_l] = 0$, we have

$$\begin{aligned}
&[[V_j, D_j], D_l]\\
&= \sum_{i,r=1}^{n}\frac{\partial\omega_{ji}}{\partial x_r}[[D_rD_i, D_j], D_l] - \sum_{r=1}^{n}[[C(j,r)D_r, D_j], D_l] + [[p_4, D_j], D_l]\\
&= \sum_{i,r=1}^{n}\frac{\partial\omega_{ji}}{\partial x_r}\left(-\frac{\partial\omega_{jr}}{\partial x_l}D_i - \frac{\partial\omega_{ji}}{\partial x_l}D_r + \omega_{ji}\omega_{lr} + \omega_{jr}\omega_{li}\right) \qquad \text{by (14)}\\
&-\sum_{r=1}^{n}\left(\frac{\partial^2C(j,r)}{\partial x_j\partial x_l}D_r - \frac{\partial(\omega_{jr}C(j,r))}{\partial x_l} - \frac{\partial C(j,r)}{\partial x_j}\omega_{lr}\right) \qquad \text{by (12)}\\
&+\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial\omega_{ji}}{\partial x_j}\frac{\partial^2\eta}{\partial x_l\partial x_i} + \frac{\partial\omega_{ji}}{\partial x_l}\frac{\partial^2\eta}{\partial x_j\partial x_i} + \omega_{ji}\frac{\partial^3\eta}{\partial x_j\partial x_i\partial x_l}\right) \qquad \text{by (10).}
\end{aligned} \tag{3.12}$$

So, $[[V_j, D_j], D_l]$ is a first-order differential operator in $\mathcal{E}$ with the coefficients of $D_1, \dots, D_n$ being constants. In view of Lemma 2.1 and Theorem 2.1, we derive from (3.12) that

$$\begin{aligned}
&\sum_{i,r=1}^{n}\frac{\partial\omega_{ji}}{\partial x_r}(\omega_{ji}\omega_{lr} + \omega_{jr}\omega_{li}) + \sum_{i=1}^{n}\left(\frac{\partial(\omega_{ji}C(j,i))}{\partial x_l} + \frac{\partial C(j,i)}{\partial x_j}\omega_{li}\right)\\
&\quad+\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial\omega_{ji}}{\partial x_j}\frac{\partial^2\eta}{\partial x_l\partial x_i} + \frac{\partial\omega_{ji}}{\partial x_l}\frac{\partial^2\eta}{\partial x_j\partial x_i} + \omega_{ji}\frac{\partial^3\eta}{\partial x_j\partial x_i\partial x_l}\right) \in \mathcal{E}\cap\mathcal{P}_2.
\end{aligned} \tag{3.13}$$

In view of the notation (3.1) and (3.2), we have

$$\begin{aligned}\sum_{i,r=1}^{n}\Bigg(&A_r(j,i)\omega_{jr}\omega_{li}+A_r(j,i)\omega_{ji}\omega_{lr}-A_j(j,i)\omega_{lr}\omega_{ri}\\ &-\frac{\partial(\omega_{ji}\omega_{jr}\omega_{ri})}{\partial x_l}\Bigg)+\sum_{i=1}^{n}\left(\frac{\partial(C(j,i))}{\partial x_j}\omega_{li}-A_j(j,i)C(l,i)\right)\in\mathcal{E}\cap\mathcal{P}_2.\end{aligned}\tag{3.14}$$

For $l=1,\ldots,n$, the coefficient of x_j^2 $(j\geq k+1)$ in (3.14) is

$$\begin{aligned}&\sum_{i,r=1}^{n}(A_r(j,i)A_j(j,r)A_j(l,i)+A_r(j,i)A_j(j,i)A_j(l,r)\\ &-A_j(j,i)A_j(l,r)A_j(r,i)-A_l(j,i)A_j(j,r)A_j(r,i)\\ &-A_j(j,i)A_l(j,r)A_j(r,i)-A_j(j,i)A_j(j,r)A_l(r,i))\\ &+\sum_{i=1}^{n}\left(C_{jj}(j,i)A_j(l,i)-\frac{1}{2}A_j(j,i)C_{jj}(l,i)\right)\\ =&\sum_{i,r=1}^{n}[-A_j(j,r)A_r(j,i)A_j(i,l)+A_j(j,i)A_j(i,r)A_j(r,l)\\ &-A_j(j,i)A_j(i,r)A_j(r,l)+A_j(j,r)A_j(r,i)A_l(i,j)\\ &-A_j(j,i)A_j(i,r)A_l(r,j)-A_j(j,i)A_l(i,r)A_j(r,j)]\\ &\left(\text{since }\sum_{i,r=1}^{n}A_r(j,i)A_j(j,i)=\sum_{i,r=1}^{n}A_j(r,i)A_j(j,i)\quad\text{by (3.4)},\right.\\ &\left.\text{and since }C_{jj}(j,i)=0=C_{jj}(l,i)\quad\text{by (3.8)}\right)\\ =&-\sum_{i,r=1}^{n}A_j(j,r)A_r(j,i)A_j(i,l)+A_j^3(j,l)-A_j^3(j,l)\\ &+(A_j^2A_l)(j,j)-(A_j^2A_l)(j,j)-(A_jA_lA_j)(j,j)\\ =&-\sum_{i,r=1}^{n}A_j(j,r)A_r(j,i)A_j(i,l)-(A_jA_lA_j)(j,j)\\ =&-\sum_{i,r=1}^{n}A_j(j,r)A_r(j,i)A_j(i,l)\\ &\text{(since }A_jA_lA_j\text{ is skew-symmetric).}\end{aligned}\tag{3.15}$$

This coefficient should be zero as $j\geq k+1$ by Theorem 2.3. So,

$$\begin{aligned}&\sum_{i,r=1}^{n}A_j(j,r)A_r(j,i)A_j(i,l)=0,\\ &j=k+1,\ldots,n,\quad l=1,\ldots,n.\qquad\square\end{aligned}\tag{3.16}$$

Lemma 3.3. *We have*

$$A_j^4(j,j) = \frac{1}{4}\sum_{i,l,r=1}^{n} A_j(j,r)A_r(j,i)A_j(i,l)A_j(l,j), \quad j = k+1, \dots, n. \tag{3.17}$$

Proof of Lemma 3.3. Introduce W_j and compute $[[[W_j, D_j], D_j], D_j]$. Recall that $\omega_{ji} \in \mathcal{P}_1$, so that all of the terms

$$\frac{\partial \omega_{ji}}{\partial x_l}, \quad i,\ j,\ l = 1, \dots, n,$$

are constants. Also recall that V_j is given by (3.11) and note that it can be rewritten as

$$V_j = \sum_{i,l=1}^{n} \frac{\partial \omega_{ji}}{\partial x_l} D_l D_i - \sum_{i=1}^{n} C(j,i) D_i + p_4 + p_1 \in \mathcal{E}.$$

Therefore, we have

$$\begin{aligned}
W_j :=& [L_0, V_j] \\
=& \frac{1}{2}\sum_{i,l,r=1}^{n} \frac{\partial \omega_{ji}}{\partial x_l}[D_r^2, D_l D_i] + \frac{1}{2}\sum_{i,l=1}^{n} \frac{\partial \omega_{ji}}{\partial x_l}[D_l D_i, \eta] \\
& - \frac{1}{2}\sum_{i,r=1}^{n}[D_r^2, C(j,i)D_i] - \frac{1}{2}\sum_{i=1}^{n}[C(j,i)D_i, \eta] + \frac{1}{2}\sum_{r=1}^{n}[D_r^2, p_4 + p_1] \\
=& \frac{1}{2}\sum_{i,l,r=1}^{n} \frac{\partial \omega_{ji}}{\partial x_l}\left(2\omega_{ir}D_l D_r + 2\omega_{lr}D_r D_i + 2\frac{\partial \omega_{ir}}{\partial x_l}D_r + \frac{\partial \omega_{ir}}{\partial x_r}D_l + \frac{\partial \omega_{lr}}{\partial x_r}D_i\right) \quad \text{by (7)} \\
& + \frac{1}{2}\sum_{i,l=1}^{n} \frac{\partial \omega_{ji}}{\partial x_l}\left(\frac{\partial \eta}{\partial x_i}D_l + \frac{\partial \eta}{\partial x_l}D_i + \frac{\partial^2 \eta}{\partial x_l \partial x_i}\right) \qquad \text{by (9)} \\
& - \sum_{i,r=1}^{n} \frac{\partial (C(j,i))}{\partial x_r}D_r D_i - \sum_{i,r=1}^{n} C(j,i)\omega_{ir}D_r - \frac{1}{2}\sum_{i,r=1}^{n} \frac{\partial^2 (C(j,i))}{\partial x_r^2}D_i \\
& - \frac{1}{2}\sum_{i,r=1}^{n} C(j,i)\frac{\partial \omega_{ir}}{\partial x_r} - \frac{1}{2}\sum_{i=1}^{n} C(j,i)\frac{\partial \eta}{\partial x_i} \qquad \text{by (5) and (2)} \\
& + \frac{1}{2}\sum_{r=1}^{n}\left(2\frac{\partial (p_4 + p_1)}{\partial x_r}D_r + \frac{\partial^2 p_4}{\partial x_r^2}\right) \qquad \text{by (4).}
\end{aligned}$$

Then W_j can be written in the following form:

$$W_j = \sum_{l,r=1}^{n} B_{lr}^j D_l D_r + \sum_{r=1}^{n}(b_r^j + \overline{b}_r^j)D_r + p_5 + p_2 \in \mathcal{E}, \tag{3.18}$$

where

$$
\begin{aligned}
B_{lr}^j :=& \sum_{i=1}^n \left(\frac{\partial \omega_{ji}}{\partial x_l} \omega_{ir} + \frac{\partial \omega_{jr}}{\partial x_i} \omega_{il} \right) - \frac{\partial (C(j,r))}{\partial x_l} \in \mathcal{P}_1, \\
b_r^j :=& \sum_{i=1}^n \left(\frac{\partial \omega_{ji}}{\partial x_r} \frac{\partial \eta}{\partial x_i} + \frac{1}{2} \omega_{ji} \frac{\partial^2 \eta}{\partial x_r \partial x_i} - C(j,i)\omega_{ir} + \frac{1}{2} \frac{\partial \omega_{jr}}{\partial x_i} \frac{\partial \eta}{\partial x_i} \right) \in \mathcal{P}_3, \\
\overline{b}_r^j :=& \sum_{i,l=1}^n \left(\frac{\partial \omega_{ji}}{\partial x_l} \frac{\partial \omega_{ir}}{\partial x_l} + \frac{1}{2} \frac{\partial \omega_{ji}}{\partial x_r} \frac{\partial \omega_{il}}{\partial x_l} + \frac{1}{2} \frac{\partial \omega_{jr}}{\partial x_l} \frac{\partial \omega_{li}}{\partial x_i} \right) \\
& - \frac{1}{2} \sum_{l=1}^n \frac{\partial^2 (C(j,r))}{\partial x_l^2} + \frac{\partial p_1}{\partial x_r} \equiv \text{a constant}, \\
p_5 :=& -\frac{1}{2} \sum_{i=1}^n C(j,i) \frac{\partial \eta}{\partial x_i} \in \mathcal{P}_5, \\
p_2 :=& \frac{1}{2} \sum_{r=1}^n \frac{\partial^2 p_4}{\partial x_r^2} + \frac{1}{2} \sum_{i,r=1}^n \frac{\partial \omega_{ji}}{\partial x_r} \frac{\partial^2 \eta}{\partial x_r \partial x_i} - \frac{1}{2} \sum_{i,r=1}^n C(j,i) \frac{\partial \omega_{ir}}{\partial x_r} \in \mathcal{P}_2.
\end{aligned}
\tag{3.19}
$$

Next we use the formulas (11), (13), and (15) in Lemma 2.2 to compute $[[[W_j, D_j], D_j], D_j]$ as follows. Noting that

$$
\begin{aligned}
[[[D_r, D_j], D_j], D_j] &= \frac{\partial^2 \omega_{jr}}{\partial x_j^2} = 0, \\
[[[p_2, D_j], D_j], D_j] &= -\frac{\partial^3 p_2}{\partial x_j^3} = 0,
\end{aligned}
$$

we have

$$
\begin{aligned}
& [[[W_j, D_j], D_j], D_j] \\
& \qquad = \sum_{l,r=1}^n [[[B_{lr}^j D_l D_r, D_j], D_j], D_j] + \sum_{r=1}^n [[[b_r^j D_r, D_j], D_j], D_j] \\
& + \sum_{r=1}^n \overline{b}_r^j [[[D_r, D_j], D_j], D_j] + [[[p_5, D_j], D_j], D_j] \\
& \qquad = \sum_{l,r=1}^n 3 \frac{\partial B_{lr}^j}{\partial x_j} \left(\frac{\partial \omega_{jr}}{\partial x_j} D_l + \frac{\partial \omega_{jl}}{\partial x_j} D_r \right) \\
& - 3 \sum_{l,r=1}^n \left(\frac{\partial B_{lr}^j}{\partial x_j} \omega_{jr} \omega_{jl} + \frac{\partial (B_{lr}^j \omega_{jr} \omega_{jl})}{\partial x_j} \right) \qquad \text{since } B_{lr}^j \in \mathcal{P}_1 \text{ and then by (15)} \\
& - \sum_{r=1}^n \frac{\partial^3 b_r^j}{\partial x_j^3} D_r + \sum_{r=1}^n \left(3 \frac{\partial^2 b_r^j}{\partial x_j^2} \omega_{jr} + 3 \frac{\partial b_r^j}{\partial x_j} \frac{\partial \omega_{jr}}{\partial x_j} \right) \qquad \text{by (13)} \\
& - \frac{\partial^3 p_5}{\partial x_j^3} \in \mathcal{E} \qquad \text{by (11).}
\end{aligned}
\tag{3.20}
$$

So, $[[[W_j, D_j], D_j], D_j]$ is a first-order differential operator in $\mathcal{E}$. Note that the first and third sums in (3.20) are linear combinations of $D_1, \ldots, D_n$ and are elements of $\mathcal{E}$ since $b^j_{lr} \in \mathcal{P}_3$. Hence,

$$\begin{aligned} &-3\sum_{l,r=1}^{n}\left(\frac{\partial B^j_{lr}}{\partial x_j}\omega_{jr}\omega_{jl} + \frac{\partial(B^j_{lr}\omega_{jr}\omega_{jl})}{\partial x_j}\right) \\ &\quad + \sum_{r=1}^{n}\left(3\frac{\partial^2 b^j_r}{\partial x_j^2}\omega_{jr} + 3\frac{\partial b^j_r}{\partial x_j}\frac{\partial \omega_{jr}}{\partial x_j}\right) - \frac{\partial^3 p_5}{\partial x_j^3} \in \mathcal{E}\cap\mathcal{P}_2. \end{aligned} \tag{3.21}$$

Let $j > k$. The coefficients of x_j^2 in (3.21) should be zero by Theorem 2.3; that is,

$$-12\sum_{l,r=1}^{n}\frac{\partial B^j_{lr}}{\partial x_j}A_j(j,r)A_j(j,l) + \frac{9}{2}\sum_{r=1}^{n}\frac{\partial^3 b^j_r}{\partial x_j^3}A_j(j,r) - \frac{1}{2}\frac{\partial^5 p_5}{\partial x_j^5} = 0. \tag{3.22}$$

In view of (3.19), we easily check that for $j > k$

$$\begin{aligned} \frac{\partial B^j_{lr}}{\partial x_j} &= \sum_{i=1}^{n}(A_l(j,i)A_j(i,r) + A_i(j,r)A_j(i,l)) - C_{jl}(j,r) \\ &= \sum_{i=1}^{n}(A_l(j,i)A_j(i,r) + A_i(j,r)A_j(i,l)) \qquad \text{by (3.8)}, \\ \frac{\partial^3 b^j_r}{\partial x_j^3} &= \sum_{i=1}^{n}\Bigg(A_r(j,i)\frac{\partial^4\eta}{\partial x_j^3\partial x_i} + \frac{3}{2}A_j(j,i)\frac{\partial^4\eta}{\partial x_j^2\partial x_r\partial x_i} \\ &\qquad + \frac{1}{2}A_i(j,r)\frac{\partial^4\eta}{\partial x_j^3\partial x_i} - 3C_{jj}(j,i)A_j(i,r)\Bigg) \\ &= \sum_{i,l=1}^{n}(-4A_r(j,i)A_j(j,l)A_j(l,i) - 6A_j(j,i)A_j(r,l)A_j(l,i) \\ &\qquad - 2A_i(j,r)A_j(j,l)A_j(l,i)) \qquad \text{by (3.8)}, \\ \frac{\partial^5 p_5}{\partial x_j^5} &= -\sum_{i=1}^{n}\frac{1}{2}\times 10C_{jj}(j,i)\frac{\partial^4\eta}{\partial x_j^3\partial x_i} = 0 \qquad \text{by (3.8)}. \end{aligned} \tag{3.23}$$

Note that the following two terms

$$\frac{\partial^4\eta}{\partial x_j^2\partial x_r\partial x_i}, \qquad \frac{\partial^4\eta}{\partial x_j^3\partial x_i},$$

which have appeared in (3.23), can be computed in the following way. We use (3.9) with i, q, l, j being replaced, respectively, by j, j, r, i, so as to get the equation

$$\frac{\partial^4\eta}{\partial x_j^2\partial x_r\partial x_i} = -4\sum_{l=1}^{n}A_j(r,l)A_j(l,i).$$

Then, letting r be equal to j in the above equation, we have

$$\frac{\partial^4\eta}{\partial x_j^3\partial x_i} = -4\sum_{l=1}^{n}A_j(j,l)A_j(l,i).$$

Substituting (3.23) into (3.22), the left-hand side of (3.22) is equal to

$$
\begin{aligned}
(3.24)\qquad & \\
& -12\sum_{i,l,r=1}^{n}(A_l(j,i)A_j(i,r)+A_i(j,r)A_j(i,l))A_j(j,r)A_j(j,l)\\
& +\frac{9}{2}\sum_{i,l,r=1}^{n}(-4A_r(j,i)A_j(j,l)A_j(l,i)-6A_j(j,i)A_j(r,l)A_j(l,i)\\
& -2A_i(j,r)A_j(j,l)A_j(l,i))A_j(j,r)\\
=&12\sum_{i,l,r=1}^{n}(A_j(j,l)A_l(j,i)A_j(i,r)A_j(r,j)-A_j(j,l)A_j(l,i)A_i(j,r)A_j(r,j))\\
& +\sum_{i,l,r=1}^{n}(-18A_j(j,r)A_r(j,i)A_j(i,l)A_j(l,j)+27A_j(j,r)A_j(r,l)A_j(l,i)A_j(i,j)\\
& +9A_j(j,l)A_j(l,i)A_i(j,r)A_j(r,j))\\
=&\sum_{i,l,r=1}^{n}(12A_j(j,l)A_l(j,i)A_j(i,r)A_j(r,j)-3A_j(j,l)A_j(l,i)A_i(j,r)A_j(r,j))\\
& +\sum_{i,l,r=1}^{n}(-18A_j(j,r)A_r(j,i)A_j(i,l)A_j(l,j)+27A_j(j,r)A_j(r,l)A_j(l,i)A_j(i,j))\\
=&12\sum_{i,l,r=1}^{n}A_j(j,l)A_l(j,i)A_j(i,r)A_j(r,j)-3A_j^4(j,j)\\
& \qquad\left(\text{since }\sum_{r=1}^{n}A_i(j,r)A_j(r,j)=\sum_{r=1}^{n}A_j(j,r)A_j(r,i)\qquad\text{by (3.4)}\right)\\
& -18\sum_{i,l,r=1}^{n}A_j(j,r)A_r(j,i)A_j(i,l)A_j(l,j)+27A_j^4(j,j)\\
=&24A_j^4(j,j)-6\sum_{i,l,r=1}^{n}A_j(j,r)A_r(j,i)A_j(i,l)A_j(l,j),
\end{aligned}
$$

with the last equality holding since

$$\sum_{i,l,r=1}^{n}A_j(j,l)A_l(j,i)A_j(i,r)A_j(r,j)=\sum_{i,l,r=1}^{n}A_j(j,r)A_r(j,i)A_j(i,l)A_j(l,j).$$

This completes the proof. □

Lemmas 3.2 and 3.3 serve to prove Lemma 3.4, which is a new crucial result and will play a key role in the proof of Lemma 3.5.

LEMMA 3.4. *Let $\mathcal{E}$ be a finite-dimensional estimation algebra of maximal rank. Then we have*

$$A_j(j,r)=0,\qquad j=k+1,\dots,n,\quad r=1,\dots,n. \tag{3.25}$$

Proof of Lemma 3.4. We have obtained two sets of elegant equations (3.10) and (3.17). Combining them, we get

$$A_j^4(j,j) = \frac{1}{4}\sum_{l=1}^{n} A_j(l,j)\left(\sum_{r,i=1}^{n} A_j(j,r)A_r(j,i)A_j(i,l)\right) = 0, \qquad j = k+1,\dots,n. \tag{3.26}$$

Since A_j^2 is symmetric, we have

$$\begin{aligned} 0 = & A_j^4(j,j) = (A_j^2A_j^2)(j,j) = \sum_{i=1}^{n} A_j^2(j,i)A_j^2(i,j) \\ = & \sum_{i=1}^{n}[A_j^2(j,i)]^2, \qquad j = k+1,\dots,n \end{aligned} \tag{3.27}$$

and, therefore,

$$A_j^2(j,i) = 0, \quad j = k+1,\dots,n, \quad i = 1,2,\dots,n. \tag{3.28}$$

In particular,

$$A_j^2(j,j) = 0, \quad j = k+1,\dots,n. \tag{3.29}$$

While by the skew-symmetry of A_j,

$$A_j^2(j,j) = \sum_{r=1}^{n} A_j(j,r)A_j(r,j) = -\sum_{r=1}^{n}[A_j(j,r)]^2. \tag{3.30}$$

We have

$$\sum_{r=1}^{n}[A_j(j,r)]^2 = 0, \tag{3.31}$$

which immediately implies (3.25). □

REMARK 3.1. *In the original version of this paper, to prove Lemma* 3.4, *we considered the second-order differential operator W_j and then used Theorem* 3.3 *of Chen and Yau* [6, p. 1122] *to conclude that* (*for $j > k$*)

$$0 = \frac{\partial B_{jj}^j}{\partial x_j} = \sum_{i=1}^{n} \frac{\partial \omega_{ji}}{\partial x_j}\frac{\partial \omega_{ij}}{\partial x_j} = -\sum_{i=1}^{n}[A_j(j,i)]^2, \tag{3.32}$$

which immediately implies Lemma 3.4. *Unfortunately, the proof of Theorem* 3.3 *of Chen and Yau* [6] *is incorrect. In the second version, we do not use it any more. Instead, we consider the two elements* $[[V_j, D_j], D_l]$ *and* $[[[W_j, D_j], D_j], D_j]$ (*which turn out to be first-order differential operators*) *in* $\mathcal{E}$, *and through analyzing the coefficients of* x_j^2 *in these two elements, we obtain Lemmas* 3.2 *and* 3.3, *which together imply* (3.32) *and thus Lemma* 3.4 *as well.*

After Lemma 3.4 has been obtained, it is easy to deduce from (3.3) of Lemma 3.1 the following lemma.

Lemma 3.5. *Let the estimation algebra $\mathcal{E}$ be both of finite dimension and of maximal rank. Then, ω_{ij}, $k+1 \leq i,\ j \leq n$, are constants.*

Proof of Lemma 3.5. Thanks to Lemma 3.4, (3.3) of Lemma 3.1 gives

$$\sum_{r=1}^{n}[A_i(j,r)]^2 = \sum_{r=1}^{n}[A_j(i,r)]^2 = \frac{1}{2}\sum_{r=1}^{n}A_j(i,r)A_i(j,r), \qquad i,\ j = k+1,\dots,n. \tag{3.33}$$

Therefore, for $i,\ j = k+1,\dots,n$,

$$\begin{aligned} &\sum_{r=1}^{n}[A_i(j,r)]^2 + \sum_{r=1}^{n}[A_j(i,r)]^2 + \sum_{r=1}^{n}[A_j(i,r) - A_i(j,r)]^2 \\ =&2\sum_{r=1}^{n}[A_i(j,r)]^2 + 2\sum_{r=1}^{n}[A_j(i,r)]^2 - 2\sum_{r=1}^{n}A_j(i,r)A_i(j,r) = 0, \end{aligned} \tag{3.34}$$

which immediately implies

$$A_i(j,r) = 0 \quad \forall\ i,\ j = k+1,\dots,n, \quad r = 1,\dots,n. \tag{3.35}$$

Noting the cyclic relation (2.7), we get

$$A_r(i,j) = -A_j(r,i) - A_i(j,r) = A_j(i,r) - A_i(j,r) \quad \forall\, i,j = k+1,\dots,n, \quad r = 1,\dots,n,$$

which, together with (3.35), gives

$$A_r(i,j) = 0 \quad \forall\ i,\ j = k+1,\dots,n, \quad r = 1,\dots,n. \tag{3.36}$$

Recall that $\omega_{ij} \in \mathcal{P}_1$ for $i,\ j = 1,\dots,n$, so that (3.36) implies that $\omega_{i,j},\ k+1 \leq i,\ j \leq n$ are constants. □

Remark 3.2. *Here we realize a new scheme to prove Lemma* 3.5. *First show $A_j(j,r) = 0$ (Lemma* 3.4*) and then show $A_i(j,r) = 0$ for $i \neq j$ and $i \neq r$. This is different from Chen, Yau, and Leung* [7, pp. 1137–1138]*, who first used the dimension assumption $n \leq 4$ and Lemma* 3.1 *to show $A_i(j,r) = 0$ for $i \neq j$ and $i \neq r$, and then to show $A_j(j,r) = 0$.*

Remark 3.3. *Lemma* 3.5 *can also be derived from the inequality* (3.16) *of Chen, Yau, and Leung* [7, p. 1138] *by applying our new result, Lemma* 3.4*. However, Chen, Yau, and Leung* [7] *did not obtain Lemma* 3.4 *for arbitrary dimension n, and nothing shows that they had realized the importance of proving Lemma* 3.4*. A crucial point of this paper is that we have realized* (*since the original version of this paper*) *the key role of Lemma* 3.4 *and that we succeed in establishing Lemma* 3.4*.*

Proof of Theorem 1.1. From Lemma 3.5 and Theorem 2.4, we conclude that all ω_{ij}, $1 \leq i,\ j \leq n$, are constants. Then Theorem 2.2 concludes our proof. □

Acknowledgments. The author would like to thank one anonymous referee for his or her careful reading of the original and the revised versions of this paper. The author would like to acknowledge the support from the International Mathematical Union and the local German organizing committee, who provided the author an opportunity to present the result of this paper (as a short communication) at the International Congress of Mathematicians of 1998 in Berlin. The third and the fourth (the final) versions of this paper were done during the author's visit to Germany in the years of 1999 and 2000, and the author gratefully acknowledges the support from the Alexander von Humboldt Foundation. In addition, the author thanks Professor Tyrone E. Duncan, Professor Xunjing Li, and Professor Shige Peng for their helpful discussions, and the author thanks Professor Michael Kohlmann for his help in preparing the third version of this paper. The author would like to thank the editor Professor Steven I. Marcus for calling the author's attention on May 2, 2000 to the paper of Yau, Wu, and Wong [23]. The paper of Yau, Wu and Wong [23] is quite related to this paper. This paper had been submitted before the publication of [23], and it is independent of [23]. And last but not least, special gratitude should be given to Ms. Luis Pfferkorn, who works at the Göthe Institute in Schwäbisch Hall, and who helped the author so much in preparing the third version of this paper.

REFERENCES

[1] R. W. Brockett, *Nonlinear control theory and differential geometry*, in Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983) PWN, Warsaw, 1984, pp. 1357–1368.

[2] R. W. Brockett, *Nonlinear systems and nonlinear estimation theory*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981, pp. 441–477.

[3] R. W. Brockett and J. M. C. Clark, *The geometry of the conditional density functions*, in Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs et al., eds., Academic Press, New York, 1980, pp. 299–309.

[4] J. Chen, S. S.-T. Yau, and C.-W. Leung, *Finite-dimensional filters with nonlinear drift* IV: *Classification of finite-dimensional estimation algebras of maximal rank with state-space dimension* 3, SIAM J. Control Optim., 34 (1996), pp. 179–198.

[5] J. Chen and S. S.-T. Yau, *Finite-dimensional filters with nonlinear drift* VI: *Linear structure of* Ω, Math. Control Signals Systems, 9 (1996), pp. 370–385.

[6] J. Chen and S. S.-T. Yau, *Finite-dimensional filters with nonlinear drift* VII: *Mitter conjecture and structure of* η, SIAM J. Control Optim., 35 (1997), pp. 1116–1131.

[7] J. Chen, S. S.-T. Yau, and C.-W. Leung, *Finite-dimensional filters with nonlinear drift* VIII: *Classification of finite-dimensional estimation algebras of maximal rank with state-space dimension* 4, SIAM J. Control Optim., 35 (1997), pp. 1132–1141.

[8] W.-L. Chiou and S. S.-T. Yau, *Finite-dimensional filters with nonlinear drift* II: *Brockett's problem on classification of finite-dimensional estimation algebras*, SIAM J. Control Optim., 32 (1994), pp. 297–310.

[9] M. Cohen De Lara, *Finite-dimensional filters. Part* I: *The Wei–Norman technique*, SIAM J. Control Optim., 35 (1997), pp. 980–1001.

[10] M. Cohen De Lara, *Finite-dimensional filters. Part* II: *Invariance group techniques*, SIAM J. Control Optim., 35 (1997), pp. 1002–1029.

[11] M. H. A. Davis and S. I. Marcus, *An introduction to nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981, pp. 53–75.

[12] R. T. Dong, L. F. Tam, W. S. Wong, and S. S.-T. Yau, *Structure and classification theorems of finite-dimensional exact estimation algebras*, SIAM J. Control Optim., 29 (1991), pp. 866–877.

[13] T. E. Duncan, *Filtering and estimation, nonlinear*, in Encyclopedia of Electrical and Electronics Engineering, Volume 7, John G. Webster, ed., John Wiley and Sons, New York, 1999, pp. 480–493.

[14] S. K. Mitter, *On the analogy between mathematical problems of nonlinear filtering and quantum physics*, Ricerche Automat., 10 (1979), pp. 163–216.

[15] D. L. Ocone, *Finite dimensional estimation algebras in nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981, pp. 629–636.

[16] S. I. Marcus, *Algebraic and geometric methods in nonlinear filtering*, SIAM J. Control Optim., 22 (1984), pp. 817–844.

[17] L.-F. Tam, W. S. Wong, and S. S.-T. Yau, *On a necessary and sufficient condition for finite dimensionality of estimation algebras*, SIAM J. Control Optim., 28 (1990), pp. 173–185.

[18] S. Tang, *Brockett's problem of classification of finite-dimensional estimation algebras for nonlinear filtering systems*, in Abstracts of Short Communications and Poster Sessions, International Congress of Mathematicians, Berlin, 1998, p. 352; also available online from http://www.mathematik.uni-bielefeld.de/icm98.

[19] W. S. Wong, *New classes of finite dimensional nonlinear filters*, Systems Control Lett., 3 (1983), pp. 155–164.

[20] W. S. Wong, *On a new class of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 79–83.

[21] W. S. Wong, *Theorems on the structure of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 117–124.

[22] S. S.-T. Yau, *Finite dimensional filters with nonlinear drift* I: *A class of filters including both Kalman-Bucy filters and Benes filters*, J. Math. Systems Estim. Control, 4 (1994), pp. 181–203.

[23] S. S.-T. Yau, X. Wu, and W. S. Wong, *Hessian matrix non-decomposition theorem*, Math. Res. Lett., 6 (1999), pp. 663–673.

SIAM J. Control Optim.
Vol. 39, No. 3, pp. 917–927

ASYMPTOTIC OPTIMALITY OF APPROXIMATE FILTERS IN STOCHASTIC SYSTEMS WITH COLORED NOISES*

ALAIN LE BRETON† AND MARIE-CHRISTINE ROUBAUD‡

Abstract. Approximate filters are proposed for semilinear and nonlinear stochastic systems with colored noises. Basically these filters are defined as those which are optimal when the noises are white. Their long time behavior is investigated and their asymptotic optimality is shown in two cases under some reasonable assumptions.

Key words. nonlinear filtering, colored noises, approximate filters, asymptotic optimality

AMS subject classifications. 93E11, 60G35

PII. S0363012998333906

1. Introduction.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ be a filtered probability space. Let $X = (X_t,\, t \geq 0)$ and $Y = (Y_t,\, t \geq 0)$ be processes, taking values in $\mathbb{R}^d$ and $\mathbb{R}^p$, respectively, uniquely defined by the system of stochastic differential equations

$$(1)\qquad \left\{ \begin{array}{lcll} dX_t &=& b(X_t)dt + d\xi_t, & t \geq 0,\ X_0, \\ dY_t &=& h(X_t)dt + d\eta_t, & t \geq 0, \quad Y_0 = 0. \end{array} \right.$$

Here X_0 stands for some $\mathcal{F}_0$-measurable random initial condition for X, with a given distribution ν_0, and $\xi = (\xi_t,\, t \geq 0)$ and $\eta = (\eta_t,\, t \geq 0)$ are $(\mathcal{F}_t)$-adapted noise processes. Supposing that only Y is observed but one wishes to know X, the classical problem of filtering the signal X at time t from the observation of Y up to time t occurs. The solution to this problem is the conditional distribution Π_t of X_t given $\mathcal{Y}_t = \sigma(\{Y_s\,,\, 0 \leq s \leq t\})$, which we shall call the *exact filter.*

If in model (1) the functions b and h are linear and the processes ξ and η are $(\mathcal{F}_t)$-Brownian motions, then it is known from Ocone and Pardoux [9] that, under stabilizability and detectability assumptions, for any distribution of ν_0 such that $\mathbb{E}[|X_0|^2] < +\infty$, the exact filter is, asymptotically in time, approached by a Kalman filter. In the same paper, Ocone and Pardoux proved also that in the case of a signal with limiting ergodic behavior observed in additive white noise through a nonlinear channel, the solution to the filtering equations initialized with an incorrect prior distribution approaches the exact filter as time goes to infinity.

In the present paper we consider the filtering model (1) when the processes ξ and η are possibly nonindependent and non-Gaussian colored noises. We assume that

$$(2)\qquad \left\{ \begin{array}{lcll} d\xi_t &=& b_\zeta(\zeta_t)dt + dV_t, & t \geq 0, \quad \xi_0 = 0, \\ d\eta_t &=& h_\zeta(\zeta_t)dt + dW_t, & t \geq 0, \quad \eta_0 = 0. \end{array} \right.$$

Here $V = (V_t,\, t \geq 0)$ and $W = (W_t,\, t \geq 0)$ are independent $(\mathcal{F}_t)$-Brownian motions in $\mathbb{R}^d$ and $\mathbb{R}^p$, respectively. We assume that $\mathbb{E}[V_1 V_1'] = \Sigma_V$ for some given symmetric

[*] Received by the editors February 6, 1998; accepted for publication (in revised form) June 7, 2000; published electronically October 11, 2000.

http://www.siam.org/journals/sicon/39-3/33390.html

[†] IDOPT (CNRS, UJF, INPG, INRIA) LMC–IMAG, Université J. Fourier, BP 53, 38041 Grenoble Cedex 9, France (Alain.Le-Breton@imag.fr).

[‡] SYSDYS–LATP, 38 rue Joliot–Curie, 13451 Marseille Cedex 20, France (mcroubau@sophia.inria.fr). This author's research was done in part while staying at the Université J. Fourier of Grenoble, where the author was supported by the project IDOPT up to September 1997.

nonnegative definite matrix Σ_V, and $\mathbb{E}[W_1 W_1'] = I_p$, where I_p denotes the $p \times p$ identity matrix. Moreover, $\zeta = (\zeta_t,\, t \geq 0)$ is an $(\mathcal{F}_t)$-adapted perturbation process taking values in $\mathbb{R}^k$. We suppose also that the pair (ζ, X_0) is independent of (V, W). Of course, in model (1)–(2), the exact filter Π_t for X is nothing but the marginal distribution of the exact filter for the pair (ζ, X). The determination of Π_t, which is in general an infinite dimensional problem, requires the complete knowledge of the joint probabilistic structure of ζ, X, and Y and leads to computations in a space dimension which is larger than that of the system of interest. Hence it is interesting to elaborate *approximate filters* which ignore a part of the actual structure of the system, and then are more easily computable, but may have a long time behavior close to that of the exact filter. Here this question is addressed under the assumptions that

$$(\mathrm{A}_0) \qquad b_\zeta(0) = 0, \qquad h_\zeta(0) = 0,$$

and the perturbation process ζ converges to zero in some appropriate sense which will be made precise later. We propose approximate filters which are basically the optimal filters when the noises are independent white noises, i.e., $\zeta \equiv 0$ since then, from (A_0), $\xi \equiv V$ and $\eta \equiv W$.

Actually, it is convenient to work on the canonical probability space of the process (ζ, X, Y). We suppose that the paths of ζ belong to the set $\Omega_0 = \mathcal{L}^2_{loc}(\mathbb{R}^+ \,;\, \mathbb{R}^k)$ of locally square integrable functions from $\mathbb{R}^+$ into $\mathbb{R}^k$. Let us denote by $\Omega_1 = \mathcal{C}(\mathbb{R}^+ \,;\, \mathbb{R}^d)$ (resp., $\Omega_2 = \mathcal{C}_0(\mathbb{R}^+ \,;\, \mathbb{R}^p)$) the set of continuous functions from $\mathbb{R}^+$ into $\mathbb{R}^d$ (resp., $\mathbb{R}^p$ with value zero at $t = 0$). Define $\Omega = \Omega_0 \times \Omega_1 \times \Omega_2$, $\mathcal{F}$ as its Borel field, $(\mathcal{F}_t)$ as the canonical Borel filtration on Ω, and $(\zeta_t, X_t, Y_t)(z, x, y) = (z_t, x_t, y_t)$ for any $(z, x, y) \in \Omega$. Let $\mathbb{P}$ be the distribution on Ω corresponding to the filtering model defined by (1)–(2). Similarly, let $\mathbb{P}^0$ be the distribution on Ω corresponding to the filtering model defined by (1)–(2) when the noises are white, i.e., $\zeta \equiv 0$ since we assume that (A_0) is fulfilled. Of course, $\mathbb{P}^0$ is nothing but the product probability $\delta_0 \otimes P$, where δ_0 is the Dirac measure on Ω_0 concentrated at 0 and P is the distribution on $\Omega_1 \times \Omega_2$ corresponding to the filtering model (1) when ξ and η are independent Brownian motions and the distribution of X_0 is ν_0. In what follows the symbols $\mathbb{E}[.]$ and $\mathbb{E}^0[.]$ will stand for an expectation or a conditional expectation computed with respect to $\mathbb{P}$ and $\mathbb{P}^0$, respectively. So for the exact filter Π_t of X_t, we write

$$\Pi_t(\varphi) := \mathbb{E}[\varphi(X_t) \,|\, \mathcal{Y}_t]$$

for any φ in the set $\mathcal{M}_b(\mathbb{R}^d)$ of all bounded measurable functions from $\mathbb{R}^d$ into $\mathbb{R}$. Similarly, when X_t is $\mathbb{P}$-integrable, for the conditional expectation we write $\hat{X}_t := \mathbb{E}[X_t \,|\, \mathcal{Y}_t]$.

The basic approximate filter is taken as the conditional distribution Π^0_t of X_t given $\mathcal{Y}_t$ with respect to $\mathbb{P}^0$, and we write also $\Pi^0_t(\varphi) := \mathbb{E}^0[\varphi(X_t) \,|\, \mathcal{Y}_t]$ for any function φ as above and, when it is well defined, $\hat{X}^0_t := \mathbb{E}^0[X_t \,|\, \mathcal{Y}_t]$. We shall also consider approximate filters defined similarly to Π^0 but with some more convenient prior distribution ν in place of the true one ν_0. This will be made precise later.

The paper is organized as follows. In section 2, we consider what we call the *semilinear case,* where the functions b and h are linear but functions b_ζ and h_ζ may be nonlinear. We derive a representation of the exact filter which extends a formula obtained by Makowski [8] and Beněs and Karatzas [1] (see also Haussmann and Pardoux [4]) for a linear system with white noises and non-Gaussian initial conditions. Here Kalman filters can be taken as approximate filters and their asymptotic analysis is led under some stabilizability and detectability assumptions. In section 3, a *nonlinear case* where the signal has a limiting ergodic behavior is investigated. Here the approximate filter appears as the optimal filter corresponding to an incorrect prior distribution, and the asymptotic study uses the results of Ocone and Pardoux [9] on the asymptotic stability of optimal filters with respect to their initial condition.

2. The semilinear case.

In this section, we are interested in the case of a *semilinear system* where the signal and observation dynamics are linear with respect to the state but the action of the pertubation process ζ can be nonlinear. Precisely, here we assume that in model (1)–(2) the condition (A$_0$) is fulfilled and, moreover, the following conditions hold:

(A$_1$) $b(x) = Bx,\ h(x) = Hx,\ x \in \mathbb{R}^d$, and b_ζ and h_ζ are Lipschitzian functions;

(A$_2$) $(B, \Sigma_V^{\frac{1}{2}})$ is a stabilizable pair, and (B, H) is a detectable pair;

(A$_3$) for all $t \geq 0$ $\mathbb{E}\left[\int_0^t |\zeta_s|^2 ds\right] < +\infty$, and $\lim_{t \to +\infty} \mathbb{E}[|\zeta_t|^2] = 0$.

Then the process (X, Y) taken under $\mathbb{P}$ is the solution to the filtering model

$$\left\{ \begin{array}{rcl} dX_t & = & BX_t dt + b_\zeta(\zeta_t)dt + dV_t, \quad t \geq 0,\ X_0, \\ dY_t & = & HX_t dt + h_\zeta(\zeta_t)dt + dW_t, \quad t \geq 0, \quad Y_0 = 0. \end{array} \right. \tag{3}$$

Obviously under the probability $\mathbb{P}^0$, which corresponds to $\zeta \equiv 0$, the system may be reduced to the linear system

$$\left\{ \begin{array}{rcl} dX_t & = & BX_t\, dt + dV_t, \quad t \geq 0,\ X_0, \\ dY_t & = & HX_t\, dt + dW_t, \quad t \geq 0, \quad Y_0 = 0. \end{array} \right. \tag{4}$$

In what follows, the notation $\mathcal{N}(\mu, \Lambda)$ will be used for the Gaussian law on some space $\mathbb{R}^l$ with mean $\mu \in \mathbb{R}^l$ and covariance matrix Λ. For any vector $x \in \mathbb{R}^d$ and any $d \times d$ matrix $R \geq 0$, let $\hat{X}_t^{x,R}$ and Q_t^R be the solutions to the Kalman filtering equations corresponding to model (4) initialized with (x, R), i.e.,

$$d\hat{X}_t^{x,R} = B\hat{X}_t^{x,R}\, dt + Q_t^R H^{'} [dY_t - H\hat{X}_t^{x,R} dt], \quad t \geq 0, \quad \hat{X}_0^{x,R} = x, \tag{5}$$

$$\dot{Q}_t^R = BQ_t^R + Q_t^R B^{'} + \Sigma_V - Q_t^R H^{'} HQ_t^R, \quad t \geq 0, \quad Q_0^R = R. \tag{6}$$

The Gaussian distribution $\Pi_t^{x,R} = \mathcal{N}(\hat{X}_t^{x,R}, Q_t^R)$ will be referred to as a Kalman filter initialized with (x, R). Of course, if $\nu_0 = \mathcal{N}(m_0, Q^0)$, the basic approximate filter Π_t^0 in model (3) is the Kalman filter initialized with (m_0, Q_0). Actually, whatever is the prior distribution ν_0, all Kalman filters $\Pi_t^{x,R} = \mathcal{N}(\hat{X}_t^{x,R}, Q_t^R)$, $x \in \mathbb{R}^d$, $R \geq 0$, are candidates as approximate filters. The key point for the asymptotic study of these filters relies on a representation of the exact filter Π_t which we derive now.

2.1. Representation of the exact filter.

For any $z = (z_t, t \geq 0) \in \Omega_0$ and $x_0 \in \mathbb{R}^d$, let us set

$$\zeta_t^*(z, x_0) = e^{Bt}x_0 + \int_0^t e^{B(t-s)} b_\zeta(z_s) ds, \quad t \geq 0, \tag{7}$$

$$\langle K\rangle_t^*(z,x_0) := \int_0^t |H\zeta_s^*(z,x_0)+h_\zeta(z_s)|^2 ds. \tag{8}$$

Omitting the dependencies on (z,x_0) as will be done often from now on, we introduce also the Kalman-type equations for processes $\hat{X}_t^* \in \mathbb{R}^d$, $\hat{K}_t^* \in \mathbb{R}$ and functions $Q_t^* \in \mathbb{R}^{d\times d}$, $S_t \in \mathbb{R}^d$, and $T_t \in \mathbb{R}$.

$$d\hat{X}_t^* = B\hat{X}_t^* dt + Q_t^* H^{'}[dY_t - H\hat{X}_t^* dt], \quad t \geq 0, \quad \hat{X}_0^* = 0, \tag{9}$$

$$d\hat{K}_t^* = [H\zeta_t^* + h_\zeta(z_t) + HS_t]^{'}[dY_t - H\hat{X}_t^* dt], \quad t \geq 0, \quad \hat{K}_0^* = 0, \tag{10}$$

$$\dot{Q}_t^* = BQ_t^* + Q_t^* B^{'} + \Sigma_V - Q_t^* H^{'} H Q_t^*, \quad t \geq 0, \quad Q_0^* = O, \tag{11}$$

$$\dot{S}_t = (B - Q_t^* H^{'} H)S_t - Q_t^* H^{'}[H\zeta_t^* + h_\zeta(z_t)], \quad t \geq 0, \quad S_0 = 0, \tag{12}$$

$$\dot{T}_t = -2[H\zeta_t^* + h_\zeta(z_t)]^{'} HS_t - S_t^{'} H^{'} HS_t, \quad t \geq 0, \quad T_0 = 0. \tag{13}$$

Finally, we define

$$C_t := \begin{pmatrix} Q_t^* & S_t \\ S_t^{'} & T_t \end{pmatrix}. \tag{14}$$

The following statement extends to the semilinear model (3) a formula obtained by Makowski [8] and Beněs and Karatzas [1] (see also Haussmann and Pardoux [4]) for a linear system with white noises and non-Gaussian initial conditions.

Proposition 2.1. *Let Π_t be the exact filter in model* (3) *(taken under $\mathbb{P}$). Then for any $\varphi \in \mathcal{M}^+(\mathbb{R}^d)$ the following holds:*

$$\Pi_t(\varphi) = \frac{\displaystyle\int_{\Omega_0\times\mathbb{R}^d} e^{[-\frac{1}{2}\langle K\rangle_t^* + \frac{1}{2}T_t + \hat{K}_t^*](z,x_0)} J_t(z,x_0) d\mathbb{P}_{(\zeta,X_0)}(z,x_0)}{\displaystyle\int_{\Omega_0\times\mathbb{R}^d} e^{[-\frac{1}{2}\langle K\rangle_t^* + \frac{1}{2}T_t + \hat{K}_t^*](z,x_0)} d\mathbb{P}_{(\zeta,X_0)}(z,x_0)}, \tag{15}$$

with

$$J_t(z,x_0) = \int_{\mathbb{R}^{d+1}} \varphi(\zeta_t^*(z,x_0)+u)\bar{n}_t^{z,x_0}(du,dv), \tag{16}$$

where for all $z = (z_t, t \geq 0) \in \Omega_0$, $x_0 \in \mathbb{R}^d$, and $t \geq 0$ the symbol $\bar{n}_t^{z,x_0}$ stands for the Gaussian distribution on $\mathbb{R}^{d+1}$ with mean $([\hat{X}_t^* + S_t(z,x_0)]', \hat{K}_t^*(z,x_0) + T_t(z,x_0))'$ and covariance $C_t(z,x_0)$. Here the quantities depending on (z,x_0) are defined by (7)–(14).

Proof. From (3) we may decompose the signal X as

$$X_t = \zeta_t^* + X_t^*, \quad t \geq 0, \tag{17}$$

where

$$\zeta_t^* = \zeta_t^*(\zeta, X_0); \quad X_t^* = \int_0^t e^{B(t-s)} dV_s, \quad t \geq 0.$$

Then by means of an appropriate change of probability (see Le Breton and Roubaud [7] for details), applying the Girsanov theorem and the classical Bayes formula, one can show that

$$\Pi_t(\varphi) = \frac{\displaystyle\int_{\Omega_0\times\mathbb{R}^d} e^{-\frac{1}{2}\langle K\rangle_t^*(z,x_0)} I_t^{(1)}(z,x_0)d\mathbb{P}_{(\zeta,X_0)}(z,x_0)}{\displaystyle\int_{\Omega_0\times\mathbb{R}^d} e^{-\frac{1}{2}\langle K\rangle_t^*(z,x_0)} I_t^{(2)}(z,x_0)d\mathbb{P}_{(\zeta,X_0)}(z,x_0)},$$

with

$$I_t^{(1)}(z,x_0) = \int_{\mathbb{R}^{d+1}} \varphi(\zeta_t^*(z,x_0)+u)e^v n_t^{z,x_0}(du,dv),$$
$$I_t^{(2)}(z,x_0) = \int_{\mathbb{R}^{d+1}} e^v n_t^{z,x_0}(du,dv).$$

Here n_t^{z,x_0} is the conditional distribution of $Z_t(z,x_0) = (X_t^{*\prime}, K_t^*(z,x_0))'$ given $\mathcal{Y}_t$ in the model

$$\begin{cases} dX_t^* &= BX_t^*dt + dV_t, \quad t\geq 0, \quad X_0^* = 0, \\ dK_t^*(z,x_0) &= [H\zeta_t^*(z,x_0) + h_\zeta(z_t)]'dW_t^*, \quad t\geq 0, \quad K_0^*(z,x_0) = 0, \\ dY_t &= HX_t^*dt + dW_t^*, \quad t \geq 0, \quad Y = 0, \end{cases}$$

where W^* is some Brownian motion independent of V.

Actually, n_t^{z,x_0} is nothing but the Gaussian distribution $\mathcal{N}(\hat{Z}_t(z,x_0),\ C_t(z,x_0))$, where $\hat{Z}_t(z,x_0) = (\hat{X}_t^{*'}, \hat{K}_t^*(z,x_0))'$ and $C_t(z,x_0)$ are given by (9)–(14). Hence in order to get (15)–(16) it remains to show that the probability distribution $\bar{n}_t^{z,x_0}$ defined by

$$\bar{n_t}^{z,x_0}(du,dv) := e^{[-\frac{1}{2}T_t - \hat{K}_t^*](z,x_0)}e^v n_t^{z,x_0}(du,dv)$$

is nothing but the Gaussian distribution on $\mathbb{R}^{d+1}$ with mean $([\hat{X}_t^* + S_t(z,x_0)]', \hat{K}_t^*(z,x_0) + T_t(z,x_0))'$ and covariance $C_t(z,x_0)$. A direct calculation of the characteristic functional of $\bar{n}_t^{z,x_0}$ gives the result. □

Remark 2.2. Notice that substituting $\mathbb{P}^0$ for $\mathbb{P}$ in Proposition 2.1, i.e., taking $\zeta\equiv 0$ and then also $\zeta_t^*(z,x_0)\equiv e^{Bt}x_0$, gives a representation of the approximate filter Π_t^0. Of course, the corresponding formula reduces to the one already known for the optimal filter in a linear system with non-Gaussian initial condition.

Remark 2.3. Comparing (9) and (11) to (5)–(6), it is clear that $\hat{X}_t^* = \hat{X}_t^{0,O}$ and $Q_t^* = \hat{Q}_t^O$, where $\hat{X}_t^{0,O}$ and Q_t^O are the solutions of (5) and (6) corresponding to initial conditions 0 and O, respectively. In other words, the marginal distribution of n_t^{z,x_0} on $\mathbb{R}^d$ is nothing but the approximate Kalman filter $\Pi_t^{0,O}$ for X_t, defined by (5)–(6), initialized with $(0,O)$.

2.2. Asymptotic optimality of the approximate filter. Now we can show that an approximate Kalman filter $\Pi_t^{x,R} = \mathcal{N}(\hat{X}_t^{x,R}, Q_t^R)$, defined by (5)–(6) and initialized with arbitrary $x\in\mathbb{R}^d$ and $R\geq 0$, is asymptotically optimal when $\mathbb{E}[|X_0|^2] < +\infty$. The following extends to the case of colored noises in the signal and in the observation, the statement obtained for white noises by Ocone and Pardoux [9, Theorem 2.6].

PROPOSITION 2.4. *Assume that in model* (3)*, the conditions* $(A_0), (A_1), (A_2)$*, and* (A_3) *are fulfilled. Suppose also that* $\mathbb{E}[|X_0|^2] < +\infty$*. Then for any* $x\in\mathbb{R}^d$ *and any* $d\times d$ *matrix* $R\geq 0$*,*

$$\lim_{t\,\to\,\infty} \mathbb{E}[|\hat{X}_t - \hat{X}_t^{x,R}|^2] = 0, \tag{18}$$

and for all uniformly continuous $\varphi \in \mathcal{M}_b(\mathbb{R}^d)$,

$$\lim_{t \to \infty} \mathbb{E}[|\Pi_t(\varphi) - \Pi_t^{x,R}(\varphi)|^2] = 0. \tag{19}$$

Proof. The proof parallels that of [9, Theorem 2.6], and we refer to Le Breton and Roubaud [7] for details. Concerning the limiting property (19), let us point out the fact that the key point is to show that

$$\lim_{t \to \infty} \mathbb{E}[|\Pi_t(\varphi) - \mathcal{N}(\hat{X}_t^*, Q_t^*)(\varphi)|^2] = 0 \tag{20}$$

for any uniformly continuous function $\varphi \in \mathcal{M}_b(\mathbb{R}^d)$. Observe that

$$\Pi_t(\varphi) - \mathcal{N}(\hat{X}_t^*, Q_t^*)(\varphi) = \mathbb{E}[\varphi(X_t) \,|\, \mathcal{Y}_t] - \mathbb{E}[\varphi(\hat{X}_t^* + U_t)],$$

where for each $t \geq 0$, U_t is a Gaussian vector with mean zero and covariance matrix Q_t^*. From (15)–(16) we have

$$\begin{aligned} &\Pi_t(\varphi) - \mathcal{N}(\hat{X}_t^*, Q_t^*)(\varphi) \\ &\qquad = \frac{\displaystyle\int_{\Omega_0 \times \mathbb{R}^d} e^{\left(-\frac{1}{2}\langle K\rangle_t^* + \frac{1}{2}T_t + \hat{K}_t^*\right)(z,x_0)} \Delta_t(z, x_0) d\mathbb{P}_{(\zeta, X_0)}(z, x_0)}{\displaystyle\int_{\Omega_0 \times \mathbb{R}^d} e^{\left(-\frac{1}{2}\langle K\rangle_t^* + \frac{1}{2}T_t + \hat{K}_t^*\right)(z,x_0)} d\mathbb{P}_{(\zeta, X_0)}(z, x_0)}, \end{aligned} \tag{21}$$

where

$$\Delta_t(z, x_0) = \mathbb{E}[\varphi(\hat{X}_t^* + (S_t + \zeta_t^*)(z, x_0) + U_t)] - \mathbb{E}[\varphi(\hat{X}_t^* + U_t)].$$

Let $\varepsilon > 0$. Choose $\eta > 0$ such that for all y and $y' \in \mathbb{R}^d$, if $|y - y'| \leq \eta$, then

$$|\varphi(y) - \varphi(y')|^2 < \frac{\varepsilon}{4}.$$

Decompose the integral in the numerator of (21) into the sum of the integral over the region $\{|(\zeta_t^* + S_t)(z, x_0)| < \eta\}$ and the integral over the region $\{|(\zeta_t^* + S_t)(z, x_0)| \geq \eta\}$. We obtain

$$|\Pi_t(\varphi) - \mathcal{N}(\hat{X}_t^*, Q_t^*)(\varphi)| \leq \sup_{|y - y'| < \eta} |\varphi(y) - \varphi(y')| + 2||\varphi||_\infty \mathbb{E}[\mathbb{1}_{\{|(\zeta_t^* + S_t)| \geq \eta\}} \,|\, \mathcal{Y}_t],$$

and therefore,

$$\mathbb{E}[|\Pi_t(\varphi) - \mathcal{N}(\hat{X}_t^*, Q_t^*)(\varphi)|^2] \leq \frac{\varepsilon}{2} + \frac{8||\varphi||_\infty^2}{\eta^2} \mathbb{E}[|\zeta_t^* + S_t|^2].$$

Since actually $\lim_{t \to \infty} \mathbb{E}[|\zeta_t^* + S_t|^2] = 0$, there is a $T > 0$ such that for all $t > T$,

$$\mathbb{E}[|\Pi_t(\varphi) - \mathcal{N}(\hat{X}_t^*, Q_t^*)(\varphi)|^2] \leq \varepsilon.$$

Thus (20) is obtained, which achieves the proof. □

Remark 2.5. One consequence of Proposition 2.4 is that in the considered case of a nonlinear action of the pertubation process ζ on the basically linear dynamics of the signal-observation process, the approximate filter is "asymptotically insensitive to perturbations of its initial condition."

Remark 2.6. Actually, it is possible to allow some correlation between the signal X and the observation noise in model (3). The statement of Proposition 2.4 can be easily extended to the case where in the state equation for X a term of the form $D\,dY_t$, $D \in \mathbb{R}^{d \times p}$, is added.

3. The nonlinear case. In this section we consider the case where in model (1)–(2) all functions b, b_ζ, h, and h_ζ may be nonlinear but the system has a limiting ergodic behavior. Here, additionally to (A_0), we assume that the following conditions hold.

(A_1^*) b, b_ζ, and h_ζ are Lipschitzian; h is continuous; h_ζ and h are bounded.

(A_2^*) There exist two constants $r > 0$ and $\alpha > 0$ such that for $x \in \mathbb{R}^d$,

$$x'b(x) \leq -\alpha|x| \quad \text{if} \quad |x| > r.$$

We suppose also that the process ζ is generated by the ordinary differential equation

$$\dot{\zeta}_t = a(\zeta_t), \quad t \geq 0, \quad \zeta_0,$$

where a is a Lipschitzian function from $\mathbb{R}^k$ into $\mathbb{R}^k$. Concerning the deterministic flow $\phi = (\phi_t, t \geq 0)$ associated with that equation and the initial condition ζ_0, we assume that

(A_3^*) $\phi(0) \equiv 0$ and ϕ is contracting with some exponential rate $\lambda > 0$, i.e.,

$$|\phi_t(z_1) - \phi_t(z_2)| \leq e^{-\lambda t}|z_1 - z_2|, \quad t \geq 0, \quad z_1, z_2 \in \mathbb{R}^k.$$

The random variable ζ_0 is $\mathcal{F}_0$-measurable and $\mathbb{E}[\exp c_0|\zeta_0|^2] < +\infty$ for $c_0 = K_\zeta^2/\lambda$, where K_ζ is the maximum of the Lipschitz constants for b_ζ and h_ζ. Finally here we assume that the Brownian motion V is nondegenerate, i.e., the matrix Σ_V is positive.

The concerned process (ζ, X, Y) taken under $\mathbb{P}$ on Ω (where actually Ω_0 may be reduced to $\mathcal{C}(\mathbb{R}^+; \mathbb{R}^k)$) is the $(\mathcal{F}_t)$-Markov diffusion process solution of the system

$$(22) \qquad \begin{cases} d\zeta_t &= a(\zeta_t)dt, \, t \geq 0, \; \zeta_0, \\ dX_t &= b(X_t)dt + b_\zeta(\zeta_t)dt + dV_t, \quad t \geq 0, \; X_0, \\ dY_t &= h(X_t)dt + h_\zeta(\zeta_t)dt + dW_t, \quad t \geq 0, \quad Y_0 = 0. \end{cases}$$

The determination of the exact filter Π_t is, in general, an infinite dimensional problem. Actually, the evolution in time of Π_t can be described by the so-called Zakai equation (cf. [11] or, e.g., [2], [10]), where, of course, the spatial dimension is $k + d$ and the initial condition is $\mathbb{P}_{(\zeta_0, X_0)}$, which will be denoted by π_0.

Under the probability $\mathbb{P}^0$, which here corresponds to $\zeta_0 = 0$, the system may be reduced to

$$(23) \qquad \begin{cases} dX_t &= b(X_t)dt + dV_t, \quad t \geq 0, \; X_0, \\ dY_t &= h(X_t)dt + dW_t, \quad t \geq 0, \quad Y_0 = 0. \end{cases}$$

Then the evolution in time of the approximate filter Π_t^0 is described by a Zakai equation for which numerical computations are easier since the spatial dimension is d only and the initial condition is the distribution ν_0 of X_0.

Here the asymptotic study of the approximate filter will use results of Ocone and Pardoux [9] on the asymptotic stability of optimal filters with respect to perturbations to their initial condition for systems with ergodic limiting behavior. So now we summarize the material we need from [9], restricting ourselves to a specific situation which is appropriate to our purpose.

3.1. Stability of filters with respect to their initial condition. Consider a filtering model specified by a signal-observation pair (Z, Y) such that the following hold.

- The signal $Z = (Z_t,\, t \geq 0)$ is a continuous, $\mathbb{R}^n$-valued homogeneous Markov process with some initial distribution π on $\mathbb{R}^n$.
- The $\mathbb{R}^p$-valued observation process Y is defined by

$$Y_t = \int_0^t H(Z_s)\, ds \,+\, W_t, \quad t \geq 0,$$

where W is a $\mathbb{R}^p$-valued standard Brownian motion independent of Z and H is a bounded continuous function from $\mathbb{R}^n$ into $\mathbb{R}^p$.

Let us work as before on the canonical stochastic basis, which here can be taken as $\Omega = \mathcal{C}(\mathbb{R}^+; \mathbb{R}^n) \times \mathcal{C}_0(\mathbb{R}^+; \mathbb{R}^p)$. We denote by $\mathbb{P}^\pi$ the distribution on Ω corresponding to the above model specified by the initial condition π. We denote also by $\mathbb{E}^\pi[.]$ an expectation computed with respect to $\mathbb{P}^\pi$ so that the optimal filter Π_t^π for Z_t corresponding to the prior distribution π is defined by

$$\Pi_t^\pi(\psi) = \mathbb{E}^\pi[\psi(Z_t)|\mathcal{Y}_t], \quad \psi \in \mathcal{M}_b(\mathbb{R}^n).$$

Given two different initial conditions π_0 (the *true prior distribution*, say) and $\bar{\pi}_0$ (an *incorrect prior distribution*), Ocone and Pardoux [9] studied the long time behavior of the difference $\Pi_t^{\pi_0}(\psi) - \Pi_t^{\bar{\pi}_0}(\psi)$ under the probability $\mathbb{P}^{\pi_0}$. They introduce some conditions which in the present case may be stated as follows. Let $\mathcal{C}_b(\mathbb{R}^n)$ be the set of those functions in $\mathcal{M}_b(\mathbb{R}^n)$ which are continuous, and let $(S_t, t \geq 0)$ denote the transition semigroup of the process Z.

Assume the following.

(H$_1$) $(S_t, t \geq 0)$ is a strongly continuous semigroup satisfying the Feller property, i.e., $S_t(\mathcal{C}_b(\mathbb{R}^n)) \subset \mathcal{C}_b(\mathbb{R}^n)$ for all $t \geq 0$, and admitting a unique invariant measure μ such that for all $\psi \in \mathcal{C}_b(\mathbb{R}^n)$

$$\limsup_{t \,\to\, +\infty} \int_{\mathbb{R}^n} |S_t\psi(z) - \mu(\psi)|\, \mu(dz) = 0.$$

Then, denoting by πS_t the distribution of the random variable Z_t under $\mathbb{P}^\pi$, it is said that $(S_t, t \geq 0)$ forgets π for μ if

$$(\mathrm{H}_2(\pi)) \qquad \pi S_t \,\to\, \mu \;\text{ weakly as }\; t \,\to\, +\infty.$$

In what follows, R^π denotes the marginal distribution of $\mathbb{P}^\pi$ on $\Omega_2 = \mathcal{C}_0(\mathbb{R}^+; \mathbb{R}^p)$, i.e., the distribution of the observation process Y corresponding to the initial distribution π for the signal Z. Then, applying [9, Theorem 3.2] to the just described situation, we get the following lemma.

LEMMA 3.1. *Assume that*

(i) $(S_t, t \geq 0)$ *satisfies* (H1),

(ii) $(\mathrm{H}_2(\pi_0))$ *and* $(\mathrm{H}_2(\bar{\pi}_0))$ *are both satisfied, and*

(iii) R^{π_0} *is absolutely continuous with respect to* $R^{\bar{\pi}_0}$.

Then, for every continuous $\psi \in \mathcal{M}_b(\mathbb{R}^n)$,

$$\lim_{t \,\to\, +\infty} \mathbb{E}^{\pi_0}[|\Pi_t^{\pi_0}(\psi) - \Pi_t^{\bar{\pi}_0}(\psi)|^2] = 0.$$

3.2. Asymptotic optimality of the approximate filter. Now we show that the version above of the result of Ocone and Pardoux is appropriate to address our problem concerning the asymptotic optimality of the approximate filter in model (22). Of course, we may look at this model as a signal-observation model of the type which has just been discussed in section 3.1. The process $Z = (\zeta', X')'$ is the $\mathbb{R}^{k+d}$-valued Markovian signal with initial condition $\pi_0 = \mathbb{P}_{(\zeta_0, X_0)}$. The exact (resp., approximate) filter Π_t (resp., Π_t^0) is the optimal filter $\Pi_t^{\pi_0}$ (resp., $\Pi_t^{\bar{\pi}_0}$) corresponding to the true (resp., incorrect) prior distribution π_0 (resp., $\bar{\pi}_0 = \delta_0 \otimes \nu_0$, where ν_0 is the distribution of X_0, i.e., the second marginal of π_0).

Now we can prove the asymptotic optimality of the approximate filter Π_t^0 in the following sense.

PROPOSITION 3.2. *Assume that in model* (22), *the conditions* $(\mathrm{A}_0), (\mathrm{A}_1^*), (\mathrm{A}_2^*)$, *and* (A_3^*) *are fulfilled. Then for every continuous* $\varphi \in \mathcal{M}_b(\mathbb{R}^d)$,

$$\lim_{t \to +\infty} \mathbb{E}[|\Pi_t(\varphi) - \Pi_t^0(\varphi)|^2] = 0. \tag{24}$$

Proof. The proof consists of showing that the conditions of Lemma 3.1 are all satisfied.

Condition (i). At first let us notice that, from Hasminski [3], condition (A_2^*) is a sufficient condition for the ergodicity of a Markov diffusion process X^0 associated with the first stochastic differential equation in (23). Under this condition the corresponding semigroup $(S_t^0, t \geq 0)$ admits a unique invariant measure μ^0 and $(S_t^0, t \geq 0)$ forgets any probability measure ν on $\mathbb{R}^d$ for μ^0, i.e.,

$$\nu S_t^0 \to \mu^0 \text{ weakly as } t \to +\infty, \tag{25}$$

where νS_t^0 denotes the law of X_t^0 when the distribution of X_0^0 is ν.

Now let $(S_t, t \geq 0)$ be the transition semigroup of the diffusion process $Z = (\zeta^{'}, X^{'})^{'}$ generated by (22). Under our assumptions, it is well known that $(S_t)_{t \geq 0}$ is a strongly continuous semigroup satisfying the Feller property. Moreover, from the discussion above about the consequences of (A_2^*) for the first equation in (23), it is straightforward to show that the measure $\mu = \delta_0 \otimes \mu^0$, where μ^0 is the unique invariant measure of $(S_t^0, t \geq 0)$, is an invariant measure of $(S_t, t \geq 0)$. The uniqueness of μ^0 implies the uniqueness of μ, and from (25), (H_1) can be obtained easily.

Condition (ii). Since the law $\pi_0 S_t$ of Z_t under $\mathbb{P}^0 = \mathbb{P}^{\bar{\pi}_0}$ coincides with the law $\delta_0 \otimes (\nu_0 S_t^0)$ on $\Omega_0 \times \Omega_1$, the condition $(\mathrm{H}_2(\bar{\pi}_0))$ is directly obtained from (25). It remains to verify that $(\mathrm{H}_2(\pi_0))$ is fulfilled. Since $(S_t, t \geq 0)$ admits a unique invariant measure, it suffices to show that the family of laws $\{\mathbb{P}_{Z_t},\, t \geq 0\}$, where $\mathbb{P}_{Z_t} = \pi_0 S_t$, is uniformly tight, i.e., for any $\varepsilon > 0$, there is a compact set $K \subset \mathbb{R}^n$ such that

$$\mathbb{P}([Z_t \in K]) \geq 1 - \varepsilon \text{ for all } t \geq 0.$$

But from (A_3^*), ζ converges to 0 in probability under $\mathbb{P} = \mathbb{P}^{\pi_0}$. Then, denoting $\mathbb{P}_{X_t}$ the distribution of X_t under $\mathbb{P}$, it suffices to prove that the family $\{\mathbb{P}_{X_t}, t \geq 0\}$ is uniformly tight. Let us introduce the new probability $\widetilde{\mathbb{P}}$ which is locally absolutely continuous with respect to $\mathbb{P}$ with the local Radon–Nikodým derivative L_t, where

$$\begin{aligned}
L_t &:= \exp\left\{-K_t - \frac{1}{2}\langle K \rangle_t\right\}, \\
K_t &:= \int_0^t b_\zeta(\zeta_s)^{'}\, dV_s + \int_0^t h_\zeta(\zeta_s)^{'}\, dW_s, \quad t \geq 0, \\
\langle K \rangle_t &:= \int_0^t |b_\zeta(\zeta_s)|^2\, ds + \int_0^t |h_\zeta(\zeta_s)|^2\, ds, \quad t \geq 0.
\end{aligned}$$

Then also $\mathbb{P}$ is locally absolutely continuous with respect to $\widetilde{\mathbb{P}}$ with the local Radon–Nikodým derivative L_t^{-1}. Then, for any $r > 0$, we have

$$\mathbb{P}([|X_t| > r]) = \widetilde{\mathbb{E}}[\mathbb{1}_{\{|X_t|>r\}} L_t^{-1}].$$

But under $\widetilde{\mathbb{P}}$, the process X satisfies the stochastic differential equation

$$dX_t = b(X_t)\,dt + d\widetilde{V}_t, \quad t \geq 0, X_0,$$

where $\widetilde{V}$ is a Brownian motion and X_0 has still distribution ν_0 and is independent of $\widetilde{V}$. Therefore, under $\widetilde{\mathbb{P}}$ the distribution of the random variable X_t coincides with $\nu_0 S_t^0$. Hence by the Cauchy–Schwarz inequality, for any $r > 0$ we get

$$\mathbb{P}([|X_t| > r]) \leq [\widetilde{\mathbb{P}}([|X_t| > r])]^{\frac{1}{2}} [\widetilde{\mathbb{E}}(L_t^{-2})]^{\frac{1}{2}},$$

or, equivalently,

$$\mathbb{P}([|X_t| > r]) \leq \left[\nu_0 S_t^0(\{x \in \mathbb{R}^d : |x| > r\})\right]^{\frac{1}{2}} \left[\mathbb{E}(L_t^{-1})\right]^{\frac{1}{2}}. \tag{26}$$

But, taking into account the independence of ζ and (V, W) under $\mathbb{P}$, conditioning on ζ, it is easy to check that

$$\mathbb{E}(L_t^{-1}) = \mathbb{E}(\exp\langle K\rangle_t).$$

Then, making use of assumptions $(\mathrm{A}_0), (\mathrm{A}_1^*)$, and (A_3^*), it is readily seen that there exists some positive constant C such that

$$\mathbb{E}(L_t^{-1}) \leq C < +\infty. \tag{27}$$

Moreover, assumption (A_2^*) implies that the family of distributions $\{\nu_0 S_t^0, t \geq 0\}$ is uniformly tight. Therefore, for given $\varepsilon > 0$ and $C > 0$, there exists an $r > 0$ such that

$$\nu_0 S_t^0(\{x \in \mathbb{R}^d : |x| > r\}) \leq \frac{\varepsilon^2}{C} \text{ for all } t \geq 0.$$

Hence, due to (26) and (27), for this r it follows that

$$\mathbb{P}([|X_t| > r]) \leq \varepsilon \text{ for all } t \geq 0.$$

This means that the family of laws $\{\mathbb{P}_{X_t}, t \geq 0\}$ is uniformly tight and we can conclude that $(\mathrm{H}_2(\pi_0))$ is fulfilled.

Condition (iii). Recall that R^{π_0} and $R^{\bar{\pi}_0}$ are the marginal distributions on Ω_2 of $\mathbb{P} = \mathbb{P}^{\pi_0}$ and $\mathbb{P}^0 = \mathbb{P}^{\bar{\pi}_0}$, respectively. The probability $\widetilde{\mathbb{P}}$, already used above, has the same marginal distribution on $\Omega_1 \times \Omega_2$ as the probability $\mathbb{P}^0$, and hence its marginal distribution on Ω_2 is nothing but $R^{\bar{\pi}_0}$. Therefore, to prove that R^{π_0} is absolutely continuous with respect to $R^{\bar{\pi}_0}$, it is sufficient to show that, in fact, $\mathbb{P}$ is absolutely continuous with respect to $\widetilde{\mathbb{P}}$ on $(\Omega, \mathcal{F})$. Thanks to [5, Proposition III.3.5], to conclude it is enough to show that the local density process L_t^{-1} of $\mathbb{P}$ with respect to $\widetilde{\mathbb{P}}$ is a square-integrable martingale with respect to $\widetilde{\mathbb{P}}$, i.e.,

$$\sup_t \widetilde{\mathbb{E}}(L_t^{-2}) < +\infty,$$

or, equivalently,

$$\sup_t \mathbb{E}(L_t^{-1}) < +\infty.$$

Actually, this is true because of the bound (27) obtained above, and so condition (iii) is fulfilled. □

Remark 3.3. It is interesting to emphasize that the constant c_0 appearing in the condition (A_3^*), which ensures the existence of an exponential moment for ζ_0, depends on the convergence rate λ of the flow ϕ and on the Lipschitz constants of b_ζ and h_ζ in (A_1^*). In particular, not surprisingly, it appears that the larger λ is, the smaller the constant c_0 is.

Remark 3.4. With slight modifications in the proof of Proposition 3.2, one can show that the same asymptotic result holds for any initial law of the form $\bar{\pi}_0 = \delta_0 \otimes \nu$ in place of $\bar{\pi}_0 = \delta_0 \otimes \nu_0$, provided that the incorrect prior distribution ν for X_0 dominates the true one ν_0, i.e., $\nu_0 \ll \nu$. So for two different initial conditions of this form the asymptotic behavior of the approximate filter is unchanged. In the particular case of the filtering model (23), this reduces to the fact that, under the assumptions "b Lipschitzian," "h continuous bounded," and (A_2^*), a filter initialized with an erroneous prior distribution ν such that $\nu_0 \ll \nu$ has the same asymptotic behavior as the optimal filter (initialized with ν_0). Actually, that property in model (23) is a direct consequence of [9, Theorem 3.2, Remark 3.3].

Acknowledgments. We are very grateful to Marina Kleptsyna for valuable discussions concerning ergodic properties of diffusion processes and also to the reviewers and the editors for their comments, which helped to improve the presentation of this paper.

REFERENCES

[1] V. E. Beněs and I. Karatzas, *Estimation and control for linear, partially observable systems with non-Gaussian initial distribution*, Stochastic Process. Appl., 14 (1983), pp. 233–248.

[2] R. J. Elliott, *Stochastic Calculus and Applications*, Springer-Verlag, New York, 1982.

[3] R. Z. Hasminski, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, Rockville, MD, 1980.

[4] U. G. Haussmann and E. Pardoux, *A conditionally almost linear filtering problem with non-Gaussian initial condition*, Stochastics, 23 (1988), pp. 241–275.

[5] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin, 1987.

[6] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[7] A. Le Breton and M.-C. Roubaud, *Asymptotic Optimality of Approximate Filters in Stochastic Systems with Colored Noises*, Rapport de Recherche 3801, INRIA, Le Chesnay, France, 1999; also available online from http://www-sop.inria.fr/rapports/sophia/RR-3081.html.

[8] A. M. Makowski, *Filtering formulae for partially observed linear systems with non-Gaussian initial conditions*, Stochastics, 16 (1986), pp. 1–24.

[9] D. Ocone and E. Pardoux, *Asymptotic stability of the optimal filter with respect to its initial condition*, SIAM J. Control Optim., 34 (1996), pp. 226–243.

[10] E. Pardoux, *Filtrage non linéaire et équations aux dérivées partielles stochastiques associées*, in Ecole d'été de Probabilités de Saint–Flour XIX, Lecture Notes in Math. 1464, P. L. Hennequin, ed., Springer-Verlag, Berlin, 1991, pp. 67–163.

[11] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 11 (1969), pp. 230–243.

SIAM J. Control Optim.
Vol. 39, No. 3, pp. 928–949

MARKOV PROPERTY AND ERGODICITY OF THE NONLINEAR FILTER[*]

A. G. BHATT[†], A. BUDHIRAJA[‡], AND R. L. KARANDIKAR[†]

Abstract. In this paper we first prove, under quite general conditions, that the nonlinear filter and the pair (signal, filter) are Feller–Markov processes. The state space of the signal is allowed to be nonlocally compact and the observation function h can be unbounded. Our proofs, in contrast to those of Kunita [*J. Multivariate Anal.,* 1 (1971), pp. 365–393; *Spatial Stochastic Processes,* Birkhäuser, 1991, pp. 233–256] and Stettner [*Stochastic Differential Equations,* Springer-Verlag, 1989, pp. 279–292], do not depend upon the uniqueness of the solutions to the filtering equations. We then obtain conditions for existence and uniqueness of invariant measures for the nonlinear filter and the pair process. These results extend those of Kunita and Stettner, which hold for locally compact state space and bounded h, to our general framework. Finally we show that the recent results of Ocone and Pardoux [*SIAM J. Control Optim.,* 34 (1996), pp. 226–243] on asymptotic stability of the nonlinear filter, which use the Kunita–Stettner setup, hold for the general situation considered in this paper.

Key words. nonlinear filtering, invariant measures, asymptotic stability, measure valued processes

AMS subject classifications. 60G35, 60J05, 60H15

PII. S0363012999357707

1. Introduction.

Kunita's pioneering paper [9] showed the way for analyzing asymptotic behavior of the nonlinear filter. Kunita showed for the case of compact state space valued Markov signal process that the nonlinear filter itself is a Markov process. This proof was based on the fact that the filter is the unique solution to the Kushner–Stratonovich equation (or the Fujisaki–Kallianpur–Kunita equation). Kunita also showed that if the signal is ergodic, then so is the filter. These results were extended to the case of locally compact state space (with a bounded observation function h) by Kunita [10] and Stettner [12]. The proofs of these results also crucially used the uniqueness of the solution to the filtering equation.

In this article, we first show that in the signal and noise independent case (which is also the case treated in the above-stated papers), when the signal is a Markov process, then so is the filter. This proof does not depend upon uniqueness of filtering equations. It just uses the Bayes formula to explicitly evaluate various conditional expectations and deduce that the filter is Markov. Using recent results in Bhatt, Kallianpur, and Karandikar [2], we deduce that under fairly general conditions on h, the filter is a Feller–Markov process. We also show that under the same conditions, the pair process (signal, filter) is a Feller–Markov process. Once this is done, we can obtain results of Kunita [10], Stettner [12], and Ocone and Pardoux [11] on ergodicity and asymptotic stability of the filter in our framework, where the state space is a

[*] Received by the editors June 1, 1999; accepted for publication (in revised form) June 26, 2000; published electronically October 20, 2000.
http://www.siam.org/journals/sicon/39-3/35770.html

[†] Indian Statistical Institute, Statistical Math Unit, 7 S.J.S. Sansanwal Marg, New Delhi, 110016, India. The research of these authors was supported by the Department of Science and Technology, Government of India.

[‡] Department of Statistics, University of North Carolina, Chapel Hill, NC 27599 (budhiraj@email.unc.edu). The research of this author was supported in part by the NSF grant DMI 9812857 and the University of Notre Dame Faculty Research Program.

complete separable metric space and the function h is allowed to be unbounded and minimal conditions are imposed on it.

Let us begin with listing some common notation used in this article. For a complete separable metric space S, let $C(S)$ be the class of continuous functions on S, $C_b(S)$ be the class of bounded continuous functions on S, $\mathcal{B}(S)$ be the Borel σ-field on S, $\mathcal{P}(S)$ be the space of probability measures on $(S, \mathcal{B}(S))$ endowed with the weak convergence topology, $\mathcal{M}_+(S)$ be the class of positive finite measures on $(S, \mathcal{B}(S))$ with the weak convergence topology, $D([0,\infty), S)$ be the class of functions, which are right continuous and have left limits (r.c.l.l.), from $[0,\infty)$ into S with Skorokhod topology, and $C([0,\infty), S)$ be the class of continuous functions from $[0,\infty)$ into S with topology of uniform convergence on compact subsets of $[0,\infty)$.

2. Preliminaries.

Consider the nonlinear filtering model

$$Y_t = \int_0^t h(X_u)du + W_t, \tag{2.1}$$

where (X_t) is the signal process, taking values in a complete separable metric space E; $h : E \to \mathbb{R}^d$ is a continuous mapping; and (W_t) is an $\mathbb{R}^d$-valued standard Wiener process, assumed to be independent of (X_t). (Y_t) is the observation process. (X_t), (W_t) are defined on a probability space $(\Omega, \mathcal{F}, P)$.

The object of interest in filtering theory is the conditional distribution π_t of X_t given the observations up until time t, i.e., given $\sigma(Y_u; u \le t)$.

We are going to assume that the signal process (X_t) is a Markov process with transition probability function $p(s,x,t,B)$: for $0 \le s < t$, $x \in E$, $B \in \mathcal{B}(E)$,

$$P(X_t \in B | \sigma(X_u : u \le s)) = p(s, X_s, t, B) \text{ almost surely (a.s.).} \tag{2.2}$$

The initial distribution of (X_t) will be denoted by γ, i.e.,

$$\gamma = P \circ (X_0)^{-1}. \tag{2.3}$$

Let $\xi_t(\cdot)$ be the coordinate process on $\mathcal{D} \doteq D([0,\infty), E)$, i.e., $\xi_t(\theta) \doteq \theta(t)$ for $\theta \in \mathcal{D}$. We assume that for all $(s,x) \in [0,\infty) \times E$ there exists a probability measure $P_{s,x}$ on $\mathcal{D}$ such that for $0 \le s < t < \infty$ and $U \in \mathcal{B}(E)$,

$$P_{s,x}(\xi_t \in U | \sigma(\xi_u : u \le s)) = p(s, \xi_s, t, U) \text{ a.s. } P_{s,x} \tag{2.4}$$

and

$$P_{s,x}(\xi_u = x, 0 \le u \le s) = 1. \tag{2.5}$$

Further, we assume that the mapping

$$(s,x) \longrightarrow P_{s,x} \text{ is continuous.} \tag{2.6}$$

As usual, for $\nu \in \mathcal{P}(E)$ and $B \in \mathcal{B}(\mathcal{D})$, let

$$P_{s,\nu}(B) \doteq \int P_{s,x}(B)\nu(dx), \quad B \in \mathcal{B}(\mathcal{D}).$$

Note that

$$P \circ (X)^{-1} = P_{0,\gamma}. \tag{2.7}$$

The Markov property implies that (T_t) defined below is a semigroup:

$$(T_t f)(s,x) \doteq \int f(s+t,y)p(s,x,s+t,dy), \quad (s,x) \in [0,\infty) \times E \tag{2.8}$$

for $f \in C_b([0,\infty)\times E)$, $0 \le t < \infty$. The assumption (2.6) implies that (T_t) is a Feller semigroup, i.e.,

$$T_t\left(C_b([0,\infty)\times E)\right) \subset C_b([0,\infty)\times E). \tag{2.9}$$

It is well known (and easy to verify) that (2.9) implies that (X_t) is continuous in probability. As a consequence, the filter (π_t) admits a continuous version (see [3]).

In order to study the Markov properties of (π_t), we first present the following preliminary result.

3. Pathwise integration formula. We cite the following result from [8]. As we shall see in the next section, this formulation of a stochastic integral is very useful when dealing with a family of measures on a measurable space.

The result is Theorem 3 in [8]. The mapping $\mathcal{I}$ is explicitly constructed in that paper.

THEOREM 3.1. *There exists a measurable mapping* $\mathcal{I} : D([0,\infty),\mathbb{R}) \times D([0,\infty),\mathbb{R}) \to D([0,\infty),\mathbb{R})$ *with the property that if* (U_t) *is a semimartingale (assumed to have r.c.l.l. paths) on a probability space* $(\Omega',\mathcal{F}',P')$ *with respect to a filtration* $(\mathcal{F}'_t)$ *and if* (V_t) *is an r.c.l.l.* $(\mathcal{F}'_t)$*-adapted process, then*

$$Z_t(w') := \mathcal{I}(V.(w'),U.(w'))(t), \quad w' \in \Omega',$$

is a version of the stochastic integral $\int_0^t V_- dU$*, i.e.,*

$$Z_t = \int_0^t V_{s-} dU_s \quad \textit{for all } t \textit{ a.s. } P'.$$

4. Markov properties of the filter. In this section we will study the Markov properties of the filter. The first main result is Theorem 4.5 in which we show that $\{\pi_t, \sigma\{Y_u : u \le t\}\}$ is a Markov process on $(\Omega,\mathcal{F},P)$. The second central result of this section is Theorem 4.8 in which we show that if $\overline{\pi}_t$ is a suboptimal filter defined via an incorrect initialization of the filtering process (to be made precise later in the section), then $\{(\overline{\pi}_t, X_t), \mathcal{F}_t\}$ is a Markov process on $(\Omega,\mathcal{F},P)$ with state space $\mathcal{P}(E)\times E$, where $\mathcal{F}_t \doteq \sigma\{(X_s,Y_s) : s \le t\}$. In particular, this result implies that $\{(\pi_t, X_t), \mathcal{F}_t\}$ is a Markov process on $(\Omega,\mathcal{F},P)$.

Let $\beta_t(\cdot)$ be the coordinate process on $\mathcal{C} \doteq C([0,\infty),\mathbb{R}^d)$, i.e., $\beta_t(\eta) \doteq \eta(t)$ for $\eta \in \mathcal{C}$. Let Q be the standard Wiener measure on $(\mathcal{C},\mathcal{B}(\mathcal{C}))$. Let

$$(\hat{\Omega},\hat{\mathcal{F}}) \doteq (\mathcal{D},\mathcal{B}(\mathcal{D})) \otimes (\mathcal{C},\mathcal{B}(\mathcal{C})),$$

and for $0 \le s < \infty, \nu \in P(E)$,

$$R_{s,\nu} \doteq P_{s,\nu} \otimes Q.$$

Let

$$Z_t(\theta,\eta) \doteq \sum_{j=1}^d \mathcal{I}(h^j(\xi(\theta)),\beta^j(\eta))(t), \tag{4.1}$$

where $h(x) \equiv (h^1(x),\ldots,h^d(x))$ and $\beta_t(\eta) \equiv (\beta^1_t(\eta),\ldots,\beta^d_t(\eta))$ for $x \in E$ and $\eta \in \mathcal{C}$.

Since (β_t), considered as a process on $(\hat{\Omega}, \hat{\mathcal{F}})$, is a Wiener process under $R_{s,\nu}$, it follows from Theorem 3.1 that

$$Z_t - Z_s = \sum_{j=1}^{d} \int_s^t h^j(\xi_u) d\beta_u^j \quad \text{a.s. } R_{s,\nu} \tag{4.2}$$

for every $0 \leq s < t < \infty$ and $\nu \in \mathcal{P}(E)$. The main thing to note is that we have been able to construct a common version of the stochastic integral appearing in (4.2) for the family of probability measures $\{R_{s,\nu}\}$.

For $0 \leq s < t < \infty$, let

$$q_{st}(\theta, \eta) \doteq \exp\left(Z_t(\theta, \eta) - Z_s(\theta, \eta) - \frac{1}{2} \sum_{j=1}^{d} \int_s^t (h^j(\xi_u(\theta))^2 du \right). \tag{4.3}$$

It is well known (and easy to verify) that $\{q_{st} : t \geq s\}$ is a $\{R_{s,\nu}\}$ martingale for every $\nu \in \mathcal{P}(E)$.

Let us note that for $0 \leq s < t$,

$$q_{0t}(\theta, \eta) = q_{0s}(\theta, \eta) q_{st}(\theta, \eta) \quad \text{for all } (\theta, \eta) \in \hat{\Omega}. \tag{4.4}$$

For $0 \leq s < t < \infty$, $\eta \in \mathcal{C}$, and $\nu \in \mathcal{M}_+(E)$, let $\Gamma_{st}(\nu, \cdot)(\eta) \in \mathcal{M}_+(E)$ and $\Lambda_{st}(\nu, \cdot)(\eta) \in \mathcal{P}(E)$ be defined as follows. For $B \in \mathcal{B}(E)$

$$\Gamma_{st}(\nu, B)(\eta) \doteq \int_E \int_{\mathcal{D}} 1_B(\xi_t(\theta)) q_{st}(\theta, \eta) dP_{s,x}(\theta) d\nu(x) \tag{4.5}$$

and

$$\Lambda_{st}(\nu, B)(\eta) \doteq \Gamma_{st}(\nu, B)(\eta) / \Gamma_{st}(\nu, E)(\eta). \tag{4.6}$$

The measure $\Gamma_{st}(\nu, \cdot)(\eta)$ will also be denoted by $\Gamma_{st}(\nu)(\eta)$ and likewise, $\Lambda_{st}(\nu, \cdot)(\eta)$ will be denoted by $\Lambda_{st}(\nu)(\eta)$. Since $\mathbb{E}_{R_{s,\nu}}[q_{st}] = 1$, it follows that

$$\mathbb{E}_Q[\Gamma_{st}(\nu, E)] = \nu(E). \tag{4.7}$$

Note that for $\nu \in \mathcal{M}_+(E)$, if $\hat{\nu}$ is defined by

$$\hat{\nu}(B) \doteq \frac{\nu(B)}{\nu(E)},$$

then

$$\Gamma_{st}(\nu, B)(\eta) = \nu(E) \Gamma_{st}(\hat{\nu}, B)(\eta) \tag{4.8}$$

and

$$\Lambda_{st}(\nu, B)(\eta) = \Lambda_{st}(\hat{\nu}, B)(\eta). \tag{4.9}$$

As a consequence of the Kallianpur–Striebel formula (see [1], [2], [6], [7]) it follows that

$$\pi_t(\omega)(B) \doteq \Lambda_{0t}(\gamma, B)(Y_{\cdot}(\omega)) \tag{4.10}$$

is a version of the filter $\mathbb{E}_P[1_B(X_t)|\sigma(Y_u : u \le t)]$. Furthermore, a.s. P, π_t has continuous paths (see [3]). Here is a technical result needed later.

THEOREM 4.1. *Fix $0 \le s < t$, $\nu \in \mathcal{P}(E)$. Let $\mathcal{F}^*$ be the completion of $\hat{\mathcal{F}}$ under $R_{0,\nu}$. Let $\mathcal{N}$ be the class of $R_{0,\nu}$ null sets in $\mathcal{F}^*$. Considering ξ, β as processes defined on $(\hat{\Omega}, \mathcal{F}^*, R_{0,\nu})$, let us define the following sub-σ fields of $\mathcal{F}^*$:*

$$\begin{aligned} \mathcal{G}_s^t &= \sigma(\sigma(\xi_u : s \le u \le t) \cup \mathcal{N}), \\ \mathcal{E}_0 &= \sigma(\sigma(\beta_u : u \ge 0) \cup \mathcal{N}), \\ \mathcal{E}_1 &= \sigma(\mathcal{G}_0^s \cup \mathcal{E}_0), \\ \mathcal{E}_2 &= \sigma(\mathcal{G}_s^\infty \cup \mathcal{E}_0). \end{aligned} \tag{4.11}$$

Let g be a $\mathcal{E}_2$-measurable, $R_{0,\nu}$-integrable random variable. Then

$$\mathbb{E}_{R_{0,\nu}}[g|\mathcal{E}_1] = g_1 \;\; a.s.,$$

where

$$g_1(\theta, \eta) = \int_{\mathcal{D}} g(\theta_1, \eta) dP_{s,\xi_s(\theta)}(\theta_1). \tag{4.12}$$

Remark 4.1. In (4.11) above, when $t = \infty$, the right-hand side is to be interpreted as $\sigma(\sigma(\xi_u : s \le u < \infty) \cup \mathcal{N})$.

Proof. When g is $\mathcal{G}_s^\infty$-measurable bounded random variable, the result follows from the Markov property of the family $\{P_{s,x}\}$ and independence of β and ξ under $R_{0,\nu}$. When g is of the form

$$g = \sum_{j=1}^{k} g_j f_j, \tag{4.13}$$

where g_j are $\mathcal{G}_s^\infty$-measurable bounded functions and f_j are $\mathcal{E}_0$-measurable bounded functions, the result follows from the preceding observation. Since the class of functions as in (4.13) forms an algebra that generates the σ field $\mathcal{E}_2$, the result follows. □

The next result connects $\{\Gamma_{st}\}$ with each other and is a key step in the proof of the Markov property.

THEOREM 4.2. *Fix $0 \le s < t < \infty$, $\nu \in \mathcal{P}(E)$. Then*

$$\Gamma_{0t}(\nu, B)(\eta) = \Gamma_{st}(\Gamma_{0s}(\nu)(\eta), B)(\eta) \;\; for\ all\ B \in \mathcal{B}(E), \;\; \eta - \;\; a.s.\ [Q] \tag{4.14}$$

and

$$\Lambda_{0t}(\nu, B)(\eta) = \Lambda_{st}(\Lambda_{0s}(\nu)(\eta), B)(\eta) \;\; for\ all\ B \in \mathcal{B}(E), \;\; \eta - \;\; a.s.\ [Q]. \tag{4.15}$$

Proof. Fix $A \in \mathcal{B}(\mathcal{C}), B \in \mathcal{B}(E)$. Let

$$G(A, B) \doteq \int_{\mathcal{C}} \Gamma_{0t}(\nu, B)(\eta) 1_A(\eta) dQ(\eta).$$

Using Fubini's theorem, the definition of Γ_{st}, and the relation (4.4), it follows that

$$G(A, B) = \int_E \int_{\hat{\Omega}} 1_A(\eta) q_{0s}(\theta, \eta) g(\theta, \eta) dR_{0,x}(\theta, \eta) d\nu(x), \tag{4.16}$$

where

$$g(\theta, \eta) = 1_B(\xi_t(\theta)) q_{st}(\theta, \eta).$$

Let $\mathcal{E}_1, \mathcal{E}_2$ be as in Theorem 4.1. Note that $1_A(\eta) q_{0s}(\theta, \eta)$ is $\mathcal{E}_1$-measurable and g is $\mathcal{E}_2$-measurable. It follows from Theorem 4.1 that

$$\mathbb{E}_{R_{0,x}}[g|\mathcal{E}_1](\theta, \eta) = f(\xi_s(\theta), \eta) \quad \text{a.s. } (\theta, \eta) \ [R_{0,x}], \tag{4.17}$$

where

$$\begin{aligned} f(x, \eta) &\doteq \int_{\mathcal{D}} g(\theta_1, \eta) dP_{s,x}(\theta_1) \\ &= \int_{\mathcal{D}} 1_B(\xi_t(\theta_1)) q_{st}(\theta_1, \eta) dP_{s,x}(\theta_1). \end{aligned}$$

Before proceeding, let us note that for $\nu \in P(E)$

$$\int_E f(y, \eta) d\nu(y) = \Gamma_{st}(\nu, B)(\eta). \tag{4.18}$$

Using (4.16), (4.17), and the fact that $1_A(\eta) q_{0s}(\theta, \eta)$ is $\mathcal{E}_1$-measurable, it follows that

$$G(A, B) = \int_E \int_{\hat{\Omega}} 1_A(\eta) q_{0s}(\theta, \eta) f(\xi_s(\theta), \eta) dR_{0,x}(\theta, \eta) d\nu(x).$$

Now applying the definition of Γ_{0s}, we have that

$$G(A, B) = \int_{\mathcal{C}} 1_A(\eta) \left[\int_E f(y, \eta) \Gamma_{0s}(\nu, dy)(\eta) \right] dQ(\eta)$$

and as a consequence on using (4.18) we have that

$$G(A, B) = \int_{\mathcal{C}} 1_A(\eta) \Gamma_{st}(\Gamma_{0s}(\nu)(\eta), B)(\eta) dQ(\eta). \tag{4.19}$$

Since (4.19) holds for all $A \in \mathcal{B}(\mathcal{C})$ and the σ field $\mathcal{B}(E)$ is countably generated, the relation (4.14) follows. The identity (4.15) is a consequence of (4.14) and (4.8)–(4.9). □

We are now in a position to prove that $\{\Gamma_{0t} : t \geq 0\}$, $\{\Lambda_{0t} : t \geq 0\}$ are Markov processes on $(\mathcal{C}, \mathcal{B}(\mathcal{C}), Q)$ with state spaces $\mathcal{M}_+(E)$ and $\mathcal{P}(E)$, respectively. Let $\tilde{\mathcal{F}}$ be the Q-completion of $\mathcal{B}(\mathcal{C})$ and $\tilde{\mathcal{N}}$ be the class of Q null sets in $\tilde{\mathcal{F}}$. For $0 \leq s \leq t \leq \infty$, let $\mathcal{A}_s^t$ be the sub-σ fields of $\tilde{\mathcal{F}}$ defined by

$$\mathcal{A}_s^t = \sigma(\sigma(\beta_u - \beta_s : s \leq u \leq t) \cup \tilde{\mathcal{N}}). \tag{4.20}$$

Here and in what follows, we will consider β and ξ as processes on $\hat{\Omega}$. It is easy to see that

$$\Gamma_{0t}(\nu), \Lambda_{0t}(\nu) \text{ are } \mathcal{A}_0^t\text{-measurable} \tag{4.21}$$

and for $s < t$

$$\Gamma_{st}(\nu), \Lambda_{st}(\nu) \text{ are } \mathcal{A}_s^\infty\text{-measurable.} \tag{4.22}$$

These observations lead us to the following theorem.

THEOREM 4.3. *Let $\nu \in \mathcal{P}(E)$. Then $(\Gamma_{0t}(\nu), \mathcal{A}_0^t)$ and $(\Lambda_{0t}(\nu), \mathcal{A}_0^t)$ are Markov processes on $(\mathcal{C}, \tilde{\mathcal{F}}, Q)$. Furthermore, for fixed $0 \le s < t < \infty$ and real valued Borel measurable functions ψ and φ on $\mathcal{M}_+(E)$ and $\mathcal{P}(E)$, respectively, which satisfy*

$$\mathbb{E}_Q[|\psi(\Gamma_{ut}(\lambda))|] < \infty$$

and

$$\mathbb{E}_Q[|\varphi(\Lambda_{ut}(\lambda))|] < \infty$$

for all $0 \le u \le t$ and $\lambda \in \mathcal{M}_+(E)$ we have that

$$\mathbb{E}_Q[\psi(\Gamma_{0t}(\nu))|\mathcal{A}_0^s] = \psi_1(\Gamma_{0s}(\nu)), \tag{4.23}$$

$$\mathbb{E}_Q[\varphi(\Lambda_{0t}(\nu))|\mathcal{A}_0^s] = \varphi_1(\Lambda_{0s}(\nu)), \tag{4.24}$$

where

$$\psi_1(\lambda) \doteq \mathbb{E}_Q[\psi(\Gamma_{st}(\lambda))], \quad \lambda \in \mathcal{M}_+(E), \tag{4.25}$$

and

$$\varphi_1(\nu) \doteq \mathbb{E}_Q[\varphi(\Lambda_{st}(\nu))], \quad \nu \in \mathcal{P}(E). \tag{4.26}$$

Proof. We will prove only the result for the case where φ and ψ are bounded. The general case follows by the usual approximation arguments (using the observation (4.7)).

We have observed that

$$\psi(\Gamma_{0t}(\nu)) = \psi(\Gamma_{st}(\Gamma_{0s}(\nu)));$$

$\Gamma_{0s}(\nu)$ is $\mathcal{A}_0^s$-measurable and $\Gamma_{st}(\lambda)$ is $\mathcal{A}_s^\infty$-measurable (for all $\lambda \in \mathcal{M}_+(E)$). Also, $\mathcal{A}_0^s$ and $\mathcal{A}_s^\infty$ are independent under Q. This implies (4.23) with ψ_1 defined by (4.25).

For $\varphi \in C_b(\mathcal{P}(E))$ define $\psi \in C_b(\mathcal{M}_+(E))$ by

$$\psi(\lambda) \doteq \varphi\left(\frac{1}{\lambda(E)}\lambda\right).$$

Then, for $\nu \in \mathcal{P}(E)$

$$\varphi(\Lambda_{0t}(\nu)) = \psi(\Gamma_{0t}(\nu)).$$

As a result

$$\mathbb{E}_Q[\varphi(\Lambda_{0t}(\nu))|\mathcal{A}_0^s] = \psi_1(\Gamma_{0s}(\nu)), \tag{4.27}$$

where for $\lambda \in \mathcal{M}_+(E)$

$$\begin{aligned}\psi_1(\lambda) &= \mathbb{E}_Q[\psi(\Gamma_{st}(\lambda))] \\ &= \mathbb{E}_Q[\varphi(\Lambda_{st}(\lambda))].\end{aligned}$$

Given $\lambda \in \mathcal{M}_+(E)$, define $\hat{\lambda} \in \mathcal{P}(E)$ by

$$\hat{\lambda}(B) \doteq \frac{1}{\lambda(E)}\lambda(B), \quad B \in \mathcal{B}(E).$$

Then from (4.9) and the definition of ϕ_1 (see (4.26)), we have that

$$\begin{aligned}\phi_1(\hat{\lambda}) &= \mathbb{E}_Q[\varphi(\Lambda_{st}(\hat{\lambda}))]\\ &= \mathbb{E}_Q[\varphi(\Lambda_{st}(\lambda))]\\ &= \psi_1(\lambda).\end{aligned}$$

As a consequence,

$$\psi_1(\Gamma_{0s}(\nu)) = \phi_1(\Lambda_{0s}(\nu)).$$

This in view of (4.27) proves (4.24). □

As noted earlier, on $(\hat{\Omega}, \hat{\mathcal{F}}, R_{s,\nu})$, $\nu \in \mathcal{P}(E)$,

$$(q_{st})_{t\geq s} \text{ is a martingale with respect to } \sigma(\xi_u, \beta_u : u \leq t)$$

and hence that

$$\rho_{st} = \begin{cases} \Gamma_{st}(\nu, E), & t \geq s, \\ 1, & t \leq s, \end{cases} \tag{4.28}$$

is a martingale on $(\mathcal{C}, \mathcal{B}(\mathcal{C}), Q)$ with respect to $(\mathcal{A}_0^t)$. Let $Q_{s,\nu} \in P(\mathcal{C})$ be defined by

$$\frac{dQ_{s,\nu}}{dQ} = \rho_{st} \text{ on } \mathcal{A}_0^t. \tag{4.29}$$

From Girsanov's theorem, it follows that $Q_{0,\gamma}$ is the law of the observation process Y, i.e., $PoY^{-1} = Q_{0,\gamma}$ (see [6]).

We are going to prove that $\{\Gamma_{0t}(\nu)\}$, $\{\Lambda_{0t}(\nu)\}$ are Markov processes on $(\mathcal{C}, \mathcal{B}(\mathcal{C}), Q_{0,\nu})$ as well.

THEOREM 4.4. *Let $\nu \in \mathcal{P}(E)$. Then $(\Gamma_{0t}(\nu), \mathcal{A}_0^t)$ and $(\Lambda_{0t}(\nu), \mathcal{A}_0^t)$ are Markov processes on $(\mathcal{C}, \tilde{\mathcal{F}}, Q_{0,\nu})$. Furthermore, for $f \in C_b(\mathcal{M}_+(E)), g \in C_b(\mathcal{P}(E))$,*

$$\mathbb{E}_{Q_{0,\nu}}[f(\Gamma_{0t}(\nu))|\mathcal{A}_0^s] = f_1(\Gamma_{0s}(\nu)) \tag{4.30}$$

and

$$\mathbb{E}_{Q_{0,\nu}}[g(\Lambda_{0t}(\nu))|\mathcal{A}_0^s] = g_1(\Lambda_{0s}(\nu)), \tag{4.31}$$

where f_1, g_1 are defined as follows. For $\lambda \in \mathcal{M}_+(E)$, $\nu \in \mathcal{P}(E)$

$$f_1(\lambda) \doteq \mathbb{E}_{Q_{s,\hat{\lambda}}}[f(\Gamma_{st}(\lambda))], \tag{4.32}$$

where $\hat{\lambda}(A) \doteq \frac{1}{\lambda(E)}\lambda(A)$ and

$$g_1(\nu) \doteq \mathbb{E}_{Q_{s,\nu}}[g(\Lambda_{st}(\nu))]. \tag{4.33}$$

Proof. We will first prove (4.30). Fix $0 \leq s < t < \infty$, $\nu \in \mathcal{P}(E)$, and $A \in \mathcal{A}_0^s$. Then

$$\begin{aligned}\int_A f(\Gamma_{0t}(\nu))dQ_{0,\nu} &= \int_A f(\Gamma_{0t}(\nu))\Gamma_{0t}(\nu, E)dQ\\ &= \int_A \psi(\Gamma_{0t}(\nu))dQ\\ &= \int_A \psi_1(\Gamma_{0s}(\nu))dQ,\end{aligned}$$

where $\psi(\lambda) \doteq f(\lambda)\lambda(E)$ and ψ_1 is given by (4.25). Thus defining $f_1(\lambda) = \frac{\psi_1(\lambda)}{\lambda(E)}$, it follows that

$$\begin{aligned} \int_A f(\Gamma_{0t}(\nu)) dQ_{0,\nu} &= \int_A f_1(\Gamma_{0s}(\nu))\Gamma_{0s}(\nu, E) dQ \\ &= \int_A f_1(\Gamma_{0s}(\nu)) dQ_{0,\nu} \end{aligned}$$

and hence (4.30) holds. Further note that

$$\begin{aligned} f_1(\lambda) &= \frac{1}{\lambda(E)}\psi_1(\lambda) \\ &= \frac{1}{\lambda(E)} \int_{\mathcal{C}} f(\Gamma_{st}(\lambda))\Gamma_{st}(\lambda, E) dQ \\ &= \int_{\mathcal{C}} f(\Gamma_{st}(\lambda))\Gamma_{st}(\hat{\lambda}, E) dQ \\ &= \mathbb{E}_{Q_{s,\hat{\lambda}}}[f(\Gamma_{st}(\lambda))]. \end{aligned}$$

Thus, f_1 is given by (4.32).

Now we prove (4.31). Let $g \in C_b(P(E))$ be as in the statement of the theorem. Let $f \in C_b(\mathcal{M}_+(E))$ be defined as

$$f(\lambda) \doteq g\left(\frac{1}{\lambda(E)}\lambda\right)$$

so that $f(\Gamma_{0t}(\nu)) = g(\Lambda_{0t}(\nu))$. Thus,

$$\mathbb{E}_{Q_{0,\nu}}[g(\Lambda_{0t}(\nu)) | \mathcal{A}_0^s] = f_1(\Gamma_{0s}(\nu)),$$

where f_1 is given by (4.32). Note that

$$\begin{aligned} f_1(\lambda) &= \mathbb{E}_{Q_{s,\hat{\lambda}}}[f(\Gamma_{st}(\lambda))] \\ &= \mathbb{E}_{Q_{s,\hat{\lambda}}}\left[g\left(\frac{1}{\Gamma_{st}(\lambda, E)}\Gamma_{st}(\lambda)\right)\right] \\ &= \mathbb{E}_{Q_{s,\hat{\lambda}}}[g(\Lambda_{st}(\lambda))] \\ &= \mathbb{E}_{Q_{s,\hat{\lambda}}}[g(\Lambda_{st}(\hat{\lambda}))] \\ &= g_1(\hat{\lambda}), \end{aligned}$$

where g_1 is defined by (4.33). Thus we have in particular that

$$\mathbb{E}_{Q_{0,\nu}}[g(\Lambda_{0t}(\nu)) | \mathcal{A}_0^s] = f_1(\Gamma_{0s}(\nu)) = g_1(\Lambda_{0s}(\nu)). \qquad \square$$

An immediate consequence of the above result is that for bounded measurable function G on $\mathcal{P}(E)$,

$$(\mathcal{T}_{st}G)(\nu) \doteq \mathbb{E}_{Q_{s,\nu}}[G(\Lambda_{st}(\nu))]$$

defines a two parameter semigroup, i.e., for $0 \leq s < t < u$

$$\mathcal{T}_{st} \circ \mathcal{T}_{tu} = \mathcal{T}_{su}.$$

Moreover, as noted earlier, $\pi_t(\omega) = \Lambda_{0t}(\gamma)(Y_{\cdot}(\omega))$ is a version of the filter $\mathbb{E}_P[X_t \in .|Y_u : u \leq t]$, while $Q_{0,\gamma}$ is the law of the observation process Y. Thus we have the following result.

THEOREM 4.5. *$\{\pi_t\}$ is a $\mathcal{P}(E)$-valued Markov process on $(\Omega, \mathcal{F}, P)$ with associated two parameter semigroup $\{\mathcal{T}_{st}\}$.*

Further, if the signal process (X_t) is a time homogeneous Markov process, then so is $\{\pi_t\}$ with

$$\mathcal{T}_{st} = \mathcal{T}_{0u}, \quad u = t - s.$$

Proof. The first part follows easily from the preceding theorem. For the second part, let us note that the law of $\{(\xi_{s+u}, \beta_{s+u} - \beta_s) : u \geq 0\}$ under $R_{s,\gamma}$ is $R_{0,\gamma}$. As a consequence, the law of $\Gamma_{st}(\gamma)$ under $R_{s,\gamma}$ is same as the law of $\Gamma_{0,(t-s)}(\gamma)$ under $R_{0,\gamma}$.

This yields the required result. □

Let $\nu \in \mathcal{P}(E)$. Define the process

$$\overline{\pi}_t(\omega) \doteq \Lambda_{0t}(\nu)(Y_{\cdot}(\omega)). \tag{4.34}$$

Then $(\overline{\pi}_t)$ considered as a process on $(\Omega, \mathcal{F}, P)$ represents a suboptimal filtering process which has been initiated at the incorrect initial law ν rather than the actual initial law γ. In the final part of this section we study the Markov properties of the process $(\overline{\pi}_t, X_t)$. The importance of Markov property and ergodicity of the pair process in the study of asymptotic robustness questions in nonlinear filtering has been pointed out in [4], [5]. Define the stochastic process (Ψ^{ν}_t) on $(\hat{\Omega}, \hat{\mathcal{F}})$ with values in $\mathcal{P}(E) \times E$ as follows:

$$\Psi^{\nu}_t(\theta, \eta) \doteq (\Lambda_{0t}(\nu)(\eta), \xi_t(\theta)).$$

Let $\mathcal{G}^t_s$ be as in Theorem 4.1, where $\mathcal{N}$ is the class of all $R_{0,\gamma}$ null sets. $\mathcal{G}^t_s$ are sub-σ fields of the completion $\mathcal{F}^*$ of $\mathcal{F}$ under the measure $R_{s,\gamma}$. We now define the following sub-σ fields of $\mathcal{F}^*$: for $0 \leq s \leq t \leq \infty$

$$\mathcal{H}^t_s \doteq \sigma(\sigma\{\beta_u - \beta_s : s \leq u \leq t\} \cup \mathcal{N}), \tag{4.35}$$

$$\mathcal{K}^t_s \doteq \sigma(\mathcal{G}^t_s \cup \mathcal{H}^t_s). \tag{4.36}$$

(See Remark 4.1 above for the case $t = \infty$.) The following theorem is essentially a consequence of Theorem 4.3.

THEOREM 4.6. *Let $\gamma, \nu \in \mathcal{P}(E)$. Then $(\Psi^{\nu}_t, \mathcal{K}^t_0)$ is a Markov process on $(\hat{\Omega}, \hat{F}, R_{0,\gamma})$. Furthermore, for $0 \leq s < t < \infty$ and a bounded measurable function f on $\mathcal{P}(E) \times E$ we have*

$$\mathbb{E}_{R_{0,\gamma}}[f(\Lambda_{0t}(\nu), \xi_t)q_{0t}|\mathcal{K}^s_0] = f_1(\Lambda_{0s}(\nu), \xi_s)q_{0s} \ \ a.s., \tag{4.37}$$

where $f_1 : \mathcal{P}(E) \times E \to \mathbb{R}$ is defined as follows: for $(\lambda, x) \in \mathcal{P}(E) \times E$

$$f_1(\lambda, x) \doteq \mathbb{E}_{R_{s,x}}[f(\Lambda_{st}(\lambda), \xi_t)q_{st}]. \tag{4.38}$$

Proof. The Markov property of $(\Psi^{\nu}_t, \mathcal{K}^t_0)$ is a direct consequence of Theorem 4.3 on noting that under $R_{0,\gamma}$ the σ fields $\mathcal{G}^s_0, \mathcal{H}^s_0$ are independent and $(\Lambda_{0t}(\nu), \mathcal{H}^t_0, R_{0,\gamma})$ and $(\xi_t, \mathcal{G}^t_0, R_{0,\gamma})$ are Markov processes. We now consider the second part of the theorem. Since $q_{0t} = q_{0s} q_{st}$ and q_{st} is $\mathcal{K}_s^\infty$-measurable, it suffices to prove that for a $\mathcal{K}_s^\infty$-measurable random variable Z with $\mathbb{E}_{R_{0,\gamma}}[\,|Z|\,] < \infty$

$$\mathbb{E}_{R_{0,\gamma}}[f(\Lambda_{0t}(\nu), \xi_t) Z | \mathcal{K}_0^s] = f_2(\Lambda_{0s}(\nu), \xi_s) \ \text{ a.s.,} \tag{4.39}$$

where $f_2 : \mathcal{P}(E) \times E \to \mathbb{R}$ is defined as follows: for $(\lambda, x) \in \mathcal{P}(E) \times E$

$$f_2(\lambda, x) \doteq \mathbb{E}_{R_{s,x}}[f(\Lambda_{st}(\lambda), \xi_t) Z]. \tag{4.40}$$

Using the usual approximation arguments, the proof of (4.39)–(4.40) can be reduced to the case when $f(\lambda, x) = g(\lambda)h(x)$ for bounded measurable functions g, h and $Z = UV$, with U being bounded $\mathcal{G}_s^\infty$-measurable and V being bounded $\mathcal{H}_s^\infty$-measurable.

Note that $\mathcal{H}_0^\infty = \sigma(\cup_t \mathcal{H}_0^t)$ and $\mathcal{G}_0^\infty = \sigma(\cup_t \mathcal{G}_0^t)$ are independent under $R_{0,\gamma}$. Also, $g(\Lambda_{0t}(\nu))V$ is $\mathcal{H}_0^\infty$-measurable, $\mathcal{H}_s^\infty \subseteq \mathcal{H}_0^\infty$, $h(\xi_t)U$ is $\mathcal{G}_0^\infty$-measurable and $\mathcal{G}_s^\infty \subseteq \mathcal{G}_0^\infty$. Further, $\mathcal{K}_0^s = \sigma(\mathcal{G}_0^s \cup \mathcal{H}_0^s)$. These observations imply that

$$\mathbb{E}_{R_{0,\gamma}}[g(\Lambda_{0t}(\nu)) V h(\xi_t) U | \mathcal{K}_0^s] = \mathbb{E}_{R_{0,\gamma}}[g(\Lambda_{0t}(\nu)) V | \mathcal{H}_0^s] \mathbb{E}_{R_{0,\gamma}}[h(\xi_t) U | \mathcal{G}_0^s]. \tag{4.41}$$

In turn, using $\Lambda_{0t}(\nu) = \Lambda_{st}(\Lambda_{0s}(\nu))$ and the independence of $\mathcal{H}_0^s$ and $\mathcal{H}_s^\infty$, it follows that

$$\mathbb{E}_{R_{0,\gamma}}[g(\Lambda_{0t}(\nu)) V | \mathcal{H}_0^s] = g_2(\Lambda_{0s}(\nu)) \tag{4.42}$$

with

$$\begin{aligned} g_2(\lambda) &= \mathbb{E}_{R_{s,x}}[g(\Lambda_{st}(\lambda)) V] \\ &= \int_{\mathcal{C}} g(\Lambda_{st}(\lambda)) V dQ, \end{aligned} \tag{4.43}$$

where the second display above follows on recalling that $R_{s,x} = P_{s,x} \otimes Q$. Using the Markov property of (ξ_t) it follows that

$$\mathbb{E}_{R_{0,\gamma}}[h(\xi_t) U | \mathcal{G}_0^s] = h_2(\xi_s) \tag{4.44}$$

with

$$h_2(x) = \mathbb{E}_{R_{s,x}}[h(\xi_t) U]. \tag{4.45}$$

The equations (4.41)–(4.45) imply that (4.39) holds with $f_2(\lambda, x) = g_2(\lambda) h_2(x)$. Using independence of $g(\Lambda_{st}(\lambda))V$ and $h(\xi_t)U$, it follows that (4.40) also holds. As noted at the beginning of the proof, this in turn implies the second part of the theorem, namely (4.37)–(4.38). □

Now for fixed $\nu \in \mathcal{P}(E)$ and $s \geq 0$ define $\hat{R}_{s,\nu}$ on $(\hat{\Omega}, \hat{\mathcal{F}})$ as follows:

$$\frac{d\hat{R}_{s,\nu}}{dR_{s,\nu}}(\theta, \eta) \doteq \ q_{st}(\theta, \eta) \quad \text{on } \mathcal{K}_0^t,\ t \geq s. \tag{4.46}$$

Recall the definition (4.29) of $Q_{s,\nu}$ and that $R_{s,\nu} \doteq P_{s,\nu} \otimes Q$. It now follows from (4.5) that

$$\hat{R}_{s,\nu}(\mathcal{D} \times B) = Q_{s,\nu}(B), \ \ B \in \mathcal{B}(\mathcal{C}). \tag{4.47}$$

The following result is the key step in the proof of the Markov property of $((X_t, \overline{\pi}_t), \mathcal{F}_t)$ on $(\Omega, \mathcal{F}, P)$.

THEOREM 4.7. *Let $\nu \in \mathcal{P}(E)$. Then $(\Psi_t^\nu, \mathcal{K}_0^t)$ is a Markov process on $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{R}_{0,\gamma})$.*

Proof. Fix $0 \leq s < t < \infty$. Let $f : \mathcal{P}(E) \times E \to \mathbb{R}$ be a bounded measurable function and let $A \in \mathcal{K}_0^s$. Then using (4.37) it follows that

$$\begin{aligned}\int_A f(\Lambda_{0t}(\nu), \xi_t) d\hat{R}_{0,\gamma} &= \int_A f(\Lambda_{0t}(\nu), \xi_t)\ q_{0t}\ dR_{0,\gamma} \\ &= \int_A f_1(\Lambda_{0s}(\nu), \xi_s)\ q_{0s}\ dR_{0,\gamma}.\end{aligned} \tag{4.48}$$

Hence

$$\mathbb{E}_{\hat{R}_{0,\gamma}}[f(\Lambda_{0t}(\nu), \xi_t) | \mathcal{K}_0^s] = f_1(\Lambda_{0s}(\nu), \xi_s), \tag{4.49}$$

where f_1 is given by (4.38). □

Remark 4.2. The relation defining f_1 can be recast as

$$f_1(\lambda, x) \doteq \mathbb{E}_{\hat{R}_{s,x}}[f(\Lambda_{st}(\lambda), \xi_t)]. \tag{4.50}$$

Now we come to the Markov property of the filter and signal.

THEOREM 4.8. *Fix $\nu \in \mathcal{P}(E)$. Let $\overline{\pi}_t$ be as in (4.34). Then $((\overline{\pi}_t, X_t), \mathcal{F}_t)$ is a $\mathcal{P}(E) \times E$-valued Markov process on $(\Omega, \mathcal{F}, P)$ with associated two parameter semigroup $\{\mathcal{S}_{s,t}\}_{0 \leq s < t < \infty}$ defined as follows. For a real bounded measurable map f on $\mathcal{P}(E) \times E$,*

$$(\mathcal{S}_{st} f)(\lambda, x) \doteq \mathbb{E}_{\hat{R}_{s,x}}[f(\Lambda_{st}(\lambda), \xi_t)]$$

for $(\lambda, x) \in \mathcal{P}(E) \times E$. Furthermore, if (X_t) is time homogeneous, then so is $(\overline{\pi}_t, X_t)$, i.e., $\mathcal{S}_{st} = \mathcal{S}_{0,t-s}$.

Proof. The Markov property follows on applying Theorem 4.7 and observing that the law of $\{(X_t, \overline{\pi}_t, Y_t) : t \geq 0\}$ under P is the same as the law of $\{(\xi_t, \Lambda_{0t}(\nu), \beta_t) : t \geq 0\}$ under $\hat{R}_{0,\gamma}$. The identities (4.49)–(4.50) imply the assertion made about the semigroup $\mathcal{S}_{st}$. □

5. Time homogeneous signal: Feller properties of the filter.

We will now examine the case of time homogeneous signal. In this case, we write

$$P_\nu = P_{0,\nu}, R_\nu = R_{0,\nu}, Q_\nu = Q_{0,\nu}, \hat{R}_\nu = \hat{R}_{0\nu},$$

$$\Gamma_t = \Gamma_{0t}, \Lambda_t = \Lambda_{0t}, \mathcal{T}_t = \mathcal{T}_{0t}, \mathcal{S}_t = \mathcal{S}_{0t}.$$

We will now show that under a suitable condition, $\{\mathcal{T}_t\}$ and $\{\mathcal{S}_t\}$ are Feller semigroups. The key step in the proof of the Feller property is a result on robustness of the filter from [2]. It is shown there that under a suitable condition on h, ν_n converging to ν implies that

$$Q_{\nu_n} \circ (\Lambda(\nu_n))^{-1} \to Q_\nu \circ (\Lambda(\nu))^{-1}.$$

This would immediately give the Feller property of $\mathcal{T}_t$. The following is essentially proved in [2]. We include here an outline of the proof as the exact result stated below is not given in [2]. This would enable us to conclude that $\mathcal{S}_t$ is a Feller semigroup.

THEOREM 5.1. *Let h be a continuous map from E to $\mathbb{R}^d$. Let $\{\nu_n\} \subset \mathcal{P}(E)$ be a sequence converging to ν weakly. Then the following hold.*

(a) *For all $t \geq 0$, $\Lambda_t(\nu_n) \to \Lambda_t(\nu)$ in Q probability.*

(b) *For all $t > 0$, $\Gamma_t(\nu_n, E) \to \Gamma_t(\nu, E)$ in $L^1(Q)$.*

Proof. For (a) note that continuity of h and continuity of the mapping $x \mapsto P_x$ imply that

$$\lim_{N\to\infty}\left[\sup_{x\in K} P_x\left(\int_0^T \|h(\xi_t)\|^2 1_{\{\|h(\xi_t)\|\geq N\}}dt > \varepsilon\right)\right] = 0 \tag{5.1}$$

for all $\varepsilon > 0$, $T < \infty$ and for all compact subsets $K \subset E$. The result now follows from Theorem 3.2 in [2].

Since $\nu_n \to \nu$ weakly as $n \to \infty$, in view of our assumption on $\{P_x\}$ we have that $P_{\nu_n} \to P_\nu$ weakly as $n \to \infty$. Now let $\{\tilde{X}_t^n\}$ and $\{\tilde{X}_t\}$ be processes with values in $\mathcal{D}$ defined on some probability space $(\overline{\Omega}, \overline{\mathcal{F}}, \overline{P})$ such that $\mathcal{L}(\tilde{X}_\cdot^n) = P_{\nu_n}$, $\mathcal{L}(\tilde{X}_\cdot) = P_\nu$, and $\tilde{X}_\cdot^n \to \tilde{X}_\cdot$ a.s. $\overline{P}$. Define

$$(\Omega_0, \mathcal{F}_0, R_0) \doteq (\overline{\Omega} \times \mathcal{C}, \overline{\mathcal{F}} \otimes \mathcal{B}(\mathcal{C}), \overline{P} \otimes Q)$$

and the processes $Z_\cdot^n$, $Z_\cdot$ on this space as

$$Z_t^n(\overline{\omega}, \eta) \doteq q_{0t}(\tilde{X}^n(\overline{\omega}), \eta),$$

$$Z_t(\overline{\omega}, \eta) \doteq q_{0t}(\tilde{X}(\overline{\omega}), \eta).$$

Then note that

$$\Gamma_t(\nu_n, B)(\eta) = \int_{\overline{\Omega}} 1_B(\tilde{X}^n(\overline{\omega})) Z_t^n(\overline{\omega}, \eta) d\overline{P}$$

and

$$\Gamma_t(\nu, B)(\eta) = \int_{\overline{\Omega}} 1_B(\tilde{X}(\overline{\omega})) Z_t(\overline{\omega}, \eta) d\overline{P}.$$

It is shown in [2] that $Z_t^n \to Z_t$ in $L^1(R_0)$ and this implies part (b). □

As a consequence of the above theorem we have the following results.

THEOREM 5.2. *Let h be continuous map from E to $\mathbb{R}^d$. Then $(\mathcal{T}_t)$ is a Feller semigroup.*

Proof. Let $\{\nu_n\}$ be a sequence in $\mathcal{P}(E)$. Suppose that ν_n converges weakly to ν. Let G be a bounded and continuous real function on $\mathcal{P}(E)$. We need to show that

$$(\mathcal{T}_t G)(\nu_n) \to (\mathcal{T}_t G)(\nu).$$

This follows from Theorem 5.1 upon observing that

$$\begin{aligned}(\mathcal{T}_t G)(\nu_n) &= \mathbb{E}_{Q_{\nu_n}}(G(\Lambda_t(\nu_n))) \\ &= \mathbb{E}_Q(G(\Lambda_t(\nu_n))\Gamma_t(\nu_n, E))\end{aligned} \tag{5.2}$$

and similarly,

$$(\mathcal{T}_t G)(\nu) = \mathbb{E}_Q(G(\Lambda_t(\nu))\Gamma_t(\nu, E)). \qquad \square \tag{5.3}$$

In the following theorem we prove the Feller property of $(\overline{\pi}_t, X_t)$.

THEOREM 5.3. *Let h be continuous map from E to $\mathbb{R}^d$. Then $\{\mathcal{S}_t\}$ is a Feller semigroup.*

Proof. Let F be a real bounded and continuous function on $\mathcal{P}(E) \times E$. Let $\{(\nu_n, x_n)\}$ be a sequence in $\mathcal{P}(E) \times E$ such that $(\nu_n, x_n) \to (\nu, x)$. We need to show that

$$(\mathcal{S}_t F)(\nu_n, x_n) \to (\mathcal{S}_t F)(\nu, x).$$

Now observe that

$$\begin{aligned}(\mathcal{S}_t F)(\nu_n, x_n) &= \mathbb{E}_{\hat{R}_{x_n}}(F(\Lambda_t(\nu_n), \xi_t)) \\ &= \mathbb{E}_{R_{x_n}}(F(\Lambda_t(\nu_n), \xi_t) q_{0t}) \\ &= \mathbb{E}_{R_0}(F(\Lambda_t(\nu_n), \tilde{X}^n_t) Z^n_t), \qquad (5.4)\end{aligned}$$

where $\tilde{X}^n_\cdot, Z^n, R_0$ are as in the proof of Theorem 5.1 with ν_n replaced by δ_{x_n}. Now the result follows once more by an application of Theorem 5.1 and recalling that $\tilde{X}^n_\cdot \to \tilde{X}_\cdot$ a.s. as $n \to \infty$. □

In the final result of this section we study the connection between the invariant measures for the semigroups $\mathcal{T}_t$ and $\mathcal{S}_t$. For a real measurable function φ on a Polish space S and a measure ν on $(S, \mathcal{B}(S))$ we will write $\int_S \varphi(x) d\nu(x)$ as $\nu(\varphi)$ when the former makes sense.

PROPOSITION 5.4. *Let $\overline{M}$ be an $(\mathcal{S}_t)$ invariant probability measure. Define the probability measures μ on $(E, \mathcal{B}(E))$ and M on $(\mathcal{P}(E), \mathcal{B}(\mathcal{P}(E)))$ as follows.*

$$\mu(B) \doteq \overline{M}(\mathcal{P}(E) \times B), \quad B \in \mathcal{B}(E),$$

and

$$M(C) \doteq \overline{M}(C \times E), \quad C \in \mathcal{B}(\mathcal{P}(E)).$$

Then μ is a (T_t) invariant measure. Furthermore, if for all real bounded measurable functions f on E and F on $\mathcal{P}(E)$

$$\int_{\mathcal{P}(E) \times E} F(\nu) f(x) \overline{M}(d\nu, dx) = \int_{\mathcal{P}(E)} \nu(f) F(\nu) M(d\nu), \qquad (5.5)$$

then M is a $(\mathcal{T}_t)$ invariant measure.

Proof. Let f be a bounded measurable function on E. Define $\tilde{f} : \mathcal{P}(E) \times E \to \mathbb{R}$ as $\tilde{f}(\nu, x) \doteq f(x)$. Then

$$\begin{aligned}\int_E (T_t f)(x) \mu(dx) &= \int_{\mathcal{P}(E) \times E} (T_t f)(x) \overline{M}(d\nu, dx) \\ &= \int_{\mathcal{P}(E) \times E} \mathbb{E}_{\hat{R}_x}(\tilde{f}(\Lambda_{0t}(\nu), \xi_t)) \overline{M}(d\nu, dx) \\ &= \int_{\mathcal{P}(E) \times E} (S_t \tilde{f})(\nu, x) \overline{M}(d\nu, dx) \\ &= \int_{\mathcal{P}(E) \times E} \tilde{f}(\nu, x) \overline{M}(d\nu, dx) \\ &= \int_E f(x) \mu(dx).\end{aligned}$$

This proves that μ is (T_t) invariant.

Next let G be a real bounded measurable function on $\mathcal{P}(E)$ and suppose that (5.5) holds. Then for all real bounded measurable functions Φ on $\mathcal{P}(E) \times E$,

$$\int_{\mathcal{P}(E)\times E} \Phi(\nu, x)\overline{M}(d\nu, dx) = \int_{\mathcal{P}(E)} \left(\int_E \Phi(\nu, x)\nu(dx)\right) M(d\nu). \tag{5.6}$$

Define $\tilde{G} : \mathcal{P}(E) \times E \to \mathbb{R}$ as $\tilde{G}(\nu, x) \doteq G(\nu)$. Then, recalling (4.47)

$$\begin{aligned}
\int_{\mathcal{P}(E)} (\mathcal{T}_t G)(\nu) M(d\nu) &= \int_{\mathcal{P}(E)} \mathbb{E}_{Q_\nu}(G(\Lambda_t(\nu)))M(d\nu)\\
&= \int_{\mathcal{P}(E)} \mathbb{E}_{\hat{R}_\nu}(G(\Lambda_t(\nu)))M(d\nu)\\
&= \int_{\mathcal{P}(E)} \left(\int_E \mathbb{E}_{\hat{R}_x}(G(\Lambda_t(\nu, \cdot)))\nu(dx)\right) M(d\nu)\\
&= \int_{\mathcal{P}(E)} \left(\int_E (\mathcal{S}_t\tilde{G})(\nu, x)\nu(dx)\right) M(d\nu)\\
&= \int_{\mathcal{P}(E)\times E} (\mathcal{S}_t\tilde{G})(\nu, x)\overline{M}(d\nu, dx)\\
&= \int_{\mathcal{P}(E)\times E} \tilde{G}(\nu, x)\overline{M}(d\nu, dx)\\
&= \int_{\mathcal{P}(E)} G(\nu)M(d\nu).
\end{aligned}$$

This proves that M is $(\mathcal{T}_t)$ invariant. □

6. Ergodic properties of the filter. In this section we will use the notation of section 5. We will also assume throughout the rest of the paper that h is a continuous function. Thus, $\{\mathcal{T}_t\}$ and $\{\mathcal{S}_t\}$ are Feller semigroups. We will obtain conditions for uniqueness of invariant measures corresponding to these semigroups. These questions have been studied in great detail for compact [9] and locally compact [12, 10, 11] state spaces. The proofs of the results in the above papers rely on the uniqueness of the solution to the Fujisaki–Kallianpur–Kunita or Kushner–Stratonovich equations; in fact even the proof of the Feller–Markov property of the filtering process crucially uses the uniqueness of the solution to the above-mentioned equation. Nevertheless, using the methods and results of sections 4 and 5 many statements and theorems in the above papers carry over to our general setup with almost identical proofs. We will now present the main ergodicity results for the filtering model considered in this paper. In order to avoid tedious repetition, we will provide proofs only when they differ from the proofs of the analogous statements in [9, 12, 10, 11].

Let $C_c(E)$ be the class of all convex functions in $C_b(E)$. Following Stettner [12], denote for $\nu \in \mathcal{P}(E)$ and $A \in \mathcal{B}(\mathcal{P}(E))$

$$m_t^\nu(A) \doteq (\mathcal{T}_t\mathcal{I}_A)(\nu) = E_{Q_\nu}(\mathcal{I}_A(\Lambda_t(\nu, \cdot)))$$

and

$$M_t^\nu(A) \doteq \int_E (\mathcal{T}_t\mathcal{I}_A)(\delta_x)\nu(dx),$$

where $\mathcal{I}_A$ is the indicator function of the set A. We will now give alternative representations for m_t^ν and M_t^ν as the laws of certain filtering processes.

We begin with the following setup. Let μ be a (T_t) invariant measure. $\mathcal{D}_{\mathbb{R}} \equiv D((-\infty, \infty); E)$ will denote the space of r.c.l.l. functions from $(-\infty, \infty)$ into E with Skorokhod topology and $\mathcal{C}_{\mathbb{R}} \equiv C((-\infty, \infty); \mathbb{R}^d)$ will denote the space of continuous functions from $(-\infty, \infty)$ into $\mathbb{R}^d$ with topology of uniform convergence on compact subsets of $(-\infty, \infty)$. Let the coordinate processes on $\mathcal{D}_{\mathbb{R}}$ and $\mathcal{C}_{\mathbb{R}}$ be denoted once more by $(\xi_t(\cdot))$ and $(\beta_t(\cdot))$, respectively. Let $P_\mu^{(1)}$ be the unique measure on $(\mathcal{D}_{\mathbb{R}}, \mathcal{B}(\mathcal{D}_{\mathbb{R}}))$ which satisfies for $E_1, \ldots, E_n \in \mathcal{B}(\mathbb{R})$ and $-\infty < t_1 < t_2 \cdots < t_n < \infty$

$$\begin{aligned} &P_\mu^{(1)}(\xi_{t_1} \in E_1, \ldots, \xi_{t_n} \in E_n) \\ &\quad = \int_{E_1 \times \cdots \times E_n} \mu(dx_1) p(t_1, x_1, t_2, dx_2) \cdots p(t_{n-1}, x_{n-1}, t_n, dx_n). \end{aligned}$$

Now let $Q^{(1)}$ be a probability measure on $(C_{\mathbb{R}}, \mathcal{B}(C_{\mathbb{R}}))$ such that for $-\infty < t_0 < t_1 \cdots < t_n < \infty$,

$$\left(\frac{1}{\sqrt{t_1 - t_0}} (\beta_{t_1} - \beta_{t_0}), \ldots, \frac{1}{\sqrt{t_n - t_{n-1}}} (\beta_{t_n} - \beta_{t_{n-1}}) \right)$$

are independent $N(0, I_{d \times d})$.

Now let $\Omega^1 \doteq \mathcal{D}_{\mathbb{R}} \times C_{\mathbb{R}}$. Let $R_\mu^{(1)} = P_\mu^{(1)} \otimes Q^{(1)}$. In this section, we will consider the coordinate processes $(\xi_t), (\beta_t)$ to be defined on the product space $(\Omega^1, \mathcal{B}(\Omega^1), R_\mu^{(1)})$. Define the observation process

$$\alpha_t - \alpha_s \doteq \int_s^t h(\xi_u) du + \beta_t - \beta_s$$

and the observation σ fields

$$\mathcal{Z}_s^t \doteq \sigma(\alpha_v - \alpha_u; s \le u \le v \le t),$$

where $-\infty \le s < t \le \infty$. Further, for s, t such that $-\infty \le s < t \le \infty$, let $\mathcal{G}_s^t, \mathcal{H}_s^t$ be defined, respectively, by (4.11) and (4.35). The cases $s = -\infty$ and $t = \infty$ are treated as in Remark 4.1. Here, these are sub-σ fields of $\mathcal{B}(\Omega^1)$. Further, let $\mathcal{G}_{-\infty}^{-\infty}$ be defined by

$$\mathcal{G}_{-\infty}^{-\infty} = \cap_{-\infty < t < \infty} \mathcal{G}_{-\infty}^t.$$

Now define for $-\infty < s < t < \infty$,

$$\overline{\pi}_{s,t}^{(0)} \doteq \Lambda_{t-s}(\mu)(\alpha^s),$$

where $\alpha^s : \Omega^1 \to C([0, \infty); \mathbb{R}^d)$ is defined as $\alpha_u^s(\omega) \doteq \alpha_{s+u}(\omega) - \alpha_s(\omega)$. Also define

$$\overline{\pi}_{s,t}^{(1)} \doteq \Lambda_{t-s}(\delta_{\xi_s})(\alpha^s).$$

Observe that for the bounded and continuous function f on E

$$\overline{\pi}_{s,t}^{(0)}(f) = \mathbb{E}_{R_\mu^{(1)}}[f(\xi_t) | \mathcal{Z}_s^t]$$

and

$$\overline{\pi}_{s,t}^{(1)}(f) = \mathbb{E}_{R_\mu^{(1)}}[f(\xi_t) | \mathcal{Z}_s^t \vee \sigma(\xi_s)]$$

(for two σ fields $\mathcal{L}_1$ and $\mathcal{L}_2$, $\mathcal{L}_1 \vee \mathcal{L}_2 \doteq \sigma(\mathcal{L}1 \cup \mathcal{L}_2)$). Also note that for the real bounded measurable function F on $\mathcal{P}(E)$

$$\begin{aligned}
\mathbb{E}_{R_\mu^{(1)}}[F(\overline{\pi}_{s,t}^{(1)})] &= \mathbb{E}_{R_\mu^{(1)}}[F(\Lambda_{t-s}(\delta_{\xi_s})(\alpha^s))] \\
&= \int_E \mathbb{E}_{Q_x} F(\Lambda_{t-s}(\delta_x))\mu(dx) \\
&= \int_E (\mathcal{T}_{t-s}F)(\delta_x)\mu(dx) \qquad (6.1) \\
&= M_{t-s}^\mu(F). \qquad (6.2)
\end{aligned}$$

In a similar manner it is seen that

$$\mathbb{E}_{R_\mu^{(1)}}[F(\overline{\pi}_{s,t}^{(0)})] = m_{t-s}^\mu(F). \tag{6.3}$$

A straightforward application of martingale convergence theorem shows that as $s \to -\infty$, a.s. the measure $\overline{\pi}_{s,t}^{(0)}$ converges weakly to the measure $\overline{\pi}_t^{(0)}$ defined as follows: for a bounded and continuous function f on E

$$\overline{\pi}_t^{(0)}(f) \doteq \mathbb{E}_{R_\mu^{(1)}}[f(\xi_t)|\mathcal{Z}_{-\infty}^t].$$

Furthermore we have that (cf. Lemma 3.3 of Kunita [9])

$$\overline{\pi}_{s,t}^{(1)}(f) = \mathbb{E}_{R_\mu^{(1)}}[f(\xi_t)|\mathcal{Z}_{-\infty}^t \vee \mathcal{G}_{-\infty}^s]$$

and thus by the reverse martingale convergence theorem we have that as $s \to -\infty$, $\overline{\pi}_{s,t}^{(1)}$ converges weakly to the measure $\overline{\pi}_t^{(1)}$ defined as

$$\overline{\pi}_t^{(1)}(f) \doteq \mathbb{E}_{R_\mu^{(1)}}[f(\xi_t)|\mathcal{Z}_{-\infty}^t \vee \mathcal{G}_{-\infty}^{-\infty}].$$

Thus in view of (6.2) and (6.3) we have that M_s^μ and m_s^μ converge weakly as $s \to \infty$ to the law of $\overline{\pi}_t^{(0)}$, $\overline{\pi}_t^{(1)}$, respectively, which also shows that the laws of $\overline{\pi}_t^{(0)}$ $\overline{\pi}_t^{(1)}$ are independent of t. Denote these laws as m^μ and M^μ, respectively. Also note that since m^μ and M^μ are the limits of m_t^μ and M_t^μ as $t \to \infty$ and $(\mathcal{T}_t)$ is a Feller semigroup, both m^μ and M^μ have to be $(\mathcal{T}_t)$ invariant. We now recall the following definition from [9].

Definition 6.1. *A measure $\nu \in \mathcal{P}(E)$ is the barycenter of the measure $\Phi \in \mathcal{P}(\mathcal{P}(E))$ if and only if for every $\varphi \in C_b(E)$,*

$$\nu(\varphi) = \int_{\mathcal{P}(E)} \nu'(\varphi)\Phi(d\nu').$$

From Theorem 3.1 of Kunita [9] we have that both m^μ and M^μ have barycenters μ and if Φ is another $\mathcal{T}_t$ invariant measure with barycenter μ, then for all $F \in C_c(E)$,

$$m^\mu(F) \le \Phi(F) \le M^\mu(F). \tag{6.4}$$

Next recalling that m^μ and M^μ are the laws of $\overline{\pi}_t^{(0)}$ and $\overline{\pi}_t^{(1)}$, respectively, we have that m^μ equals M^μ if $\mathcal{G}_{-\infty}^{-\infty}$ is trivial. From (6.4) we observe that the equality of m^μ and M^μ implies that $\mathcal{T}_t$ admits a unique invariant measure with barycenter μ. We thus have the following result.

THEOREM 6.2. *Suppose that there is a unique* (T_t) *invariant measure* μ. *Then there is a unique* $\mathcal{T}_t$ *invariant measure if for all* $f \in C_b(E)$

$$\limsup_{t\to\infty} \int_E |T_t f(x) - \mu(f)|\mu(dx) = 0. \tag{6.5}$$

Proof. As pointed out in [9], (6.5) is equivalent to the condition that $\mathcal{G}_{-\infty}^{-\infty}$ is trivial. This proves the uniqueness of $(\mathcal{T}_t)$ invariant measure. □

We remark that our proof of the theorem above is different from the proof in [9]. The proof in [9] requires showing that $(\overline{\pi}_t^{(0)})$ and $(\overline{\pi}_t^{(1)})$ are Markov with semigroup $\mathcal{T}_t$ which is shown in [9] by appealing to the uniqueness of the solution to Kushner–Stratonovich equation. Although we don't need the Markov properties for our proof above, using the Feller property of $\mathcal{T}_t$ nevertheless, they are seen to hold as argued below.

PROPOSITION 6.3. *Let* μ *be a* (T_t) *invariant measure. For* $i = 0, 1$, $(\overline{\pi}_t^{(i)})$ *is a stationary Markov process on* $(\Omega^1, \mathcal{B}(\Omega^1), R_\mu^{(1)})$ *with semigroup* $\mathcal{T}_t$.

Proof. Let $g \in C_b(\mathcal{P}(E))$ and fix $-\infty < u < t < \infty$. Then from an application of martingale convergence theorem and the observation that a.s., $\overline{\pi}_{s,t}^{(0)}$ converges weakly to the measure $\overline{\pi}_t^{(0)}$ as $s \to -\infty$ we have that

$$\begin{aligned}
\mathbb{E}\left(g(\overline{\pi}_t^{(0)})|\mathcal{Z}_{-\infty}^u\right) &= \lim_{s\to-\infty} \mathbb{E}\left(g(\overline{\pi}_{s,t}^{(0)})|\mathcal{Z}_s^u\right)\\
&= \lim_{s\to-\infty} \mathbb{E}\left(g(\Lambda_{t-s}(\mu)(\alpha^s))|\sigma(\alpha_v^s; 0 \le v \le u-s)\right)\\
&= \lim_{s\to-\infty} g_1(\Lambda_{u-s}(\mu)(\alpha^s))\\
&= \lim_{s\to-\infty} g_1(\overline{\pi}_{s,u}^{(0)})\\
&= g_1(\overline{\pi}_u^{(0)}),
\end{aligned} \tag{6.6}$$

where all the limits in (6.6) are taken in probability and g_1 is the real valued function on $\mathcal{P}(E)$ defined in (4.33) with s there replaced by u, i.e.,

$$g_1(\nu) \doteq \mathbb{E}_{Q_\nu}[g(\Lambda_{t-u}(\nu))]. \tag{6.7}$$

Note that g_1 is independent of s and the last step in the above display follows on observing that the Feller property of $\mathcal{T}_t$ implies that g_1 is continuous. This proves the Markov property of $(\overline{\pi}_t^{(0)})$ and that the semigroup for this Markov process is $(\mathcal{T}_t)$. As is already seen, the law of $(\overline{\pi}_t^{(0)})$ does not depend on t and thus this Markov process is stationary. Finally consider the process $(\overline{\pi}_t^{(1)})$. Let g be as before. Then

$$\begin{aligned}
&\mathbb{E}_{R_\mu^{(1)}}[g(\overline{\pi}_t^{(1)})|\mathcal{Z}_{-\infty}^u \vee \mathcal{G}_{-\infty}^{-\infty}]\\
&\quad= \lim_{s\to-\infty} \mathbb{E}_{R_\mu^{(1)}}[g(\overline{\pi}_{s,t}^{(1)})|\mathcal{Z}_{-\infty}^u \vee \mathcal{G}_{-\infty}^s]\\
&\quad= \lim_{s\to-\infty} \mathbb{E}_{R_\mu^{(1)}}[g(\overline{\pi}_{s,t}^{(1)})|\mathcal{Z}_s^u \vee \mathcal{G}_{-\infty}^s]\\
&\quad= \lim_{s\to-\infty} \mathbb{E}_{R_\mu^{(1)}}[g(\Lambda_{t-s}(\delta_{\xi_s})(\alpha^s))|\sigma(\alpha_v^s : 0 \le v \le u-s) \vee \mathcal{G}_{-\infty}^s]\\
&\quad= \lim_{s\to-\infty} g_1(\Lambda_{u-s}(\delta_{\xi_s})(\alpha^s))\\
&\quad= \lim_{s\to-\infty} g_1(\overline{\pi}_{s,u}^{(1)})\\
&\quad= g_1(\overline{\pi}_u^{(1)}),
\end{aligned} \tag{6.8}$$

where g_1 is as in (6.7). Hence, $(\overline{\pi}_t^{(1)})$ is a Markov process with semigroup $(\mathcal{T}_t)$. The stationarity of this process follows as before. □

We now turn our attention to the semigroup $(\mathcal{S}_t)$. Let μ, as before, be a (T_t) invariant probability measure. Define the probability measures $\overline{m}_t^\mu$ and $\overline{M}_t^\mu$ on $(\mathcal{P}(E)\times E)$ as follows. For $A\in\mathcal{B}(\mathcal{P}(E))\otimes\mathcal{B}(E)$,

$$\overline{m}_t^\mu(A)=\int_E(\mathcal{S}_t\mathcal{I}_A)(\mu,x)\mu(dx)$$

and

$$\overline{M}_t^\mu(A)=\int_E(\mathcal{S}_t\mathcal{I}_A)(\delta_x,x)\mu(dx).$$

Observe that for real measurable bounded function F on $\mathcal{P}(E)\times E$ and $-\infty<s<t<\infty$,

$$\begin{aligned}\mathbb{E}_{R_\mu^{(1)}}[F(\overline{\pi}_{s,t}^{(1)},\xi_t)]&=\mathbb{E}_{R_\mu^{(1)}}[F(\Lambda_{t-s}(\delta_{\xi_s})(\alpha^s),\xi_t)]\\&=\int_E\mathbb{E}_{\hat{R}_x}F(\Lambda_{t-s}(\delta_x),\xi_{t-s})\mu(dx)\\&=\int_E(\mathcal{S}_{t-s}F)(\delta_x,x)\mu(dx)\\&=\overline{M}_{t-s}^\mu(F).\end{aligned}$$

In a similar way it is seen that

$$\mathbb{E}_{R_\mu^{(1)}}[F(\overline{\pi}_{s,t}^{(0)},\xi_t)]=\overline{m}_{t-s}^\mu(F).$$

Now the almost sure convergence, as $s\to-\infty$, of $(\overline{\pi}_{s,t}^{(0)},\xi_t)$ and $(\overline{\pi}_{s,t}^{(1)},\xi_t)$ to $(\overline{\pi}_t^{(0)},\xi_t)$ and $(\overline{\pi}_t^{(1)},\xi_t)$, respectively, implies that $\overline{m}_u^\mu$ and $\overline{M}_u^\mu$ converge weakly to the law of $(\overline{\pi}_t^{(0)},\xi_t)$ and $(\overline{\pi}_t^{(1)},\xi_t)$, respectively, and also that these laws don't depend on t. Denote the law of $(\overline{\pi}_t^{(0)},\xi_t)$ by $\overline{m}^\mu$ and the law of $(\overline{\pi}_t^{(1)},\xi_t)$ by $\overline{M}^\mu$. Once more the Feller property of the semigroup $(\mathcal{S}_t)$ implies that since $\overline{m}_t^\mu$ and $\overline{M}_t^\mu$ converge weakly to $\overline{m}^\mu$ and $\overline{M}^\mu$, respectively, these latter measures are $(\mathcal{S}_t)$ invariant. Next note that for $C\in\mathcal{B}(\mathcal{P}(E))$

$$\overline{m}^\mu(C\times E)=m^\mu(C)$$

and

$$\overline{M}^\mu(C\times E)=M^\mu(C).$$

Further note that for real bounded measurable functions f and F on E and $\mathcal{P}(E)$, respectively,

$$\begin{aligned}\int_{\mathcal{P}(E)\times E}F(\nu)f(x)\overline{m}^\mu(d\nu,dx)&=\mathbb{E}_{R_\mu^{(1)}}[(F(\overline{\pi}_t^{(0)})f(\xi_t))]\\&=\mathbb{E}_{R_\mu^{(1)}}[F(\overline{\pi}_t^{(0)})\mathbb{E}_{R_\mu^{(1)}}(f(\xi_t)|\mathcal{Z}_{-\infty}^t)]\\&=\int_E F(\nu)\nu(f)m^\mu(d\nu).\end{aligned}$$

Similarly,

$$\int_{\mathcal{P}(E)\times E} F(\nu)f(x)\overline{M}^{\mu}(d\nu,dx) = \int_E F(\nu)\nu(f)M^{\mu}(d\nu).$$

Thus $\overline{m}^{\mu}, \overline{M}^{\mu}$ are $(\mathcal{S}_t)$ invariant measures for which (5.5) holds. Also, in the class of $(\mathcal{S}_t)$ invariant probability measures for which (5.5) holds, $\overline{m}^{\mu}$ is the minimal and $\overline{M}^{\mu}$ is the maximal in the sense of Kunita (cf. Theorem 2.2 in [10]). Finally note that if $\mathcal{G}_{-\infty}^{-\infty}$ is trivial, $(\overline{\pi}_t^{(0)}, \xi_t) = (\overline{\pi}_t^{(1)}, \xi_t)$ a.s. and thus $\overline{m}^{\mu} = \overline{M}^{\mu}$. Hence the following theorem holds.

THEOREM 6.4. *Suppose that there is a unique* (T_t) *invariant measure* μ *and suppose that* (6.5) *holds for all* $f \in C_b(E)$. *Then there is a unique* $(\mathcal{S}_t)$ *invariant measure for which* (5.5) *holds.*

7. Asymptotic stability. In this section, we will study the asymptotic behavior of the filter when it is initialized at an incorrect initial condition. The results extend the results of Ocone and Pardoux [11] to the case of unbounded h and we also do away with the assumption in [11] that the state space is locally compact.

We will once more use the notations of section 5 and assume continuity of h. In particular, recall that the signal process X is an E-valued time homogeneous Markov process with a Feller semigroup T_t and initial distribution (law of X_0) γ. Further, we assume that T_t admits a unique invariant probability measure μ and that (6.5) holds. We have noted in the previous section that in this case the semigroup $\mathcal{T}_t$ is Feller and also admits a unique invariant probability measure. Here is the first result on asymptotic stability

THEOREM 7.1. *Suppose that* γ *satisfies*

$$\gamma T_t \to \mu. \tag{7.1}$$

Let

$$\pi_t(B)(\omega) = \Lambda_t(\gamma, B)(Y_{\cdot}(\omega)) \tag{7.2}$$

be the filter. Then the law of π_t *on* $(\Omega, \mathcal{F}, P)$ *converges to* M*—the invariant measure for the semigroup* $\mathcal{T}_t$.

Proof. This result can be proved exactly following the steps in the proof of Theorem 3 in Stettner [12]. Note that we have already proved the Feller property of $\mathcal{T}_t$ and hence it follows that if ν_n converges to ν, then $m_t^{\nu_n}$ converges in law to m_t^{ν}. This is a crucial step in the proof of Theorem 3 in [12]. The tightness of $\{m_t^{\gamma} : t \geq 0\}$ follows from (7.1) exactly as in Stettner. Also, we have already proved that there exists a unique $\mathcal{T}_t$ invariant probability measure M. □

The previous result gives the asymptotics of the filter π_t when the signal process is a purely nondeterministic ergodic Markov process.

Ocone and Pardoux [11] consider the behavior of the filter when the initial distribution of the signal is wrongly taken to be ν instead of γ. So whereas the correct filter is given by (7.2), the incorrectly initialized filter is given as

$$\bar{\pi}_t(B)(\omega) = \Lambda_t(\nu, B)(Y_{\cdot}(\omega). \tag{7.3}$$

In [11] it was shown that under appropriate conditions

$$\mathbb{E}_P[(\langle \pi_t, \varphi\rangle - \langle \bar{\pi}_t, \varphi\rangle)^2] \to 0 \tag{7.4}$$

as t tends to ∞. The paper [11] assumes that h is bounded and that E is locally compact. We will now show that (7.4) holds in our framework.

We need the following auxiliary result for this purpose.

THEOREM 7.2. *Let $\gamma^n, \nu^n \in \mathcal{P}(E)$ be such that*

$$\gamma^n \to \mu, \quad \nu^n \to \mu.$$

Let $T > 0$ be fixed. Then for a bounded continuous function φ on E,

$$\mathbb{E}_{Q_{\gamma^n}}[|\langle \Lambda_T(\gamma^n), \varphi\rangle - \langle \Lambda_T(\nu^n), \varphi\rangle|] \to 0.$$

Proof. In view of (4.28)–(4.29)

$$\mathbb{E}_{Q_{\gamma^n}}[|\langle \Lambda_T(\gamma^n), \varphi\rangle - \langle \Lambda_T(\nu^n), \varphi\rangle|] = \mathbb{E}_Q[|\langle \Lambda_T(\gamma^n), \varphi\rangle - \langle \Lambda_T(\nu^n), \varphi\rangle|\Gamma_T(\gamma^n, E)]. \tag{7.5}$$

By Theorem 5.1

$$|\langle \Lambda_T(\gamma^n), \varphi\rangle - \langle \Lambda_T(\nu^n), \varphi\rangle| \to 0 \text{ in } Q \text{ probability}$$

and

$$\Gamma_T(\gamma^n, E) \text{ is } Q \text{ uniformly integrable.}$$

The required result follows from these observations. □

This is the analog of the result in [11] for our setup.

THEOREM 7.3. *Let $\nu \in \mathcal{P}(E)$ be given. Suppose that ν, γ satisfy*

$$\gamma T_t \to \mu \quad \nu T_t \to \mu. \tag{7.6}$$

Further, suppose that

$$Q_\gamma \textit{ is absolutely continuous with respect to } Q_\nu. \tag{7.7}$$

Then the filter π_t (defined by (7.2)) and the erroneous filter $\bar{\pi}_t$ wrongly computed with initial measure ν (defined by (7.3)) satisfy

$$\mathbb{E}_P[(\langle \pi_t, \varphi\rangle - \langle \bar{\pi}_t, \varphi\rangle)^2] \to 0. \tag{7.8}$$

Proof. As in [11], let the finite memory approximations $\pi_{t-\tau,t}$, $\bar{\pi}_{t-\tau,t}$ of π_t, $\bar{\pi}_t$, respectively, be defined by

$$\pi_{t-\tau,t} = \Lambda_\tau(\gamma T_{t-\tau}),$$

$$\bar{\pi}_{t-\tau,t} = \Lambda_\tau(\nu T_{t-\tau}).$$

Then given $\varepsilon > 0$, there exists τ_ε such that for $\tau > \tau_\varepsilon$ we have for $\varphi \in C(E)$

$$\mathbb{E}_P[(\langle \pi_t, \varphi\rangle - \langle \pi_{t-\tau,t}, \varphi\rangle)^2] \to 0 \tag{7.9}$$

and

$$\mathbb{E}_P[(\langle \bar{\pi}_t, \varphi\rangle - \langle \bar{\pi}_{t-\tau,t}, \varphi\rangle)^2] \to 0 \tag{7.10}$$

as t tends to infinity. These statements are proven in [11, Lemma 3.4]. The proof given there does not use boundedness of h or the underlying assumption in that paper that E is locally compact and hence carries over the present setup. Since φ is bounded, say, by K, we have

$$\mathbb{E}_P[(\langle \pi_{t-\tau,t}, \varphi\rangle - \langle \bar{\pi}_{t-\tau,t}, \varphi\rangle)^2] \leq 2K\mathbb{E}_P[|\langle \pi_{t-\tau,t}, \varphi\rangle - \langle \bar{\pi}_{t-\tau,t}, \varphi\rangle|]$$

and hence to complete the proof it remains to show that

$$\lim_{t\to\infty} \mathbb{E}_P[|\langle \pi_{t-\tau,t}, \varphi\rangle - \langle \bar{\pi}_{t-\tau,t}, \varphi\rangle|] = 0. \tag{7.11}$$

This in turn follows from Theorem 7.2 and the assumption (7.6) upon observing that $\pi_{t-\tau,t} = \Lambda_\tau(\gamma T_{t-\tau})$ and $\bar{\pi}_{t-\tau,t} = \Lambda_\tau(\nu T_{t-\tau})$. □

Acknowledgment. The authors would like to thank Professor T. G. Kurtz who pointed out that (5.1) is a consequence of the Feller property of (T_t). Equation (5.1) had been included as an assumption in an earlier version.

REFERENCES

[1] A. Bhatt, G. Kallianpur, and R. Karandikar, *Uniqueness and robustness of solution of measure-valued equations of nonlinear filtering*, Ann. Probab., 23 (1995), pp. 1895–1938.

[2] A. Bhatt, G. Kallianpur, and R. Karandikar, *Robustness of the nonlinear filter*, Stochastic Processes Appl., 81 (1999), pp. 247–254.

[3] A. Bhatt and R. Karandikar, *Path Continuity of the Nonlinear Filter*, preprint, 1999.

[4] A. Budhiraja and H. Kushner, *Approximation and limit results for nonlinear filters over an infinite time interval*, SIAM J. Control Optim., 37 (1999), pp. 1946–1979.

[5] A. Budhiraja and H. Kushner, *Approximation and limit results for nonlinear filters over an infinite time interval: Part* II, *random sampling algorithms*, SIAM J. Control Optim., 38 (2000), pp. 1874–1908.

[6] G. Kallianpur, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.

[7] G. Kallianpur and R. L. Karandikar, *White Noise Theory of Prediction, Filtering and Smoothing*, Gordon and Breach, New York, 1988.

[8] R. Karandikar, *On pathwise stochastic integration*, Stochastic Process. Appl., 57 (1995), pp. 11–18.

[9] H. Kunita, *Asymptotic behavior of the nonlinear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.

[10] H. Kunita, *Ergodic properties of nonlinear filtering processes*, in Spatial Stochastic Processes, K. Alexander and J. Watkins, eds., Birkhäuser, Boston, 1991, pp. 233–256.

[11] D. Ocone and E. Pardoux, *Asymptotic stability of the optimal filter with respect to its initial condition*, SIAM J. Control Optim., 34 (1996), pp. 226–243.

[12] L. Stettner, *On invariant measures of filtering processes*, in Stochastic Differential Systems, Proceedings of the 4th Bad Honnef Conference, 1988, Lecture Notes in Control and Inform Sci. 126, K. Helmes, N. Christopeit, and M. Kohlmann, eds., Springer-Verlag, Berlin, 1989, pp. 279–292.

SIAM J. Control Optim.
Vol. 39, No. 3, pp. 950–960

NONPARAMETRIC IDENTIFICATION OF CONTROLLED NONLINEAR TIME VARYING PROCESSES*

NADINE HILGERT†, RACHID SENOUSSI‡, and JEAN-PIERRE VILA†

Abstract. We are interested in the identification of an unknown time varying additive component of a controlled nonlinear autoregressive model, a problem of interest in the modeling and control of uncertain systems, such as those met in biotechnological processes. A kernel-based nonparametric estimator is proposed whose almost sure convergence is studied by means of a Lyapunov stabilizability assumption and laws of large numbers for martingales. We then adapt the general result to several classes of deterministic or random functional model uncertainties.

Key words. autoregressive process, stabilization, nonparametric estimation, convolution kernels

AMS subject classifications. 60F15, 62G05, 60G42

PII. S0363012998334456

1. Introduction.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space endowed with a filtration $\mathbb{F} = (\mathcal{F}_n)_{n\geq 0}$. We consider controlled autoregressive processes of the form

$$X_{n+1} = f_n(X_n) + F_n(X_n, U_n) + \varepsilon_{n+1} \tag{1.1}$$

with

- (X_n) a sequence of random variables taking values in $\mathbb{R}^s$,
- (U_n) a sequence of control variables taking values in $\mathbb{R}^m$,
- $(F_n(.,.))$ a sequence of known measurable functions,
- $(f_n(.))$ a sequence of unknown deterministic or random functions of $\mathcal{C}(\mathbb{R}^s, \mathbb{R}^s)$, the space of continuous functions endowed with the topology of uniform convergence on compacts,
- (ε_n) an unobservable white noise.

This setting can be extended to a more general type of model, frequently used in process control, especially in biotechnology, where $f_n(x) = H_n(x)\, g_n(x)$, H_n is a known matricial mapping from $\mathbb{R}^s$ to $\mathbb{R}^{s\times l}$, $l \leq s$, and g_n an unknown mapping from $\mathbb{R}^s$ to $\mathbb{R}^l$.

The model (1.1), thus adapted, represents, for example, the real time evolution of biomasses (microorganisms) and substrates concentrations in bioreactions (see [1]). Such reactions are very common in depollution or in the agro-food industry. In that case, $g_n(x)$ characterizes the microbial growth rate, which is a time varying quantity, influenced by many factors (biomasses and substrates concentrations, temperature, pH,...). For a given bioreaction the analytic form of $g_n(x)$ is generally not well known, in spite of its importance for a good modeling of the reaction dynamic and then its control.

The (X_n) process is assumed to be observed, and we are interested in estimating the functional sequence (f_n). A good estimation of this sequence is prerequisite to any control of the process.

*Received by the editors February 23, 1998; accepted for publication (in revised form) June 23, 2000; published electronically October 20, 2000.
http://www.siam.org/journals/sicon/39-3/33445.html

†Laboratoire d'Analyse des Systèmes et de Biométrie, INRA-ENSAM, 2 Place Viala, 34060 Montpellier cedex 1, France (hilgert@ensam.inra.fr, vila@ensam.inra.fr).

‡Laboratoire de Biométrie, INRA, Domaine Saint Paul, Site Agroparc, 84914 Avignon cedex 9, France (senoussi@avignon.inra.fr).

To the best of our knowledge, the identification of a random functional sequence (f_n) as in (1.1) has not yet been studied. This stochastic point of view allows a realistic treatment of the modeling of time varying dynamics. Until now only the estimation of unknown deterministic functions has been considered.

Georgiev [4] considered the model

$$X_{n+1} = f(X_n, U_n) + \varepsilon_{n+1},$$

where (X_n, U_n) is a specific stationary Markov process and $f(.,.)$ is the unknown functional component. He proved the weak consistency of a kernel estimator of f.

The use of kernel estimators for the adaptive control of nonlinear autoregressive processes was pioneered by Hernandez-Lerma and Doukhan [6] who faced the problem of the unavailability of any mixing property. Then, for the model

$$X_{n+1} = f(X_n) + F(X_n, U_n) + \varepsilon_{n+1},$$

where f is an unknown mapping from $\mathbb{R}^s$ to $\mathbb{R}^s$, Senoussi [9] showed the uniform convergence on compact subsets of a kernel estimator of f through a notion of stabilization of the process (X_n) and produced the related rates of convergence. For the same model, Portier [7] investigated adaptive control, i.e., determination of (U_n) through an adaptive functional estimator of f and a control objective. He got encouraging results in simple cases.

The identification of a sequence of unknown deterministic functions (f_n), rather than a unique function f, has been studied by Rutkowski [8] in the regression framework

$$Y_n = f_n(Z_n) + \varepsilon_n.$$

Under some restrictive assumptions (among them the convergence of f_n), he showed the strong point consistency of a kernel estimator of f.

The classes of sequences (f_n) we consider are more general.

In the context of model (1.1), we prove (section 2) the strong uniform consistency of a kernel estimator of the deterministic or random sequence (f_n). Results are then adapted to several types of sequences (f_n) of interest (section 2.4 and section 2.5).

2. Nonparametric approach of the identification problem. Convolution kernels were first used for probability density function estimation (Rosenblatt (1956), Parzen (1962), Wolverton and Wagner (1969) in [10]) and then for regression function estimation (Nadaraya (1964), Watson (1964), Härdle (1990) in [5]). A recent synthesis is in Bosq [2]. We consider the following estimator of $f_n(x)$ for all $x \in \mathbb{R}^s$:

$$\widehat{f}_n(x) = \frac{\sum_{i=0}^{n-1} h_i^{-s} K(\frac{x-X_i}{h_i})(X_{i+1} - F_i(X_i, U_i))}{\sum_{i=0}^{n-1} h_i^{-s} K(\frac{x-X_i}{h_i})}, \tag{2.1}$$

where $K(.)$ is an s-dimensional kernel, i.e., a bounded symmetric density with respect to Lebesgue measure, and h_i is a bandwidth parameter.

The kernel estimator (2.1) is said to be recursive since the bandwidth h_i depends on the current step i in the weighted sum and not on the number n of observations taken into account in the estimation. This recursive property is useful in the adaptive control framework, because $\widehat{f}_n(x)$ can easily be updated when a new observation is available. Besides, it allows the use of martingale tools [2, 3, 7, 9], which is, to our knowledge, the only way to provide asymptotic results with controlled processes.

2.1. Definitions. We define a control policy, or strategy, as a sequence of deterministic mappings $\delta = (\delta_k)$, $k \geq 0$, from $(\mathbb{R}^s)^k$ to the space of controls $\mathcal{U}$, such that $U_k = \delta_k(X_1, \ldots, X_k)$. For all $x \in \mathbb{R}^s$ we shall consider the set of admissible controls, with respect to x, to be a subset $A(x)$ of $\mathcal{U}$, for which $\delta_k(x_1, \ldots, x_{k-1}, x) \in A(x)$. A policy (δ_k) will be said to be *A-admissible*, or *admissible* for short, if for all k, $U_k \in A(X_k)$.

Moreover, model (1.1) is said to be *stabilized* by the use of any admissible policy, if we can choose a class $\mathcal{D}$ of strategies such that, for any $\varepsilon > 0$, there exists a compact set $\mathcal{C}$ of $\mathbb{R}^s$ satisfying the following property: for any initial law of X_0 and any strategy $\delta \in \mathcal{D}$,

$$\liminf_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} 1\!\!1_{\mathcal{C}}(X_i) \geq 1 - \varepsilon \quad \text{almost surely (a.s.),}$$

where $1\!\!1_{\mathcal{C}}$ stands for the indicator function of the set $\mathcal{C}$.

2.2. Main result. We assume that the control sequence (U_n) is adapted to the filtration $\mathbb{F}$, and that X_0 is $\mathcal{F}_0$-measurable. In the case of random functions f_n, the sequence (f_n) is also assumed to be $\mathbb{F}$-adapted. Moreover, (ε_n) is supposed to be an $\mathbb{F}$-adapted white noise. Let us recall that a random sequence (ξ_n) is said to be an $\mathbb{F}$-*adapted white noise* if for all $n \in \mathbb{N}$, ξ_{n+1} is $\mathcal{F}_{n+1}$-measurable and independent of $\mathcal{F}_n$.

The problem we are concerned with is to show the almost sure convergence of the kernel estimator (2.1). We shall require four different sets of hypotheses.

ASSUMPTION 2.1. *The noise $\varepsilon = (\varepsilon_n)$ is a sequence of independent and identically distributed (i.i.d. for short) random vectors with mean* 0 *and covariance matrix* Γ. *Its distribution is absolutely continuous (with respect to the Lebesgue measure), with a probability density function p supposed to be positive and C^1-class, p and its gradient are bounded.*

ASSUMPTION 2.2. *There exists a constant $\theta < 1$ such that*

$$\limsup_{||x||\to\infty} \frac{\sup_{i\in\mathbb{N}} \sup_{u\in A(x)}(||f_i(x) + F_i(x,u)||)}{||x||} \leq \theta \quad a.s. \tag{2.2}$$

ASSUMPTION 2.3. (a) *The sequence (f_n) is a.s. equicontinuous on compacts.*

(b) *There exists a function f such that, for all $x \in \mathbb{R}^s$,*

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n} i^{\alpha s} K(i^\alpha(X_i - x)) f_i(x)}{\sum_{i=0}^{n} i^{\alpha s} K(i^\alpha(X_i - x))} = f(x) \quad a.s. \tag{2.3}$$

Assumptions 2.1 and 2.2 are quite standard for nonlinear autoregressive control models; see, for instance, [2, 3, 5, 7, 9]. Let us note that Assumption 2.1 is satisfied with a classical nondegenerate Gaussian white noise. Assumption 2.2, also known as a stability criterion of Lyapunov, characterizes the stabilization ability of the model. It can be softened in numerous cases (see [3]), but generally requires exponential moments of the noise ε. Moreover, Assumption 2.3 obviously holds if the sequence (f_n) is a constant and continuous function f. Other cases that verify this hypothesis are developed as corollaries in section 2.4. Finally, the following Assumption 2.4 is made on the kernel K and the bandwidth (h_i).

ASSUMPTION 2.4. *The kernel K has a compact support and is Lipschitz continuous. The bandwidth $h_i := i^{-\alpha}$, $i \in \mathbb{N}$, is used, with $\alpha \in (0, 1)$.*

We shall now state our main result, which is proved in section 2.3.

Theorem 2.5. *Suppose that Assumptions* 2.1 *to* 2.4 *hold. Then*

(a) *for any* $0 < \alpha < 1/s$, *any admissible control policy and any initial probability distribution* ν *of* X_0, *we have for all* $x \in \mathbb{R}^s$

$$\widehat{f}_n(x) \longrightarrow f(x) \quad \text{a.s.} \quad \text{as } n \to \infty, \tag{2.4}$$

(b) *for* $0 < \alpha < 1/2s$, *if* ε *admits a finite moment of order strictly greater than* 2, *the almost sure convergence of* $\widehat{f}_n$ *to* f *is uniform on compacts.*

2.3. Proof. The proof of Theorem 2.5 requires two preliminary results. For all $x \in \mathbb{R}^s, \alpha \in (0, 1/s)$ and $n \in \mathbb{N}$, let us first denote by $D_n(x)$ the denominator of (2.1), i.e.,

$$D_n(x) := \sum_{i=1}^{n} i^{\alpha s} K\left(i^{\alpha}(x_i - x)\right).$$

Moreover, let $B(0, r)$ be the ball of radius r $(r > 0)$ centered on 0. The empirical measure of $B(0, r)$ associated to the transitions of the process (X_i) is given by

$$\Lambda_n(B(0, r)) := \sum_{i=0}^{n} 1\!\!1_{(||X_i|| \leq r)}. \tag{2.5}$$

Lemma 2.6. (a) *Under Assumptions* 2.1 *and* 2.2, *there exists a finite constant* r_0 *such that, for any initial distribution and any policy,*

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \|X_i\|^2 \leq r_0 \quad a.s. \tag{2.6}$$

(b) *It follows that for all* $r > r_0$,

$$\liminf_{n\to\infty} \frac{1}{n+1} \Lambda_n(B(0, r)) \geq 1 - r_0/r^2 \quad a.s., \tag{2.7}$$

that is to say, the process (X_n) *is stabilized by the use of any admissible policy.*

Proof of Lemma 2.6. (a) Let us define the new control variable for all $k \in \mathbb{N}$, $\widetilde{U}_k := (U_k, k) \in \widetilde{\mathcal{U}} = \mathcal{U} \times \mathbb{N}$ and the new admissible set $\widetilde{A}(x) := A(x) \times \mathbb{N}$. If δ_k is A-admissible, then $\widetilde{\delta}_k(x_1, \ldots, x_{k-1}, x) = (\delta_k(x_1, \ldots, x_{k-1}, x), k)$ is $\widetilde{A}$-admissible and model (1.1) can be rewritten as $X_{n+1} = g(X_n, \widetilde{U}_n) + \varepsilon_{n+1}$ with

$$g(x, \widetilde{u}) := g(x, u, m) = f_m(x) + F_m(x, u).$$

Assumption 2.2 is equivalently stated as follows:

$$\limsup_{\|x\|\to\infty} \frac{\sup_{\tilde{u}\in\tilde{A}(x)}(\|g(x, \widetilde{u})\|)}{\|x\|} < 1 \quad \text{a.s.},$$

and thus, applying Proposition 7.3.19 in Duflo [3] on the stabilization of controlled iterative models, there exists a finite constant r_0 such that (2.6) holds.

(b) Inequality (2.7) can easily be deduced from (2.5) and (2.6), since $\|X_i\|^2 \geq r^2 1\!\!1_{\{\|X_i\| \geq r\}}$. □

LEMMA 2.7. *Suppose that Assumptions* 2.1, 2.2, *and* 2.4 *hold. For any $a > 0$, there exist two constants m and M such that*

(a) *for $\alpha < 1/s$ and all x such that $\|x\| \leq a$,*

$$0 < m \leq \liminf_{n\to\infty} n^{-1} D_n(x) \leq \limsup_{n\to\infty} n^{-1} D_n(x) \leq M < \infty, \tag{2.8}$$

(b) *for $\alpha < 1/2s$, the previous inequalities hold uniformly on x, i.e.,*

$$0 < m \leq \liminf_{n\to\infty} \inf_{\|x\|\leq a} n^{-1} D_n(x) \leq \limsup_{n\to\infty} \sup_{\|x\|\leq a} n^{-1} D_n(x) \leq M < \infty. \tag{2.9}$$

Proof of Lemma 2.7. Since f_{i-1}, X_{i-1} and U_{i-1} are $\mathcal{F}_{i-1}$-measurable, and ε_i is independent of $\mathcal{F}_{i-1}$, we have, from (1.1),

$$\begin{aligned} &E(K(i^\alpha(X_i - x))/\mathcal{F}_{i-1}) \\ &= \int p(z - f_{i-1}(X_{i-1}) - F_{i-1}(X_{i-1}, U_{i-1}))K(i^\alpha(z-x))dz \\ &= i^{-\alpha s} \int K(z)p(i^{-\alpha}z + x - f_{i-1}(X_{i-1}) - F_{i-1}(X_{i-1}, U_{i-1}))dz. \end{aligned}$$

Let D_n^c be the compensator of D_n, i.e.,

$$\begin{aligned} D_n^c(x) &:= \sum_{i=1}^n i^{\alpha s} E(K(i^\alpha(X_i - x))/\mathcal{F}_{i-1}) \\ &= \sum_{i=1}^n \int K(z)p(i^{-\alpha}z + x - f_{i-1}(X_{i-1}) - F_{i-1}(X_{i-1}, U_{i-1}))dz. \end{aligned} \tag{2.10}$$

Let r_K be such that the compact support of K is included in the ball $B(0, r_K)$—see Assumption 2.4. We denote for $a > 0$ and $r > 0$

$$\begin{aligned} c_r &:= \sup\left(\|f_i(x) + F_i(x,u)\| \ : \ \|x\| \leq r, \ u \in A(x), \ i \in \mathbb{N}\right), \\ \gamma_{r,a} &:= \inf(p(z) \ : \ \|z\| \leq a + r_K + c_r). \end{aligned}$$

For r sufficiently large, Assumption 2.2 implies that $c_r < r$ and Assumption 2.1 yields that $\gamma_{r,a} > 0$. It follows, with $\|p\| := \sup_x \|p(x)\|$, that for all $i = 1, 2, \ldots,$

$$\begin{aligned} \|p\| &\geq \int K(z)p(i^{-\alpha}z + x - f_{i-1}(X_{i-1}) - F_{i-1}(X_{i-1}, U_{i-1}))dz \\ &\geq \int K(z)p(i^{-\alpha}z + x - f_{i-1}(X_{i-1}) - F_{i-1}(X_{i-1}, U_{i-1})) \\ &\qquad \times \mathbb{1}_{(\|x\|\leq a, \|z\|\leq r_K, \|X_{i-1}\|\leq r)} dz \\ &\geq \gamma_{r,a} \ \mathbb{1}_{(\|x\|\leq a, \|X_{i-1}\|\leq r)}, \end{aligned}$$

which, according to (2.5) and (2.10), gives

$$\|p\| \geq n^{-1} D_n^c(x) \geq \gamma_{r,a} \Lambda_n(B(0,r))$$

for all x such that $\|x\| \leq a$. Taking $r > \sqrt{2} r_0$, we then deduce by Assumption 2.2 that

$$\liminf_{n\to\infty} \inf_{\|x\|\leq a} n^{-1} D_n^c(x) \geq \gamma_{r,a}/2 = m > 0 \tag{2.11}$$

$$\text{and} \quad \limsup_{n\to\infty} \sup_{\|x\|\leq a} n^{-1} D_n^c(x) \leq \|p\| = M < \infty. \tag{2.12}$$

(a) Now, let us decompose $D_n(x)$ as $D_n(x) = M_n(x) + D_n^c(x)$, where $M_n(x) := D_n(x) - D_n^c(x)$. For all x, $M_n(x)$ is a martingale whose *hook*, $\langle M(x)\rangle_n$, defined by

$$\langle M(x)\rangle_n - \langle M(x)\rangle_{n-1} = E\Big((M_n(x) - M_{n-1}(x))^2/\mathcal{F}_{n-1}\Big), \tag{2.13}$$

satisfies

$$\begin{aligned}\langle M(x)\rangle_n &\le \sum_{i=1}^{n} i^{\alpha s} \int K(z)^2 p\big(i^{-\alpha} z + x - f_{i-1}(X_{i-1}) - F_{i-1}(U_{i-1}, X_{i-1})\big) dz\\ &\le cst\; n^{\alpha s+1},\end{aligned}$$

where *cst* stands for any finite constant. Thus $(n^{-(\alpha s+1)} E\langle M(x)\rangle_n)$ is bounded. The second law of large numbers for martingales (see [3]) then yields, for $\alpha s < 1$,

$$n^{-1}\big(D_n(x) - D_n^c(x)\big) \longrightarrow 0 \quad \text{a.s.},$$

which, together with (2.11) and (2.12), implies (2.8).

(b) To get the uniform version, let us consider the martingale $H_n(x, y) = M_n(x) - M_n(y)$ and recall that, for a Lipschitz continuous kernel K with compact support, for any $\rho > 0$, there exists a finite constant C such that

$$\|K(x) - K(y)\| \le C\|x - y\|^{\rho} \quad \text{for all } (x, y) \in \mathbb{R}^s \times \mathbb{R}^s. \tag{2.14}$$

Straightforward calculations easily yield that

$$\begin{aligned}(\langle H\rangle_n - \langle H\rangle_{n-1})(x, y) &\le n^{2\alpha s} E\Big((K(n^{\alpha}(X_n - x)) - K(n^{\alpha}(X_n - y)))^2/\mathcal{F}_{n-1}\Big)\\ &\le C^2 n^{2\alpha(s+\rho)} \|x - y\|^{2\rho}.\end{aligned}$$

It follows then, from the definition (2.13), that

$$E((M_n(x) - M_n(y))^2) \le cst \times n^{1+2\alpha(s+\rho)} \|x - y\|^{2\rho}.$$

As ρ is arbitrary, the uniform law of large numbers given in the appendix applies to M_n, i.e., if $\alpha < 1/2s$,

$$\lim_{n\to\infty} \sup_{\{\|x\|\le a\}} n^{-1} M_n(x) = 0 \quad \text{a.s.},$$

and then the inequalities (2.9) hold. □

Proof of Theorem 2.5. Let us write $\widehat{f}_n(x) - f(x)$ as follows:

$$\widehat{f}_n(x) - f(x) = \frac{1}{D_{n-1}(x)}(S_n(x) + R_{n,1}(x) + R_{n,2}(x)), \tag{2.15}$$

where

$$S_n(x) = \sum_{i=1}^{n-1} i^{\alpha s} K\big(i^{\alpha}(X_i - x)\big)\varepsilon_{i+1}, \tag{2.16}$$

$$R_{n,1}(x) = \sum_{i=1}^{n-1} i^{\alpha s} K\big(i^{\alpha}(X_i - x)\big)\big(f_i(X_i) - f_i(x)\big), \tag{2.17}$$

$$\text{and} \quad R_{n,2}(x) = \sum_{i=1}^{n-1} i^{\alpha s} K\big(i^{\alpha}(X_i - x)\big)\big(f_i(x) - f(x)\big). \tag{2.18}$$

The proof consists in studying separately S_n, $R_{n,1}$, and $R_{n,2}$ and then combining the different convergence results together with (2.8) and (2.9) in Lemma 2.7.

Study of S_n: $S_n(x)$ is for all x, a vector martingale whose hook—see (2.13)—in the sense of semipositive definite matrices, is

$$\langle S(x)\rangle_n = \sum_{i=1}^{n-1} i^{2\alpha s} K^2\big(i^\alpha(X_i - x)\big)\,\Gamma, \tag{2.19}$$

where Γ is the covariance matrix of ε. Now take the expectation of (2.19), and after an easy change of variable, we get

(2.20)

$$E\langle S(x)\rangle_n = \Gamma \sum_{i=1}^{n-1} i^{\alpha s} E\left(\int K^2(z) p\big(i^{-\alpha}z + x - f_i(X_{i-1}) - F_i(X_{i-1}, U_{i-1})\big)dz\right).$$

(a) From (2.20), we have trace $(E\langle S(x)\rangle_n) \leq cst \times n^{\alpha s+1}$. The second law of large numbers on martingales [3] then applies whenever $\alpha s < 1$, and yields that $n^{-1}S_n(x) \xrightarrow{a.s.} 0$.

(b) The proof of the uniform convergence of S_n is similar to the proof of Lemma 2.7(b), done with the martingale $M_n(x) = D_n(x) - D_n^c(x)$. Let us define the martingale $G_n(x, y) := S_n(x) - S_n(y)$. From the Lipschitz property (2.14), we deduce

$$\begin{aligned}(\langle G\rangle_{n+1} - \langle G\rangle_n)(x, y) &= n^{2\alpha s}(K(n^\alpha(X_n - x)) - K(n^\alpha(X_n - y)))^2\ \Gamma \\ &\leq C^2 n^{2\alpha(s+\rho)}\|x - y\|^{2\rho}\Gamma.\end{aligned}$$

Thus

$$E\|S_n(x) - S_n(y)\|^2 \leq cst\ n^{1+2\alpha(s+\rho)}\|x - y\|^{2\rho}.$$

As $\rho > 0$ is arbitrary, we conclude, using the uniform law of large numbers given in the appendix, that, whenever $2\alpha s < 1$ for all $a > 0$,

$$\lim_{n\to\infty} \sup_{\{\|x\|\leq a\}} n^{-1}S_n(x) = 0 \quad \text{a.s.}$$

Study of $R_{n,1}$: Since the compact support of K is included in $B(0, r_K)$, then $K(z) = 0$ when $||z|| > b$. Let $\eta > 0$ and $a > 0$. From Assumption 2.3(a), the sequence (f_n) is equicontinuous. Let us then denote by $\omega(\eta)$ the modulus of uniform continuity of (f_n), associated with η on the ball $B(0, a + r_k)$, i.e., for any pair (x, y) in the ball such that $||x - y|| < \eta$, we have

$$\sup_{i\in\mathbb{N}} ||f_i(x) - f_i(y)|| \leq \omega(\eta).$$

Then, for the first rank n_1 such that $bn_1^{-\alpha} < \eta$, we have

$$\sum_{i=n_1}^{n-1} i^{\alpha s}K\big(i^\alpha(X_i - x)\big)||f_i(X_i) - f_i(x)|| \leq \omega(\eta)\sum_{i=n_1}^{n-1} i^{\alpha s}K(i^\alpha(X_i - x)), \tag{2.21}$$

since, when $||i^\alpha(X_i - x)|| \leq b$ for $i \geq n_1$, we have

$$||X_i - x|| \leq bi^{-\alpha} \leq bn_1^{-\alpha} < \eta.$$

From (2.17) and (2.21), it follows that

$$\frac{1}{n-1}\|R_{n,1}(x)\| \leq \frac{cst}{n-1} + \frac{\omega(\eta)}{n-1}D_{n-1}(x). \tag{2.22}$$

(a) Let $\alpha < 1/s$ and x be such that $\|x\| \leq a$. From (2.8) and (2.22), we deduce

$$\limsup_{n\to\infty} \frac{1}{n-1}\|R_{n,1}(x)\| \leq \omega(\eta)\ M \quad \text{a.s.},$$

and then, $\limsup_{n\to\infty} \|R_{n,1}(x)\|/n = 0$ a.s., since $\lim_{\eta\to 0}\omega(\eta) = 0$.

(b) Similarly, for $\alpha < 1/2s$, (2.9) and (2.22) easily yield

$$\limsup_{n\to\infty} \sup_{\{\|x\|\leq a\}} \frac{1}{n}\|R_{n,1}(x)\| = 0 \quad \text{a.s.}$$

Study of $R_{n,2}$: By Assumption 2.3(a), the almost sure equicontinuity on compacts of the sequence (f_n) induces the almost sure equicontinuity on compacts of the weighted sums $R_{n,2}(x)/D_{n-1}(x)$. Then the pointwise convergence of $R_{n,2}/D_{n-1}$, given in Assumption 2.3(b), is strengthened to the uniform convergence on compacts, that is, for $\alpha s < 1$ and all $a > 0$,

$$\lim_{n\to\infty} \sup_{\{\|x\|\leq a\}} R_{n,2}(x)/D_{n-1}(x) = 0 \text{ a.s.}$$

Finally, from (2.15), combining the results obtained for each term D_n, S_n, $R_{n,1}$, and $R_{n,2}$ completes the proof of Theorem 2.5. □

2.4. Corollaries. The previous theorem can be adapted in several ways with respect to the nature of the unknown functional sequence (f_n). Recall also that only the pointwise convergence in (2.3) is required if (f_n) is a.s. equicontinuous.

2.4.1. The case of convergent sequences.

COROLLARY 2.8. *Assumption* 2.3 *holds with any sequence of continuous deterministic (resp., random) functions* f_n*, converging (resp., converging a.s.) uniformly on the compact subsets of* $\mathbb{R}^s$.

Proof. Let us note first that the uniform convergence of (f_n) implies the equicontinuity of the sequence. Thus, since $D_n = \sum_{i=0}^{\infty} i^{\alpha s}K(i^{\alpha}(X_i - x)) = \infty$ a.s., the application of the Toeplitz lemma yields (2.3) and Assumption 2.3 holds. □

REMARK 2.9.

- *The deterministic case* $f_n = f$ *includes the model of Senoussi* [9].
- *The regression model of Rutkowski* [8] *of independent sequences* (Z_n, Y_n) *is reproduced by letting* $X_n^1 = Z_{n+1}$, $X_n^2 = Y_n$, *and* $f_n(x) = E(Y_n \mid Z_n = x)$.

2.4.2. The case of martingale increments. Let (f_n) be a random sequence in $\mathcal{C}(\mathbb{R}^s, \mathbb{R}^s)$ and let us define the filtration $\mathbb{G} = (\mathcal{G}_n)_{n\geq 1}$, where $\mathcal{G}_n := \sigma(X_0, X_i, f_{i-1}, 1 \leq i \leq n)$.

COROLLARY 2.10. *Assumption* 2.3 *is satisfied if a.s. the sequence* (f_n) *is equicontinuous and if there exist two continuous and deterministic functions* f *and* γ *such that for all* $x \in \mathbb{R}^s$ *and* $n \in \mathbb{N}$,

$$f(x) := E\Big(f_n(x)/\mathcal{G}_n\Big) \text{ and } E\Big(\|f_n - f\|^2(x)/\mathcal{G}_n\Big) \leq \gamma(x). \tag{2.23}$$

Proof. For all x, the process

$$\widetilde{R}_n(x) := R_{n,2}(x) = \sum_{i=1}^{n-1} i^{\alpha s} K(i^{\alpha}(X_i - x))(f_i - f)(x)$$

is a $\mathbb{G}$-martingale of hook

$$\langle \widetilde{R} \rangle_n(x) = \sum_{i=1}^{n-1} i^{2\alpha s} K^2(i^{\alpha}(X_i - x)) E\Big((f_i - f)^t (f_i - f)(x)/\mathcal{G}_i\Big).$$

Following the proof of the asymptotic behavior of $S_n(x)$ (cf. section 2.3), we obtain

$$\begin{aligned} \text{trace} E\Big(\langle \widetilde{R} \rangle_n(x)\Big) &\leq \gamma(x) \sum_{i=1}^{n-1} i^{2\alpha s} E\Big(K^2(i^{\alpha}(X_i - x))\Big) \\ &\leq cst\ \gamma(x)\ n^{\alpha s + 1}. \end{aligned}$$

Since $\gamma(.)$ is continuous, it is finite on each compact. Thus, the second law of large numbers on martingales applies whenever $\alpha s < 1$, i.e., for all $a > 0$ and $\|x\| \leq a$, we have $n^{-1}\widetilde{R}_n(x) \to 0$, a.s. By Lemma 2.7(a), we deduce the almost sure pointwise convergence of $\widetilde{R}_n/D_{n-1}(x)$ to 0, that is to say, (2.3) holds. □

COROLLARY 2.11. *A sequence in an equicontinuous subset $\mathcal{A} \subset \mathcal{C}(\mathbb{R}^s, \mathbb{R}^s)$ of i.i.d. random functions f_n satisfying $E(\|f_i(0)\|^2) < \infty$ verifies Assumption* 2.3.

Proof. If $\omega(.)$ denotes the modulus of continuity of (f_n) on $\mathcal{A}$, then for $a > 0$ and $\|x\| \leq a$, it follows that

$$E\Big(\|f_i(x)\|^2\Big) \leq 2\omega^2(a) + 2E\Big(\|f_i(0)\|^2\Big),$$

since $\|f_i(x)\| \leq \|f_i(x) - f_i(0)\| + \|f_i(0)\| \leq \omega(a) + \|f_i(0)\|$. Thus, the functions $f(x) := E(f_i(x))$ and $\Gamma_f(x) := E((f_i - f)^t(f_i - f)(x))$ exist. They are continuous because of the almost sure equicontinuity of (f_n) and the pointwise convergence (by ergodicity) of $n^{-1}\sum_i f_i(x)$ and $n^{-1}\sum_i (f_i - f)^t(f_i - f)(x)$ to $f(x)$ and $\Gamma_f(x)$, respectively. Let us define $\gamma(x) := \text{trace}(\Gamma_f(x))$. Then (2.23) is satisfied and Corollary 2.10 holds. □

2.5. A special class of models. Let us consider models of the form $f_i(x) = H_i(x)g_i(x), i = 1, 2, \ldots$, where (H_i) is a sequence of known mappings from $\mathbb{R}^s$ into $\mathcal{M}_{s\times l}$, the set of $s \times l$ real matrices with $l \leq s$, and (g_i) is a sequence of unknown mappings from $\mathbb{R}^s$ into $\mathbb{R}^l$. This type of model is frequently met in biotechnological process control. The following lemma is needed for the definition of an appropriate estimator of the sequence (g_i).

LEMMA 2.12. *Let $H(.)$ denote a mapping from $\mathbb{R}^s$ into $\mathcal{M}_{s\times l}$, the set of $s \times l$ real matrices with $l \leq s$.*

(a) *If $H(.)$ has a constant rank l, it admits a left inverse in $\mathcal{M}_{l\times s}$, denoted by $H^-(.)$, of constant rank l, such that for all $x, H^-(x)H(x) = I_l$ the identity matrix of dimension l.*

(b) *Furthermore, if $H(.)$ is a continuous mapping, so is $H^-(.)$.*

Proof. As the rank of H is l, the square matrix ${}^tH.H$ is invertible for all x and one can take $H^- = ({}^tH.H)^{-1}\,{}^tH$. Since it is a composition of matrix transposition, product, and inversion, this mapping is continuous, irrespective of the norms considered. □

One can then define the following estimator of $g_n(x)$:

$$\widehat{g}_n(x) = \frac{\sum_{i=1}^{n-1} h_i^{-s} K(h_i^{-1}(x - X_i)) H_i^{-}(X_i)(X_{i+1} - F_i(X_i, U_i))}{\sum_{i=1}^{n-1} h_i^{-s} K(h_i^{-1}(x - X_i))} . \tag{2.24}$$

Assumption 2.3 about the sequence (g_i) can be strengthened by the following finiteness assumption.

ASSUMPTION 2.13. *For all* $r > 0$, $\sup\{||H_i^{-}(x)|| \,;\, i \geq 0, ||x|| \leq r\} := w(r) < \infty$.

COROLLARY 2.14. *Under Assumptions* 2.1, 2.2, 2.3, 2.4, *and* 2.13, *the estimator* $\widehat{g}_n$ *converges a.s. to* g *and uniformly on the compact subsets.*

Proof. It suffices to consider the martingale

$$S_n'(x) = \sum_{i=0}^{n} i^{\alpha s} K(i^{\alpha}(X_i - x)) H_i^{-}(X_i) \varepsilon_{i+1}.$$

As for $S_n(.)$, one can apply the uniform law of large numbers, by noticing that

$$\text{for all } x \,:\, ||x|| \leq r \,,\, K(i^{\alpha}(X_i - x))||H_i^{-}(X_i)|| \leq K(i^{\alpha}(X_i - x)) w(r + r_K),$$

where $B(0, r_K)$ is the ball containing the support of K. □

COROLLARY 2.15. *If* $H_i(x) = H(x)$ *is continuous, Assumption* 2.13 *is satisfied.*

Proof. H^{-} is a continuous mapping according to Lemma 2.12(b). H^{-} is then bounded on every compact. □

REMARK 2.16. *Corollaries* 2.8, 2.10, *and* 2.11 *can easily be adapted to the sequence* (g_i).

REMARK 2.17. *One could think of replacing* $H_i^{-}(X_i)$ *by* $H_n^{-}(x)$ *in the estimator given by* (2.24), *which amounts to first computing* $\widehat{f}_n(x) = \widehat{H_n(x) g_n}(x)$ *and then deducing an estimator of* $g_n(x)$ *by multiplying it by* $H_n^{-}(x)$. *But as it has been shown by simulation, the behavior of this estimator is less satisfactory than that of the previous one. This is not surprising since it does not make use of the knowledge of the* (H_i).

3. Concluding remarks. We have used convolution kernels for the nonparametric identification of an unknown component in an autoregressive nonlinear controlled process. This procedure could certainly be improved by some appropriate optimization of the bandwidth parameter and of course the type of kernel. Finally, let us point out that this nonparametric approach can be the first step towards a nonparametric adaptive process control.

Appendix. Let $(M_n(.))_{n \geq 0}$ be a random sequence of continuous functions, such that for all x, $(M_n(x))_n$ is a square-integrable martingale adapted to a filtration $\mathbb{F}$. The following theorem has been proved in Senoussi [9] and corresponds to Proposition 6.4.33 in Duflo [3].

THEOREM. *If there exist* $\alpha > 0$ *and* $\gamma > 0$ *such that*

(1) $E\big(||M_n(0)||^2\big) = \mathcal{O}(n^{\alpha})$,

(2) *given any* $A > 0$, *there exists* η *such that for all* x, y *with moduli less than* A,

$$E\Big(||M_n(x) - M_n(y)||^2\Big) \leq \eta n^{\alpha} ||x - y||^{s+\gamma},$$

then for $\beta > \alpha/2$ *and any* $A < \infty$, $\sup_{||x|| \leq A} n^{-\beta} ||M_n(x)|| \to 0$ *a.s.*

REFERENCES

[1] G. Bastin and D. Dochain, *On-line Estimation and Adaptive Control of Bioreactors*, Elsevier, New York, 1990.

[2] D. Bosq, *Nonparametric Statistics for Stochastic Processes, Estimation and Prediction*, Lecture Notes in Statist. 110, Springer-Verlag, New York, 1996.

[3] M. Duflo, *Random Iterative Models*, Springer-Verlag, New York, 1997.

[4] A. Georgiev, *Nonparametric system identification by kernel methods*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 356–358.

[5] W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK, 1989.

[6] O. Hernandez-Lerma and P. Doukhan, *Density Estimation and Adaptive Control of a Class of Discrete-Time Stochastic Systems*, Technical Report, CINVESTAV-IPN, Mexico City, Mexico, 1988.

[7] B. Portier, *Estimation non paramétrique et commande adaptative de processus Markoviens non linéaires*, Ph.D. Thesis, Université Paris-Sud, Paris, 1992.

[8] L. Rutkowski, *Nonparametric identification of quasi-stationary systems*, System Control Lett., 6, (1985), pp. 33–35.

[9] R. Senoussi, *Lois du logarithme itéré et identification*, Thèse d'état, Université Paris-Sud, Paris, 1991.

[10] C.T. Wolverton and T.J. Wagner, *Recursive estimates of probability densities*, IEEE Trans. Syst. Sci. Cyber., 5 (1969), pp. 246–247.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 961–980

UNIFORM CONVERGENCE AND MESH INDEPENDENCE OF NEWTON'S METHOD FOR DISCRETIZED VARIATIONAL PROBLEMS[*]

A. L. DONTCHEV[†], W. W. HAGER[‡], AND V. M. VELIOV[§]

Abstract. In an abstract framework, we study local convergence properties of Newton's method for a sequence of generalized equations which models a discretized variational inequality. We identify conditions under which the method is locally quadratically convergent, uniformly in the discretization. Moreover, we show that the distance between the Newton sequence for the continuous problem and the Newton sequence for the discretized problem is bounded by the norm of a residual. As an application, we present mesh-independence results for an optimal control problem with control constraints.

Key words. Newton's method, variational inequality, optimal control, sequential quadratic programming, discrete approximation, mesh independence

AMS subject classifications. 49M25, 65J15, 65K10

PII. S0363012998338570

1. Introduction.

In this paper we study local convergence properties of Newton-type methods applied to discretized variational problems. Our target problem is the variational inequality representing the first-order optimality conditions in constrained optimal control. In an abstract framework, the optimality conditions are modeled by a "generalized equation," a term coined by S. Robinson [12], where the normal cone mapping is replaced by an arbitrary map with closed graph. In this setting, Newton's method solves at each step a linearized generalized equation. When the generalized equation describes first-order optimality conditions, Newton's method becomes the well-known sequential quadratic programming (SQP) method.

We identify conditions under which Newton's method is not only locally quadratically convergent, but the convergence is uniform with respect to the discretization. Moreover, we derive an estimate for the number of steps required to achieve a given accuracy. Under some additional assumptions, which are natural in the context of the target problem, we prove that the distance between the Newton sequence for the continuous problem and the Newton sequence for the discretized problem, measured in the discrete metric, can be estimated by the norm of a residual. Normally, the residual tends to zero when the approximation becomes finer, and the two Newton sequences approach each other. In the context of the target optimal control problem, the residual is proportional to the mesh spacing h, uniformly along the Newton sequence. In particular, this implies that the least number of steps needed to reach a point at distance ε from the solution of the discrete problem does not depend on the mesh spacing; that is, the method is mesh independent.

[*] Received by the editors May 11, 1998; accepted for publication (in revised form) April 21, 1999; published electronically October 20, 2000. This research was supported by the National Science Foundation.

http://www.siam.org/journals/sicon/39-3/33857.html

[†] Mathematical Reviews, American Mathematical Society, Ann Arbor, MI 48107 (ald@ams.org).

[‡] Department of Mathematics, University of Florida, Gainesville, FL 32611 (hager@math.ufl.edu, http://www.math.ufl.edu/~hager).

[§] Institute of Mathematics and Informatics, Bulgarian Academy of Science, 1113 Sofia, Bulgaria, and University of Technology, Wiedner Hauptstr. 8–10/115, A-1040 Vienna, Austria (veliov@ uranus. tuwien.ac.at).

The convergence of the SQP method applied to nonlinear optimal control problems has been studied in several papers recently. In [5, 6] we proved local convergence of the method for a class of constrained optimal control problems. In parallel, Alt and Malanowski obtained related results for state constrained problems [3]. Along the same lines, Tröltzsch [13] studied the SQP method for a problem involving a parabolic partial differential equation.

Kelley and Sachs [10] were the first to obtain a mesh-independence result in constrained optimal control; they studied the gradient projection method. More recently, uniform convergence and mesh-independence results for an augmented Lagrangian version of the SQP method, applied to a discretization of an abstract optimization problem with equality constraints, were presented by Kunisch and Volkwein [11]. Alt [2] studied the mesh independence of Newton's method for generalized equations in the framework of the analysis of operator equations in Allgower et al. [1]. An abstract theory of mesh independence for infinite-dimensional optimization problems with equality constraints, together with applications to optimal control of partial differential equations and an extended survey of the field, can be found in the thesis of Volkwein [14].

The local convergence analysis of numerical procedures is closely tied to the problem's stability. The analysis is complicated for optimization problems with inequality constraints or for related variational inequalities. In this context, the problem solution typically depends on perturbation parameters in a nonsmooth way. In section 2 we present an implicit function theorem which provides a basis for our further analysis. In section 3 we obtain a result on uniform convergence of Newton's method applied to a sequence of generalized equations, while section 4 presents our mesh-independence results. Although in part parallel, our approach is different from the one used by Alt in [2], who adopted the framework of [1]. First, we study the uniform local convergence of Newton's method, which is not considered in [2]. In the mesh-independence analysis, we avoid consistency conditions for the solutions of the continuous and the discretized problems; instead, we consider the residual obtained when the Newton sequence of the continuous problem is substituted into the discrete necessary conditions. This allows us to obtain mesh independence under conditions weaker than those in [2] and, at the same time, to significantly simplify the analysis.

In section 5 we apply the abstract results to the constrained optimal control problem studied in our previous paper [5]. We show that under the smoothness and coercivity conditions given in [5] and assuming that the optimal control of the continuous problem is a Lipschitz continuous function of time, the SQP method applied to the discretized problem is Q-quadratically convergent, and the region of attraction and the constant of the convergence are independent of discretization, for a sufficiently small mesh size. Moreover, the l_∞ distance between the Newton sequence for the continuous problem at the mesh points and the Newton sequence for the discretized problem is of order $O(h)$. In particular, this estimate implies the mesh-independence result in Alt [2].

2. Lipschitzian localization.

Let X and Y be metric spaces. We denote both metrics by $\rho(\cdot,\cdot)$; it will be clear from the context which metric we are using. $B_r(x)$ denotes the closed ball with center x and radius r. In writing "f maps X into Y" we adopt the convention that the domain of f is a (possibly proper) subset of X. Accordingly, a set-valued map F from X to the subsets of Y may have empty values.

DEFINITION 2.1. *Let Γ map Y to the subsets of X and let $x^* \in \Gamma(y^*)$. We say that Γ has a Lipschitzian localization with constants a, b, and M around (y^*, x^*), if* *the map $y \mapsto \Gamma(y) \cap B_a(x^*)$ is single valued (a function) and Lipschitz continuous in $B_b(y^*)$ with a Lipschitz constant M.*

THEOREM 2.1. *Let G map X into the subsets of Y and let $y^* \in G(x^*)$. Let G^{-1} have a Lipschitzian localization with constants a, b, and M around (y^*, x^*). In addition, suppose that the intersection of the graph of G with $B_a(x^*) \times B_b(y^*)$ is closed and $B_a(x^*)$ is complete. Let the real numbers λ, $\bar{M}$, $\bar{a}$, m, and δ satisfy the relations*

$$\lambda M < 1, \quad \bar{M} = \frac{M}{1 - \lambda M}, \quad m + \delta < b, \quad \text{and} \quad \bar{a} + \bar{M}\delta < a. \tag{1}$$

Suppose that the function $g : B_a(x^*) \mapsto Y$ is Lipschitz continuous with a constant λ in the ball $B_a(x^*)$, that

$$\sup_{x \in B_a(x^*)} \rho(g(x), y^*) \le m, \tag{2}$$

and that the set

$$\Delta := \{x \in B_{\bar{a}}(x^*) : \operatorname{dist}(g(x), G(x)) \le \delta\} \tag{3}$$

is nonempty.

Then the set $\{x \in B_a(x^*) \mid g(x) \in G(x)\}$ consists of exactly one point, $\hat{x}$, and for each $x' \in \Delta$ we have

$$\rho(x', \hat{x}) \le \bar{M} \operatorname{dist}(g(x'), G(x')). \tag{4}$$

Proof. Let us choose positive $\lambda, \bar{M}, m, \bar{a}$, and δ such that the relations in (1) hold. We first show that the set $T := \{x \in B_a(x^*) \mid g(x) \in G(x)\}$ is nonempty. Let $x' \in \Delta$ and put $x_0 = x'$. Take an arbitrary $\varepsilon > 0$ such that

$$m + \delta + \varepsilon \le b \quad \text{and} \quad \bar{a} + \bar{M}(\delta + \varepsilon) \le a.$$

Choose an $y' \in G(x')$ such that $\rho(y', g(x')) \le \operatorname{dist}(g(x'), G(x')) + \varepsilon$. Since

$$\rho(y', y^*) \le \rho(y^*, g(x')) + \operatorname{dist}(g(x'), G(x')) + \varepsilon \le m + \delta + \varepsilon \le b$$

and

$$\rho(g(x_0), y^*) \le m \le b,$$

from the Lipschitzian localization property, there exists x_1 such that

$$g(x_0) \in G(x_1), \quad \rho(x_1, x_0) \le M\rho(y', g(x_0)) \le M(\operatorname{dist}(g(x'), G(x')) + \varepsilon). \tag{5}$$

We define inductively a sequence x_k in the following way. Let $x_0, \ldots, x_k$ be already defined for some $k \ge 1$ in such a way that

$$\rho(x_i, x_{i-1}) \le (\lambda M)^{i-1} \rho(x_1, x_0), \quad i = 1, \ldots, k, \tag{6}$$

and

$$g(x_{k-1}) \in G(x_k). \tag{7}$$

Clearly, x_0 and x_1 satisfy these relations. Using the second inequality in (5), we estimate

$$\begin{aligned} \rho(x_i, x^*) &\le \rho(x_0, x^*) + \sum_{j=1}^{i} \rho(x_j, x_{j-1}) \le \rho(x', x^*) + \sum_{j=0}^{\infty} (\lambda M)^j \rho(x_1, x_0) \\ &\le \bar{a} + \frac{1}{1 - \lambda M} M(\operatorname{dist}(g(x'), G(x')) + \varepsilon) \le \bar{a} + \bar{M}(\delta + \varepsilon) \le a. \end{aligned}$$

Thus both x_{k-1} and x_k are in $B_a(x^*)$ from which we obtain by (2)

$$\rho(g(x_i), y^*) \leq m \leq b$$

for $i = k-1$ and for $i = k$. Due to the assumed Lipschitzian localization property of G, there exists x_{k+1} such that (7), with k replaced by $k+1$, is satisfied and

$$\rho(x_{k+1}, x_k) \leq M\rho(g(x_k), g(x_{k-1})).$$

By (6) we obtain

$$\rho(x_{k+1}, x_k) \leq M\lambda\rho(x_k, x_{k-1}) \leq (\lambda M)^k \rho(x_1, x_0),$$

and hence (6) with k replaced by $k+1$, is satisfied. The definition of the sequence x_k is complete.

From (6) and the condition $\lambda M < 1$, $\{x_k\}$ is a Cauchy sequence. Since all $x_k \in B_a(x^*)$, the sequence $\{x_k\}$ has a limit $x'' \in B_a(x^*)$. Passing to the limit in (7), we obtain $g(x'') \in G(x'')$. Hence $x'' \in T$ and the set T is nonempty. Note that x'' may depend on the choice of ε. If we prove that the set T is a singleton, say $\hat{x}$, the point $x'' = \hat{x}$ would not depend on ε.

Suppose that there exist $x'' \in T$ and $\bar{x}'' \in T$ with $\rho(x'', \bar{x}'') > 0$. It follows that $\rho(g(x), y^*) \leq m \leq b$ for $x = x''$ and $x = \bar{x}''$. From the definition of the Lipschitzian localization, we obtain

$$\rho(x'', \bar{x}'') \leq M\rho(g(x''), g(\bar{x}'')) \leq M\lambda\rho(x'', \bar{x}'') < \rho(x'', \bar{x}''),$$

which is a contradiction. Thus T consists of exactly one point, $\hat{x}$, which does not depend on ε. To prove (4) observe that for any choice of $k > 1$,

$$\begin{aligned} \rho(x', x'') \leq \rho(x_0, x_k) + \rho(x_k, x'') &\leq \sum_{i=0}^{k-1} \rho(x_{i+1}, x_i) + \rho(x_k, x'') \\ &\leq \sum_{i=0}^{k-1} (\lambda M)^i \rho(x_1, x_0) + \rho(x_k, x''). \end{aligned}$$

Passing to the limit in the latter inequality and using (5), we obtain

$$\rho(x', x'') \leq \bar{M}(\operatorname{dist}(g(x'), G(x')) + \varepsilon). \tag{8}$$

Since $x'' = \hat{x}$ does not depend on the choice of ε, one can take $\varepsilon = 0$ in (8) and the proof is complete. □

3. Newton's method. Theorem 2.1 provides a basis for the analysis of the error of approximation and the convergence of numerical procedures for solving variational problems. In this and the following section we consider a sequence of so-called generalized equations. Specifically, for each $N = 1, 2, \ldots$, let X^N be a closed and convex subset of a Banach space, let Y^N be a linear normed space, let $f_N : X^N \mapsto Y^N$ be a function, and let $F_N : X^N \mapsto 2^{Y^N}$ be a set-valued map with closed graph. We denote by $\|\cdot\|_N$ the norms of both X^N and Y^N. We study the following sequence of problems:

$$\text{Find } x \in X^N \text{ such that } 0 \in f_N(x) + F_N(x). \tag{9}$$

We assume that there exist constants α, β, γ, and L, as well as points $x_N^* \in X^N$ and $z_N^* \in Y^N$, that satisfy the following conditions for each N:

(A1) $z_N^* \in f_N(x_N^*) + F_N(x_N^*)$.

(A2) The function f_N is Frechét differentiable in $B_\alpha(x_N^*)$ and the derivative ∇f_N is Lipschitz continuous in $B_\alpha(x_N^*)$ with a Lipschitz constant L.

(A3) The map

$$y \mapsto \Big(f_N(x_N^*) + \nabla f_N(x_N^*)(\cdot - x_N^*) + F_N(\cdot)\Big)^{-1}(y)$$

has a Lipschitzian localization with constants α, β, and γ around the point (z_N^*, x_N^*).

We study the Newton method for solving (9) for a fixed N which has the following form: If x^k is the current iterate, the next iterate x^{k+1} satisfies

$$0 \in f_N(x^k) + \nabla f_N(x^k)(x^{k+1} - x^k) + F_N(x^{k+1}), \quad k = 0, 1, \ldots, \tag{10}$$

where x^0 is a given starting point. If the range of the map F is just the origin, then (9) is an equation and (10) becomes the standard Newton method. If F is the normal cone mapping in a variational inequality describing first-order optimality conditions, then (10) represents the first-order optimality condition for the auxiliary quadratic program associated with the SQP method.

In the following theorem, by applying Theorem 2.1, we obtain the existence of a locally unique solution of the problem (9) which is at a distance from the reference point proportional to the norm of the residual z_N^*. We also show that the method (10) converges Q-quadratically and this convergence is uniform in N and in the choice of the initial point from a ball around the reference point x_N^* with radius independent of N. Note that for obtaining this result we do not pass to a limit and consequently we do not need to consider sequences of generalized equations.

THEOREM 3.1. *For every $\gamma' > \gamma$ there exist positive constants κ and σ such that if $\|z_N^*\| \leq \sigma$, then the generalized equation* (9) *has a unique solution x_N in $B_\kappa(x_N^*)$; moreover, x_N satisfies*

$$\|x_N - x_N^*\|_N \leq \gamma' \|z_N^*\|_N. \tag{11}$$

Furthermore, for every initial point $x^0 \in B_\kappa(x_N^*)$ there is a unique Newton sequence $\{x^k\}$, with $x^k \in B_\kappa(x_N^*)$, $k = 1, 2, \ldots$, and this Newton sequence is Q-quadratically convergent to x_N, that is,

$$\|x^{k+1} - x_N\|_N \leq \Theta \|x^k - x_N\|_N^2, \quad k = 0, 1, \ldots, \tag{12}$$

where Θ is independent of k, N and $x^0 \in B_\kappa(x_N^*)$.

Proof. Define

$$\kappa = \min\left\{\alpha, \gamma\beta, \frac{\gamma' - \gamma}{L\gamma\gamma'}, \frac{1}{5L\gamma'}\right\}, \quad \sigma = \frac{1}{\gamma'} \min\left\{\frac{\kappa}{4}, \sqrt{\frac{\kappa}{3L\gamma'}}, \frac{1}{6L\gamma'}\right\}, \quad \Theta = \frac{\gamma' L}{2}.$$

We will prove the existence and uniqueness of x_N by using Theorem 2.1 with

$$a = \kappa,\ b = \kappa/\gamma,\ M = \gamma,\ \lambda = \kappa L,\ \bar{M} = \gamma',\ \bar{a} = 0,\ m = \kappa^2 L/2 + \sigma,\ \delta = \sigma$$

and

$$g(x) = -f_N(x) + f_N(x_N^*) + \nabla f_N(x_N^*)(x - x_N^*),$$
$$G(x) = f_N(x_N^*) + \nabla f_N(x_N^*)(x - x_N^*) + F_N(x).$$

Observe that $a \le \alpha, b \le \beta$ and $\gamma b \le a$. By (A3) the map G has a Lipschitzian localization around (x_N^*, z_N^*) with constants a, b, and γ. One can check that the relations (1) are satisfied. Further, for x, x', and $x'' \in B_\kappa(x_N^*)$, we have

$$\|g(x) - z_N^*\|_N \le \|z_N^*\|_N + L\|x - x_N^*\|_N^2/2 \le \sigma + L\kappa^2/2 = m,$$

$$\begin{aligned}\|g(x') - g(x'')\|_N &\le \| - f_N(x') + f_N(x'') + \nabla f(x_N^*)(x' - x'')\|_N \\ &\le L\kappa\|x' - x''\|_N = \lambda\|x' - x''\|_N,\end{aligned}$$

$$\mathrm{dist}(g(x_N^*), G(x_N^*)) = \mathrm{dist}(0, f_N(x_N^*) + F(x_N^*)) \le \|z_N^*\|_N \le \sigma = \delta.$$

Obviously, $x_N^* \in B_0(x_N^*)$ and $x_N^* \in \Delta$, with Δ defined in (3). The assumptions of Theorem 2.1 are satisfied; hence there exists a unique x_N in $B_\kappa(x_N^*)$ for which $g(x_N) \in G(x_N)$. Hence x_N is a unique solution of (9) in $B_\kappa(x_N^*)$ and (11) holds. This completes the first part of the proof.

Given $x^k \in B_\kappa(x_N^*)$, a point x is a Newton step from x^k if and only if x satisfies the inclusion

$$g(x) \in G(x), \tag{13}$$

where G is the same as above, but now

$$g(x) = -f_N(x^k) - \nabla f_N(x^k)(x - x^k) + f_N(x_N^*) + \nabla f_N(x_N^*)(x - x_N^*).$$

The proof will be completed if we show that (13) has a unique solution x^{k+1} in $B_\kappa(x_N^*)$ and this solution satisfies (12). To this end we apply again Theorem 2.1 with $a, b, M, \bar{M}$, and λ the same as in the first part of the proof and with

$$\bar{a} = \sigma\gamma', \quad m = \sigma + \frac{5L\kappa^2}{2}, \quad \delta = \frac{L}{2}(\gamma'\sigma + \kappa)^2.$$

With these identifications, it can be checked that the assumptions (1) and (2) hold, and that g is Lipschitz continuous in $B_\kappa(x_N^*)$ with a Lipschitz constant λ. Further, by using the solution x_N obtained in the first part of the proof, we have

$$\begin{aligned}\mathrm{dist}(g(x_N), G(x_N)) &= \mathrm{dist}(0, f_N(x^k) + \nabla f_N(x^k)(x_N - x^k) + F_N(x_N)) \\ &\le \frac{L}{2}\|x_N - x^k\|_N^2 + \mathrm{dist}(0, f(x_N) + F_N(x_N)) = \frac{L}{2}\|x_N - x^k\|_N^2.\end{aligned} \tag{14}$$

The last expression has the estimate

$$\frac{L}{2}\|x_N - x^k\|_N^2 \le \frac{L}{2}\left(\|x_N - x_N^*\|_N + \|x_N^* - x^k\|_N\right)^2 \le \frac{L}{2}(\gamma'\sigma + \kappa)^2 = \delta.$$

Thus $x_N \in \Delta \ne \emptyset$ and the assumptions of Theorem 2.1 are satisfied. Hence, there exists a unique Newton step x^{k+1} in $B_\kappa(x_N^*)$ and by Theorem 2.1 and (14) it satisfies

$$\|x^{k+1} - x_N\|_N \le \frac{\gamma' L}{2}\|x^k - x_N\|_N^2 = \Theta\|x^k - x_N\|_N^2. \qquad \square$$

4. Mesh independence. Consider the generalized equation (9) under the assumptions (A1)–(A3). We present first a lemma in which, for simplicity, we suppress the dependence of N.

LEMMA 4.1. *For every $\gamma' > \gamma$, every $\mu > 0$, and every sufficiently small $\xi > 0$, there exists a positive η such that the map*

$$(y, w) \mapsto P(y, w) := (f(w) + \nabla f(w)(\cdot - w) + F(\cdot))^{-1}(y) \cap B_\alpha(x^*) \tag{15}$$

is a Lipschitz continuous function from $B_\eta(z^*) \times B_\xi(x^*)$ into $B_\xi(x^*)$ with Lipschitz constants γ' for y and μ for w.

Proof. Let $\gamma' > \gamma$ and $\mu > 0$. We choose the positive constants ξ and η as a solution of the following system of inequalities:

$$\gamma L\xi < 1, \quad \xi \le \frac{\gamma - \gamma'}{\gamma\gamma' L}, \quad 3\eta + \frac{15}{2}L\xi^2 + L\xi\alpha \le \beta,$$

$$\xi + \gamma'(2\eta + 6L\xi^2) \le \alpha, \quad 3L\xi\gamma' \le \mu, \quad \gamma'(\eta + 3L\xi^2) \le \xi.$$

This system of inequalities is satisfied by first taking ξ sufficiently small and then taking η sufficiently small. In particular, we have $\xi \le \alpha$ and $\eta \le \beta$.

Take $(y'', w'') \in B_\eta(z^*) \times B_\xi(x^*)$. We apply Theorem 2.1 with $a = \alpha$, $b = \beta$, $M = \gamma$, $\bar{a} = \xi$, $\bar{b} = \eta$, $\bar{M} = \gamma'$, $\lambda = L\xi$, $m = \eta + \frac{3}{2}L\xi^2 + L\xi\alpha$, $\delta = 2\eta + 6L\xi^2$,

$$g(x) = y'' + f(x^*) + \nabla f(x^*)(x - x^*) - f(w'') - \nabla f(w'')(x - w''),$$

and

$$G(x) = f(x^*) + \nabla f(x^*)(x - x^*) + F(x).$$

We have

$$\begin{aligned}\|g(x_1) - g(x_2)\| &= \|(\nabla f(x^*) - \nabla f(w''))(x_1 - x_2)\| \\ &\le L\|w'' - x^*\|\|x_1 - x_2\| \le L\xi\|x_1 - x_2\|\end{aligned}$$

for all $x_1, x_2 \in B_\alpha(x^*)$. Hence the function g is Lipschitz continuous with a Lipschitz constant λ. For $x \in B_\alpha(x^*)$ we have

$$\begin{aligned}\|g(x) - z^*\| &\le \|y'' - z^*\| + \|f(w'') - f(x^*) - \nabla f(x^*)(w'' - x^*)\| \\ &\quad + \|(\nabla f(x^*) - \nabla f(w''))(x - w'')\| \\ &\le \eta + \frac{L}{2}\|w'' - x^*\|^2 + L\|w'' - x^*\|\|x - w''\| \\ &\le \eta + \frac{1}{2}L\xi^2 + L\xi(\xi + \alpha) = m.\end{aligned}$$

Note that a point x is in the set $P(y'', w'')$ if and only if $g(x) \in G(x)$. Since

$$\begin{aligned}\operatorname{dist}(g(x^*), G(x^*)) &\le \|y'' - z^*\| + \operatorname{dist}(z^* - f(w'') - \nabla f(w'')(x^* - w''), F(x^*)) \\ &\le \eta + \operatorname{dist}(z^*, f(x^*) + F(x^*)) + \frac{L}{2}\|x^* - w''\|^2 \le \eta + \frac{L}{2}\xi^2 < \delta,\end{aligned}$$

the set Δ defined in (3) is not empty. Hence, from Theorem 2.1 the set $P(y'', w'') \cap B_\alpha(x^*)$ consists of exactly one point. Let us call it x''. Applying the same argument to an arbitrary point $(y', w') \in B_\eta(z^*) \times B_\xi(x^*)$, we obtain that there is exactly one point $x' \in P(y', w') \cap B_\alpha(x^*)$. Furthermore,

$$\begin{aligned}\operatorname{dist}(g(x'), G(x')) &\leq \|y' - y''\| + \|f(w'') - \nabla f(w'')(x' - w'') - f(w') - \nabla f(w')(x' - w')\| \\ &\leq \|y' - y''\| + \|f(w'') - f(w') - \nabla f(w')(w'' - w')\| \\ &\qquad\qquad\qquad + \|\nabla f(w'') - \nabla f(w')\|\|x' - w''\| \\ &\leq \|y' - y''\| + \frac{L}{2}\|w' - w''\|^2 + 2L\xi\|w' - w''\| \\ &\leq \|y' - y''\| + 3L\xi\|w' - w''\|.\end{aligned}$$

Hence $x' \in \Delta$ and we obtain

$$\rho(x', x'') \leq \gamma'(\|y' - y''\| + 3L\xi\|w' - w''\|) \leq \gamma'\|y' - y''\| + \mu\|w' - w''\|.$$

It remains to prove that P maps $B_\eta(z^*) \times B_\xi(x^*)$ into $B_\xi(x^*)$. From the last inequality with $x' = x^*$ and $w' = x^*$, we have

$$\rho(x'', x^*) \leq \gamma'(\|y'' - z^*\| + 3L\xi\|w'' - x^*\|) \leq \gamma'(\eta + 3L\xi^2) \leq \xi.$$

Thus $x'' \in B_\xi(x^*)$. □

In the remaining part of this section, we fix $\gamma' > \gamma$ and $0 < \mu < 1$, and we choose the constants κ and σ according to Theorem 3.1. For a positive ξ with $\xi \leq \kappa$, let η be the constant whose existence is claimed in Lemma 4.1. Note that η can be chosen arbitrarily small; we take $0 < \eta \leq \sigma$. Also, we assume that $\|z_N^*\| \leq \eta$ and consider Newton sequences with initial points $x^0 \in B_\xi(x_N^*)$. In such a way, the assumptions of Theorem 3.1 hold and we have a unique Newton sequence which is convergent quadratically to a solution.

Suppose that Newton's method (10) is supplied with the following stopping test: Given $\varepsilon > 0$, at the kth step the point x^k is accepted as an approximate solution if

$$\operatorname{dist}(0, f_N(x^k) + F_N(x^k)) < \varepsilon. \tag{16}$$

Denote by k_ε the first step at which the stopping test (16) is satisfied.

Theorem 4.1. *For any positive $\varepsilon < \eta$, if x^{k_ε} is the approximate solution obtained using the stopping test* (16) *at the step $k = k_\varepsilon$, then*

$$\|x^{k_\varepsilon} - x_N\|_N \leq \frac{\gamma'\varepsilon}{1 - \mu} \tag{17}$$

and

$$k_\varepsilon \leq 2 + \frac{1}{2}\log_\mu\left(\frac{\varepsilon}{2L\xi^2}\right). \tag{18}$$

Proof. Choose an ε such that $0 < \varepsilon < \eta$. If the stopping test (16) is satisfied at x^{k_ε}, then there exists v_ε^k with $\| v_\varepsilon^k \|_N \leq \varepsilon$ such that

$$v_\varepsilon^k \in f_N(x^{k_\varepsilon}) + F_N(x^{k_\varepsilon}).$$

Let P^N be defined as in (15) on the basis of f_N and F_N. Since

$$x^{k_\varepsilon} = P^N(v_\varepsilon^k, x^{k_\varepsilon}) \quad \text{and} \quad x_N \in P^N(0, x_N),$$

Lemma 4.1 implies that

$$\|x^{k_\varepsilon} - x_N\|_N \le \gamma' \parallel v_\varepsilon^k \parallel_N + \mu \|x^{k_\varepsilon} - x_N\|_N.$$

The latter inequality yields (17). For all $k < k_\varepsilon$, we obtain

$$\varepsilon \le \operatorname{dist}(0, f_N(x^k) + F_N(x^k)).$$

Since x^k is a Newton iterate, we have

$$f_N(x^k) - f_N(x^{k-1}) - \nabla f_N(x^{k-1})(x^k - x^{k-1}) \in f_N(x^k) + F_N(x^k).$$

Hence

$$\begin{aligned} \operatorname{dist}(0, f_N(x^k) + F_N(x^k)) &\le \parallel f_N(x^k) - f_N(x^{k-1}) - \nabla f_N(x^{k-1})(x^k - x^{k-1}) \parallel_N \\ &\le L\|x^k - x^{k-1}\|_N^2/2. \end{aligned} \tag{19}$$

By the definition of the map P^N, the Newton step x^1 from x^0 satisfies

$$x^1 = P^N(0, x^0),$$

while the Newton step x^2 from x^1 is

$$x^2 = P^N(0, x^1).$$

Since P^N is Lipschitz continuous with a constant μ, we have

$$\|x^2 - x^1\|_N \le \mu \|x^1 - x^0\|_N.$$

By induction, the $(k+1)$st Newton step x^{k+1} satisfies

$$\|x^{k+1} - x^k\|_N \le \mu^k \|x^1 - x^0\|_N. \tag{20}$$

Combining (19) and (20) and we obtain the estimate

$$\varepsilon \le 2L\xi^2 \mu^{2(k-1)},$$

which yields (18). □

Our next result provides a basis for establishing the mesh independence of Newton's method (10). Namely, we compare the Newton sequence x_N^k for the "discrete" problem (9) and the Newton sequence for a "continuous" problem which is again described by (9) but with index $N = 0$. Let us assume that the conditions (A1)–(A3) hold for the generalized equation (9) with $N = 0$. According to Theorem 3.1, for each starting point x_0^0 close to a solution x_0, there is a unique Newton sequence x_0^k which converges to x_0 Q-quadratically. To relate the continuous problem to the discrete one, we introduce a mapping π_N from X^0 to X^N. Having in mind the application to optimal control presented in the following section, X^0 can be thought as a space of continuous functions $x(\cdot)$ in $[0, 1]$ and, for a given natural number N, $t_0 = 0$ and $t_i = i/N$, X^N will be the space of sequences $\{x_i, i = 0, 1, \ldots, N\}$. In this case the operator π_N is the interpolation map $\pi_N(x(\cdot)) = (x(t_0), \ldots, x(t_N))$.

Theorem 4.2. *Suppose that for every k and N there exists $r_N^k \in Y^N$ such that*

$$r_N^k \in f_N(\pi_N(x_0^k)) + \nabla f_N(\pi_N(x_0^k))(\pi_N(x_0^{k+1}) - \pi_N(x_0^k)) + F_N(\pi_N(x_0^{k+1}))$$

and

$$\omega_N := \sup_k \parallel r_N^k \parallel_N < \eta. \tag{21}$$

In addition, let

$$\|\pi_N(x_0^k) - x_N^*\|_N \leq \xi$$

for all k and N. Then for all $k = 1, 2, \ldots$ and N

$$\|x_N^{k+1} - \pi_N(x_0^{k+1})\|_N \leq \frac{\gamma'}{1-\mu}\omega_N + \mu^{k+1}\|x_N^0 - \pi_N(x_0^0)\|_N. \tag{22}$$

Proof. By definition, we have

$$\pi_N(x_0^{k+1}) = P^N(r_N^k, \pi_N(x_0^k)) \quad \text{and} \quad x_N^{k+1} = P^N(0, x_N^k).$$

Using Lemma 4.1 we have

$$\|x_N^{k+1} - \pi_N(x_0^{k+1})\|_N \leq \gamma' \parallel r_N^k \parallel_N + \mu\|x_N^k - \pi_N(x_0^k)\|_N \leq \gamma'\omega_N + \mu\|x_N^k - \pi_N(x_0^k)\|_N.$$

By induction we obtain (22). □

The above result means that, under our assumptions, the distance between the Newton sequence for the continuous problem and the Newton sequence for the discretized problem, measured in the discrete metric, can be estimated by the sup-norm ω_N of the residual obtained when the Newton sequence for the continuous problem is inserted into the discretized generalized equations. If the sup-norm of the residual tends to zero when the approximation becomes finer, that is, when $N \to \infty$, then the two Newton sequences approach each other. In the next section, we will present an application of the abstract analysis to an optimal control problem for which the residual is proportional to the mesh spacing h, uniformly along the Newton sequence. For this particular problem Theorem 4.2 implies that the distance between the Newton sequences for the continuous problem and the Newton sequence for the discretized problem is $O(h)$.

For simplicity, let us assume that if the continuous Newton process starts from the point x_N^0, then the discrete Newton process starts from $\pi_N(x_0^0)$. Also, suppose that for any fixed $w, v \in X^0$,

$$\|\pi_N(w) - \pi_N(v)\|_N \to \|w - v\|_0 \quad \text{as} \quad N \to \infty. \tag{23}$$

In addition, let

$$\omega_N \to 0 \quad \text{as} \quad N \to \infty, \tag{24}$$

where ω_N is defined in (21). Letting k tend to infinity and assuming that π_N is a continuous mapping for each N, Theorem 4.2 gives us the following estimate for the distance between the solution x_N of the discrete problem and the discrete representation $\pi_N(x_0)$ of the solution x_0 of the continuous problem:

$$\|x_N - \pi_N(x_0)\|_N \leq \frac{\gamma'}{1-\mu}\omega_N. \tag{25}$$

Choose a real number ε satisfying

$$0 < \varepsilon < 1/(5\Theta), \tag{26}$$

where Θ is as in Theorem 3.1. Theorem 4.2 yields the following result.

THEOREM 4.3. *Let* (23) *and* (24) *hold and let* ε *satisfy* (26). *Then for all* N *sufficiently large,*

$$|\min\left\{k \in \mathbf{N} : \|x_0^k - x_0\|_0 < \varepsilon\right\} - \min\left\{k \in \mathbf{N} : \|x_N^k - x_N\|_N < \varepsilon\right\}| \leq 1. \tag{27}$$

Proof. Let m be such that

$$\|x_0^{m+1} - x_0\|_0 < \varepsilon \leq \|x_0^m - x_0\|_0. \tag{28}$$

Choose N so large that

$$\frac{\gamma'}{1-\mu}\omega_N < \varepsilon/2$$

and

$$\|\pi_N(x_0^{m+1}) - \pi_N(x_0)\|_N \leq \varepsilon.$$

Using Theorem 3.1, Theorem 4.2, (25), and (29), we obtain

$$\begin{aligned}\|x_N^{m+2} - x_N\|_N &\leq \Theta\|x_N^{m+1} - x_N\|_N^2\\ &\leq \Theta\left(\|x_N^{m+1} - \pi_N(x_0^{m+1})\|_N + \|\pi_N(x_0^{m+1}) - \pi_N(x_0)\|_N + \|\pi_N(x_0) - x_N\|_N\right)^2\\ &\leq \Theta(\varepsilon/2 + \varepsilon + \varepsilon/2)^2 = 4\Theta\varepsilon^2 < \varepsilon.\end{aligned}$$

This means that if the continuous Newton sequence achieves accuracy ε (measured by the distance to the exact solution) at the mth step, then the discrete Newton sequences should achieve the same accuracy ε at the $(m+1)$st step or earlier. Now we show that the latter cannot happen earlier than at the $(m-1)$st step. Choose N so large that

$$\|x_0^{m-1} - x_0\|_0^2 \leq \|\pi_N(x_0^{m-1}) - \pi_N(x_0)\|_N^2 + \varepsilon^2 \tag{29}$$

and suppose that

$$\|x_N^{m-1} - x_N\|_N < \varepsilon.$$

From Theorem 3.1, (22), (25), (28), and (29), we get

$$\begin{aligned}\varepsilon \leq \|x_0^m - x_0\|_0 \leq \Theta\|x_0^{m-1} - x_0\|_0^2 &\leq \Theta\|\pi_N(x_0^{m-1}) - \pi_N(x_0)\|_N^2 + \varepsilon^2\\ &\leq \Theta\left(\|\pi_N(x_0^{m-1}) - x_N^{m-1}\|_N + \|x_N^{m-1} - x_N\|_N + \|x_N - \pi_N(x_0)\|_N\right)^2 + \varepsilon^2\\ &\leq \Theta(\varepsilon/2 + \varepsilon + \varepsilon/2)^2 + \varepsilon^2 = 5\Theta\varepsilon^2,\end{aligned}$$

which contradicts the choice of ε in (26). □

5. Application to optimal control.

We consider the following optimal control problem:

$$\text{minimize } G(y(1)) + \int_0^1 \varphi(y(t), u(t))\, dt \tag{30}$$

$$\text{subject to} \quad \dot{y}(t) = g(y(t), u(t)) \text{ and } u(t) \in U \text{ for almost every (a.e.) } t \in [0,1],$$
$$y(0) = y_0,\ y \in W^{1,\infty}(\mathbb{R}^n), \text{ and } u \in L^\infty(\mathbb{R}^m),$$

where $\varphi : \mathbb{R}^{n+m} \to \mathbb{R}$, $g : \mathbb{R}^{n+m} \to \mathbb{R}^n$, $G : \mathbb{R}^n \to \mathbb{R}$, U is a nonempty, closed and convex set in $\mathbb{R}^m$, and y_0 is a fixed vector in $\mathbb{R}^n$. $L^\infty(\mathbb{R}^m)$ denotes the space of essentially bounded and measurable functions with values in $\mathbb{R}^m$ and $W^{1,\infty}(\mathbb{R}^n)$ is the space of Lipschitz continuous functions with values in $\mathbb{R}^n$.

We are concerned with local analysis of the problem (30) around a fixed local minimizer (y^*, u^*) for which we assume certain regularity. Our first standing assumption is the following:

Smoothness. *The optimal control u^* is Lipschitz continuous in $[0,1]$. There exists a positive number δ such that the first three derivatives of φ and g exist and are continuous in the set $\{(y,u) \in \mathbb{R}^{n+m} : |y - y^*(t)| + |u - u^*(t)| \le \delta$ for all $t \in [0,1]\}$.*

Defining the Hamiltonian H by

$$H(y,u,\psi) = \varphi(y,u) + \psi^\mathsf{T} g(y,u),$$

it is well known that the first-order necessary optimality conditions at the solution (y^*, u^*) can be expressed in the following way: There exists $\psi^* \in W^{1,\infty}(\mathbb{R}^n)$ such that (y^*, u^*, ψ^*) is a solution of the variational inequality

$$\dot{y}(t) = g(y(t), u(t)), \quad y(0) = y_0, \tag{31}$$

$$\dot{\psi}(t) = -\nabla_y H(y(t), u(t), \psi(t)), \quad \psi(1) = \nabla G(y(1)), \tag{32}$$

$$0 \in \nabla_u H(y(t), u(t), \psi(t)) + N_U(u(t)) \quad \text{for a.e. } t \in [0,1], \tag{33}$$

where $N_U(u)$ is the normal cone to the set U at the point u; that is, $N_U(u)$ is empty if $u \notin U$, while for $u \in U$,

$$N_U(u) = \{p \in \mathbb{R}^m : p^\mathsf{T}(q - u) \le 0 \text{ for all } q \in U\}.$$

Although the problem (30) is posed in L^∞ and the optimality system (31)–(33) is satisfied a.e. in $[0,1]$, the regularity we assume for the particular optimal solution implies that at (y^*, u^*, ψ^*) the relations (31)–(33) hold everywhere in $[0,1]$.

Defining the matrices

$$A(t) = \nabla_y g(z^*(t)), \quad B(t) = \nabla_u g(z^*(t)), \quad V = \nabla^2 G(y^*(1)),$$
$$Q(t) = \nabla^2_{yy} H(x^*(t)), \quad R(t) = \nabla^2_{uu} H(x^*(t)), \quad S(t) = \nabla^2_{yu} H(x^*(t)),$$

where $z^* = (y^*, u^*)$ and $x^* = (y^*, u^*, \psi^*)$, we employ the following coercivity condition.

Coercivity. *There exists $\alpha > 0$ such that*

$$y(1)^\mathsf{T} V y(1) + \int_0^1 [y(t)^\mathsf{T} Q(t) y(t) + u(t)^\mathsf{T} R(t) u(t) + 2y(t)^\mathsf{T} S(t) u(t)]\, dt \ge \alpha \int_0^1 |u(t)|^2\, dt \tag{34}$$

whenever $y \in W^{1,2}(\mathbb{R}^n)$, $y(0) = 0, u \in L^2(\mathbb{R}^n), \dot{y} = Ay + Bu, u(t) \in U - U$ for a.e. $t \in [0,1]$.

Let N be a natural number, let $h = 1/N$ be the mesh spacing, let $t_i = ih$, and let y_i' denote the forward difference operator defined by

$$y_i' = \frac{y_{i+1} - y_i}{h}.$$

We consider the following Euler discretization of the optimality system (31)–(33):

$$y_i' = \nabla_\psi H(y_i, u_i, \psi_i), \tag{35}$$

$$\psi_{i-1}' = -\nabla_y H(y_i, u_i, \psi_i), \quad \psi_{N-1} = \nabla G(y_N), \tag{36}$$

$$0 \in \nabla_u H(y_i, u_i, \psi_i) + N_U(u_i), \quad i = 0, 1, \ldots, N-1. \tag{37}$$

The system (35)–(37) is a discrete-time variational inequality depending on the step size h. It represents the first-order necessary optimality condition for the following discretization of the original problem (30):

$$\text{minimize} \quad G(y_N) + \sum_{i=0}^{N-1} h\varphi(y_i, u_i) \tag{38}$$
$$\text{subject to} \;\; y_i' = g(y_i, u_i), \;\; u_i \in U, \; i = 0, 1, \dots, N-1.$$

In this section we examine the following version of the Newton method for solving the variational system (35)–(37), which correspond to the SQP method for solving the optimization problem (38). Let $x^k = (y^k, u^k, \psi^k)$ denote the kth iterate. Let the superscript k and the subscript i attached to the derivatives of H and G denote their values at x_i^k. Then the new iterate $x^{k+1} = (y^{k+1}, u^{k+1}, \psi^{k+1})$ is a solution of the following linear variational inequality for the variable $x = (y, u, \psi)$:

$$y_i' = \nabla_\psi H_i^k + \nabla^2_{\psi x} H_i^k (x_i - x_i^k), \tag{39}$$
$$\psi_{i-1}' = -\nabla_y H_i^k - \nabla^2_{yx} H_i^k (x_i - x_i^k), \quad \psi_{N-1} = \nabla G^k + \nabla^2 G^k (y_N - y_N^k), \tag{40}$$
$$0 \in \nabla_u H_i^k + \nabla^2_{ux} H_i^k (x_i - x_i^k) + N_U(u_i), \quad i = 0, 1, \dots, N-1. \tag{41}$$

In [5, Appendix 2], it was proved that the coercivity condition (34) is stable under the Euler discretization, then the variational system (39)–(41) is equivalent, for x^k near $x^* = (y^*, u^*, \psi^*)$, to the following linear-quadratic discrete-time optimal control problem which is expressed in terms of the variables y, u, and $z = (y, u)$:

$$\begin{aligned} \text{minimize} \quad & \left(\nabla G^k + \frac{1}{2}\nabla^2 G^k (y_N - y_N^k)\right)^{\mathsf{T}} (y_N - y_N^k) \\ & + \; h \sum_{i=0}^{N-1} \left(\nabla_z \varphi_i^k + \frac{1}{2}\nabla^2_{zz} H_i^k (z_i - z_i^k)\right)^{\mathsf{T}} (z_i - z_i^k) \\ \text{subject to} \quad & y_i' = g_i^k + \nabla_z g_i^k (z_i - z_i^k), \quad u_i \in U, \quad i = 0, 1, \dots, N-1. \end{aligned}$$

A natural stopping criterion for the problem at hand is the following: Given $\varepsilon > 0$, a control $\tilde{u}^k$ obtained at the kth iteration is considered an ε-optimal solution if

$$\max_{0 \le i \le N-1} \operatorname{dist}(\nabla_u H(\tilde{y}_i^k, \tilde{u}_i^k, \tilde{\psi}_i^k), N_U(\tilde{u}_i^k)) \le \varepsilon, \tag{42}$$

where $\tilde{y}_i^k$ and $\tilde{\psi}_i^k$ are the solutions of the state and the adjoint equations (35) and (36) correspond to $u = \tilde{u}^k$.

We now apply the general approach developed in the previous section to the discrete-time variational inequality (35)–(36). The discrete L_N^∞ norm is defined by

$$\|v\|_N^\infty = \max_{0 \le i \le N-1} |v_i|.$$

The variable x is the triple (y, u, ψ) while X^N is the space of all finite sequences $x_0, x_1, \dots, x_{N-1}$, with y_0 given, equipped with the L_N^∞ norm. The space Y^N is the Cartesian product $L_N^\infty \times L_N^\infty \times \mathbb{R}^n \times L_N^\infty$ corresponding to the four components of the function f_N defined by

$$f_N(x)_i = \begin{pmatrix} y_i' - g(y_i, u_i) \\ -\psi_{i-1}' + \nabla_y H(y_i, u_i, \psi_i) \\ \psi_{N-1} - \nabla G(y_N) \\ -\nabla_u H(y_i, u_i, \psi_i) \end{pmatrix} \quad \text{and} \quad F_N(x)_i = \begin{pmatrix} 0 \\ 0 \\ 0 \\ N_U(u_i) \end{pmatrix}.$$

With the choice $(x_N^*)_i = (y^*(t_i), u^*(t_i), \psi^*(t_i))$, the general condition (A1) is satisfied by taking

$$(z_N^*)_i = \begin{pmatrix} (y^*(t_{i+1}) - y^*(t_i))/h - g(y^*(t_i), u^*(t_i)) \\ (\psi^*(t_{i-1}) - \psi^*(t_i))/h - \nabla_x H(y^*(t_i), u^*(t_i), \psi^*(t_i)) \\ 0 \\ 0 \end{pmatrix}.$$

The first component of z_N^* is estimated in the following way:

$$\sup_i \left| \frac{y^*(t_{i+1}) - y^*(t_i)}{h} - g(y^*(t_i), u^*(t_i)) \right| \\ \leq \sup_i \frac{1}{h} \int_{t_i}^{t_{i+1}} |g(y^*(t_i), u^*(t_i)) - g(y^*(t), u^*(t))| dt.$$

Since g is smooth and both y^* and u^* are Lipschitz continuous, the above expression is bounded by $O(h)$. The same bound applies for the second component of z_N^*, while the third and fourth components are zero. Thus the norm of z_N^* can be made arbitrarily small for all sufficiently large N. Condition (A2) follows from the smoothness assumption. A proof of condition (A3) is contained in the proof of Theorem 6 in [5]. Applying Theorems 3.1 and 4.1 and using the result from [5, Appendix 2], that the discretized coercivity condition is a second-order sufficient condition for the discrete problem, we obtain the following theorem.

THEOREM 5.1. *If smoothness and coercivity hold, then there exist positive constants K, c, σ, $\bar{\varepsilon}$, and $\bar{N}$ with the property that for every $N > \bar{N}$ there is a unique solution (y_h, u_h, ψ_h) of the variational system* (35)–(37) *and (y_h, u_h) is a local minimizer for the discrete problem* (38). *For every starting point (y^0, u^0, ψ^0) with*

$$\max_{0 \leq i \leq N} \left(|(y^0)_i - y^*(t_i)| + |(u^0)_i - u^*(t_i)| + |(\psi^0)_i - \psi^*(t_i)| \right) \leq \sigma,$$

there is a unique SQP sequence (y^k, u^k, ψ^k) and it is Q-quadratically convergent, with a constant K, to the solution (y_h, u_h, ψ_h). In particular, for the sequence of controls we have

$$\max_{0 \leq i \leq N-1} |(u^{k+1})_i - (u_h)_i| \leq K \left(\max_{0 \leq i \leq N-1} |(u^k)_i - (u_h)_i| \right)^2.$$

Moreover, if the stopping test (42) *is applied with an $\varepsilon \in [0, \bar{\varepsilon}]$, then the resulting ε-optimal control u^{k_ε} satisfies*

$$\max_{0 \leq i \leq N-1} |u_i^{k_\varepsilon} - u^*(t_i)| \leq c(\varepsilon + h).$$

Note that the termination step k_ε corresponding to the assumed accuracy of the stopping test can be estimated by Theorem 4.1. Combining the error in the discrete control with the discrete state equation (35) and the discrete adjoint equation (36), yield corresponding estimates for discrete state and adjoint variables.

Remark. Following the approach developed in [5] one can obtain an analogue of Theorem 5.1 by assuming that the optimal control u^* is merely bounded and Riemann integrable in $[0, 1]$ and employing the so-called averaged modulus of smoothness to obtain error estimates.. The stronger Lipschitz continuity condition for the optimal control is, however, needed in our analysis of the mesh independence.

The SQP method applied to the continuous-time optimal control problem (30) has the following form: If $x^0 = (y^0, u^0, \psi^0)$ is a starting point, the iterate $x^{k+1} = (y^{k+1}, u^{k+1}, \psi^{k+1})$ satisfies

$$\dot{y}(t) = \nabla_\psi H^k(t) + \nabla^2_{\psi x} H^k(t)(x(t) - x^k(t)),\ y(0) = y_0, \tag{43}$$

$$\dot{\psi}(t) = -\nabla_y H^k(t) - \nabla^2_{yx} H^k(t)(x(t) - x^k(t)), \tag{44}$$

$$\psi(1) = \nabla G^k(1) + \nabla^2 G^k(y(1) - y^k(1)), \tag{45}$$

$$0 \in \nabla_u H^k(t) + \nabla^2_{ux} H^k(t)(x(t) - x^k(t)) + N_U(u(t)) \tag{46}$$

for a.e. $t \in [0,1]$, where the superscript k attached to the derivatives of H and G denotes their values at x^k. In particular, (43)–(46) is a variational inequality to which we can apply the general theory from the previous sections. We attach the index $N = 0$ to the continuous problem and for $x = (y, u, \psi)$ we choose $X^0 = C^1_0(\mathbb{R}^n) \times C(\mathbb{R}^m) \times C^1(\mathbb{R}^n)$, where $C^1_0 = \{y \in C^1 \mid y(0) = y_0\}$, and $Y^0 = C(\mathbb{R}^n) \times C(\mathbb{R}^n) \times \mathbb{R}^n \times C(\mathbb{R}^m)$. Condition (A1) is clearly satisfied with $x^*_0 = x^* := (y^*, u^*, \psi^*)$ and $z^*_0 = 0$. Condition (A2) follows from the smoothness assumption. Condition (A3) follows from the coercivity assumption as proved in [9, Lemma 3] (see also [4, section 2.3.4], for an earlier version of this result in the convex case). Hence, we can apply Theorem 3.1 obtaining that for any sufficiently small ball $\mathcal{B}$ around x^* (in the norm of X^0), if the starting point x^0 is chosen from $\mathcal{B}$, then the SQP method produces a unique sequence $x^k \in \mathcal{B}$ which is Q-quadratically convergent to x^* (in the norm of X^0). Moreover, from Theorem 4.1 we obtain an estimate for the number of steps needed to achieve a given accuracy.

In order to derive a mesh-independence result from the general theory, we first study the regularity of the SQP sequence for the continuous problem.

LEMMA 5.1. *There exist positive constants p and q such that for every $x^0 \in B_p(x^*)$ with $u^0(\cdot)$ Lipschitz continuous in* $[0,1]$*, for every $k = 1, 2, \ldots,$ and for every $t_1, t_2 \in [0,1]$,*

$$|u^k(t_1) - u^k(t_2)| \le q|t_1 - t_2|.$$

Proof. In [5, section 6], extending a previous result in [7], see also [6], Lemma 2, we showed that the coercivity condition implies pointwise coercivity almost everywhere. In the present circumstances, the latter condition is satisfied everywhere in $[0,1]$; that is, there exists a constant $\alpha > 0$ such that for every $v \in U - U$ and for all $t \in [0,1]$,

$$v^\mathsf{T} R(t) v \ge \alpha v^\mathsf{T} v. \tag{47}$$

For a positive parameter p, consider the SQP sequence x^k starting from $x^0 \in B_p(x^*)$ such that the initial control u^0 is a Lipschitz continuous function in $[0,1]$. Throughout the proof we will choose p sufficiently small and check the dependence of the constants of p. By (46) the iterate x^k satisfies

$$\begin{aligned}(\nabla_u H^k(t) + \nabla^2_{uu} H(x^k(t))(u^{k+1}(t) - u^k(t)) + \nabla^2_{uy} H(x^k(t))(y^{k+1}(t) - y^k(t))\\ +\nabla^2_{u\psi} H(x^k(t))(\psi^{k+1}(t) - \psi^k(t)))^\mathsf{T}(u - u^{k+1}(t)) \ge 0\end{aligned} \tag{48}$$

for every $t \in [0,1]$ and for every $u \in U$. Let $t_1, t_2 \in [0,1]$. Note that x^k are contained in $B_p(x^*)$ for all k and therefore both y'^k and ψ'^k are bounded by a constant independent of k; hence, y^k and ψ^k are Lipschitz continuous functions in time with Lipschitz constants independent of k. We have from (48)

$$\begin{aligned}(\nabla_u H^k(t_1) + \nabla^2_{uu}H(x^k(t_1))(u^{k+1}(t_1) - u^k(t_1)) + \nabla^2_{uy}H(x^k(t_1))(y^{k+1}(t_1) - y^k(t_1))\\ + \nabla^2_{u\psi}H(x^k(t_1))(\psi^{k+1}(t_1) - \psi^k(t_1)))^{\mathsf{T}}(u^{k+1}(t_2) - u^{k+1}(t_1)) \geq 0\end{aligned}$$

and the analogous inequality with t_1 and t_2 interchanged. Adding these two inequalities and adding and subtracting the expressions $\nabla^2_{uu}H(x^k(t_1))u^{k+1}(t_2)$ and $\nabla^2_{uu}H(x^k(t_1))u^*(t_1) - \nabla^2_{uu}H(x^k(t_2))u^*(t_2)$ we obtain

$$\begin{aligned}(\theta^k(t_1) - \theta^k(t_2) - \nabla^2_{uu}H(x^k(t_1))u^*(t_1) + \nabla^2_{uu}H(x^k(t_2))u^*(t_2)\\ +(\nabla^2_{uu}H(x^k(t_1)) - \nabla^2_{uu}H(x^k(t_2)))u^{k+1}(t_2)\\ +\nabla^2_{uy}H(x^k(t_1))(y^{k+1}(t_1) - y^k(t_1))\\ +\nabla^2_{u\psi}H(x^k(t_1))(\psi^{k+1}(t_1) - \psi^k(t_1)))^{\mathsf{T}}(u^{k+1}(t_2) - u^{k+1}(t_1))\\ \geq (\nabla^2_{uu}H(x^k(t_1))(u^{k+1}(t_1) - u^{k+1}(t_2)))^{\mathsf{T}}(u^{k+1}(t_1) - u^{k+1}(t_2))\end{aligned} \tag{49}$$

where the function θ^k is defined as

$$\theta^k(t) = \nabla_u H^k(t) + \nabla^2_{uu}H(x^k(t))(u^k(t) - u^*(t)).$$

By (47), for a sufficiently small p the right-hand side of the inequality (49) satisfies

$$\begin{aligned}(\nabla^2_{uu}H(x^k(t_1))(u^{k+1}(t_1) - u^{k+1}(t_2)))^{\mathsf{T}}(u^{k+1}(t_1) - u^{k+1}(t_2))\\ \geq \frac{\alpha}{2}|u^{k+1}(t_1) - u^{k+1}(t_2)|^2.\end{aligned} \tag{50}$$

Combining (49) and (50) we obtain

$$\begin{aligned}&\frac{\alpha}{2}|u^{k+1}(t_1) - u^{k+1}(t_2)| \leq |\theta^k(t_1) - \theta^k(t_2)|\\ &+|(\nabla^2_{uu}H(x^k(t_1)) - \nabla^2_{uu}H(x^k(t_2)))(u^{k+1}(t_2) - u^*(t_2))|\\ &+|\nabla^2_{uu}H(x^k(t_1))(u^*(t_1) - u^*(t_2))|\\ &+|(\nabla^2_{uy}H(x^k(t_1)) - \nabla^2_{uy}H(x^k(t_2)))(y^{k+1}(t_1) - y^k(t_1))|\\ &+|\nabla^2_{uy}H(x^k(t_2))((y^{k+1}(t_1) - y^{k+1}(t_2)) - (y^k(t_1) - y^k(t_2)))|\\ &+|(\nabla^2_{u\psi}H(x^k(t_1)) - \nabla^2_{u\psi}H(x^k(t_2)))(\psi^{k+1}(t_1) - \psi^k(t_1))|\\ &+|\nabla^2_{u\psi}H(x^k(t_2))((\psi^{k+1}(t_1) - \psi^{k+1}(t_2)) - (\psi^k(t_1) - \psi^k(t_2)))|.\end{aligned} \tag{51}$$

Let u_k be Lipschitz continuous in time with a constant L_k. Then the function θ^k is almost everywhere differentiable and its derivative is given by

$$\begin{aligned}\dot\theta(t) = \nabla^2_{uy}H^k(t)\dot y_k(t) + \nabla^2_{u\psi}H^k(t)\dot\psi_k(t) - \nabla^3_{uuu}H^k(t)\dot u^k(t)(u^k(t) - u^*(t))\\ - \nabla^3_{uuy}H^k(t)\dot y^k(t)(u^k(t) - u^*(t)) - \nabla^3_{uu\psi}H^k(t)\dot\psi^k(t)(u^k(t) - u^*(t)) - \nabla^2_{uu}H^k(t)\dot u^*(t).\end{aligned}$$

From this expression we obtain that there exists a constant c_1, independent of k and t and bounded from above when $p \to 0$, such that

$$\|\dot\theta\|_{L^\infty} \leq cp\|\dot{u^k}\|_{L^\infty} + c_1 \leq c_1(pL_k + 1).$$

Estimating the expressions in the right-hand side of (51) we obtain that there exists a constant c_2, independent of k and t and bounded from above when $p \to 0$, such that

$$|u^{k+1}(t_1) - u^{k+1}(t_2)| \le c_2(pL_k + 1)|t_1 - t_2|.$$

Hence, u^{k+1} is Lipschitz continuous and, for some constants c of the same kind as c_1, c_2, its Lipschitz constant L_{k+1} satisfies

$$L_{k+1} \le c(pL_k + 1).$$

Since p can be chosen arbitrarily small, the sequence $L_i, i = 1, 2, \ldots,$ is bounded, i.e., by a constant q. The proof is complete. □

To apply the general mesh-independence result presented in Theorem 4.2 we need to estimate the residual r_N^k obtained when the SQP sequence of the continuous problem is substituted into the relations determining the SQP sequence of the discretized problem. Specifically, the residual is the remainder term associated with the Euler scheme applied to (43)–(46); that is,

$$r_N^k = \begin{pmatrix} \frac{1}{h}\int_{t_i}^{t_{i+1}} (\nabla_\psi H^k(t) + \nabla^2_{\psi x} H^k(t)(x^{k+1}(t) - x^k(t)) \\ \qquad - (\nabla_\psi H_i^k + \nabla^2_{\psi x} H_i^k(x_i^{k+1} - x_i^k)))dt \\ \frac{1}{h}\int_{t_i}^{t_{i+1}} (-\nabla_x H^k(t) - \nabla^2_{xx} H^k(t)(x^{k+1}(t) - x^k(t)) \\ \qquad - (-\nabla_x H_i^k - \nabla^2_{xx} H_i^k(x_i^{k+1} - x_i^k)))dt \\ \psi^{k+1}(1-h) - \psi^{k+1}(1) \\ 0 \end{pmatrix},$$

where the subscript i denotes the value at t_i. From the regularity of the Newton sequence established in Lemma 5.1, the uniform norm of the residual is bounded by ch, where c is independent of k. Note that the map $\pi_N(x)$ defined in section 4, acting on a function $x \in X^0$, gives the sequence $x(t_i), i = 0, 1, \ldots N$. Condition (23) is satisfied because the space X^0 is a subset of the space of continuous functions. Summarizing, we obtain the following result.

Theorem 5.2. *Suppose that smoothness and coercivity conditions hold. Then there exists a neighborhood $\mathcal{W}$, in the norm of X^0, of the solution $x^* = (y^*, u^*, \psi^*)$ such that for all sufficiently small step-sizes h, the following mesh-independence property holds:*

$$\sup_k \max_{0 \le i \le N-1} |u^k(t_i) - (u_h^k)_i| = O(h), \tag{52}$$

where $u^k(\cdot)$ is the control in the SQP sequence $(y^k(\cdot), u^k(\cdot), \psi^k(\cdot))$ for the continuous problem starting from a point $x^0 = (y^0, u^0, \psi^0) \in \mathcal{W}$ with $u^0(\cdot)$ Lipschitz continuous in $[0,1]$, and u_h^k is the control in the SQP sequence (y_h^k, u_h^k, ψ_h^k) for the discretized problem starting from the point $\pi_N(x^0)$.

Applying Theorem 4.3 to the optimal control problem considered we obtain the mesh-independence property (27) which relates the number of steps for the continuous and the discretized problem needed to achieve certain accuracy. The latter property can be also easily deduced from the estimate (52) in Theorem 5.2, in a way analogous to the proof of Theorem 4.3. Therefore the estimate (52) is a stronger mesh-independence property than (27).

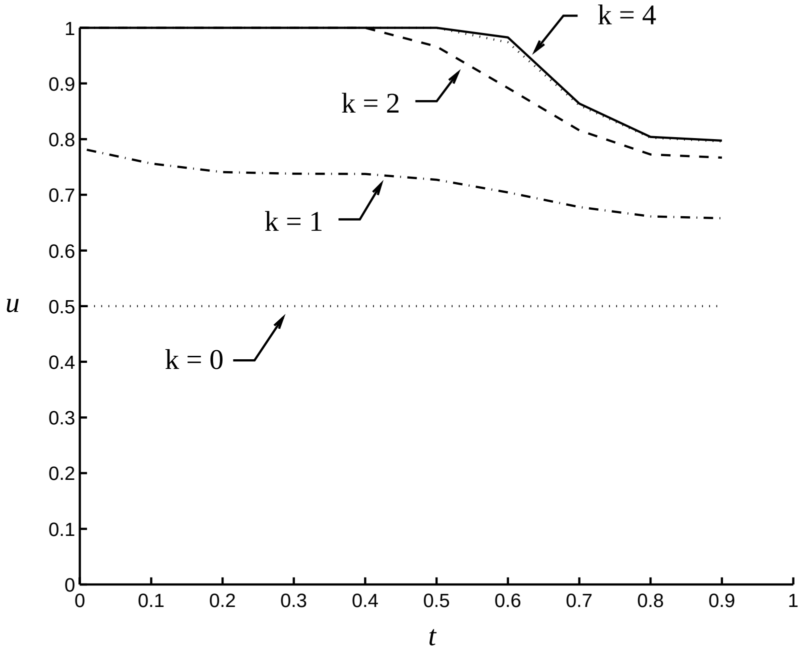


FIG. 1. *SQP iterates for the control with $N = 10$.*

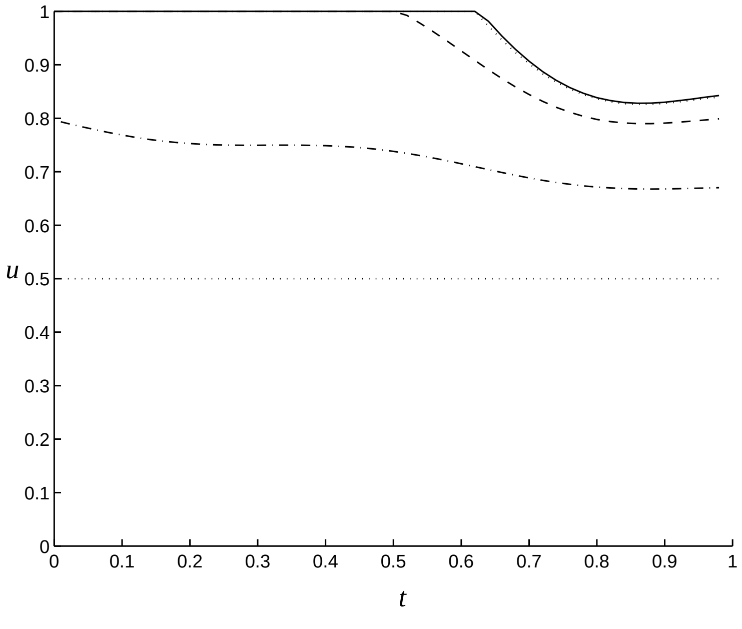


FIG. 2. *SQP iterates for the control with $N = 50$.*

6. Numerical examples. The convergence estimate of Theorem 5.2 is illustrated using the following example:

$$\text{minimize} \int_0^1 \left(\tfrac{1}{2}(y(t)^4 + u(t)^2 + u(t)y(t)) + \tfrac{1}{4}\sin(10t)u(t) + u(t)^{-1}\right) dt$$

$$\text{subject to } \ \dot{y}(t) = -u(t)/(2y(t)), \ y(0) = \sqrt{\tfrac{1+3e}{2(e-1)}}, \ u(t) \leq 1.$$

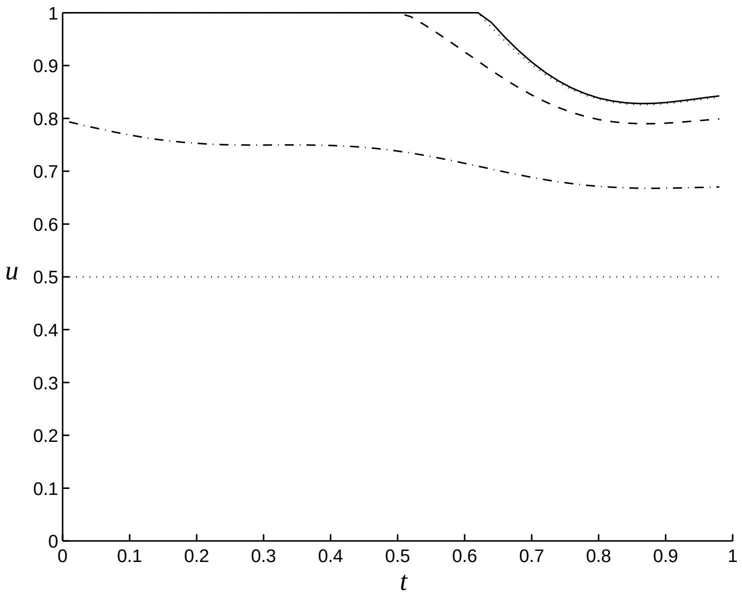


FIG. 3. *SQP iterates for the control with N=250.*

TABLE 1
L^∞ error in the control for various choices of the mesh.

Iteration	$N = 10$	$N = 50$	$N = 250$
0	.500000	.500000	.500000
1	.278473	.290428	.291671
2	.090857	.091727	.097923
3	.008928	.008971	.010185
4	.000082	.000084	.000105

TABLE 2
Error in current iterate divided by error in prior iterate squared.

Iteration	$N = 10$	$N = 50$	$N = 250$
1	1.113	1.161	1.166
2	1.171	1.087	1.151
3	1.081	1.066	1.062
4	1.027	1.039	1.013

This problem is a variation of Problem I in [8] that has been converted from a linear-quadratic problem to a fully nonlinear problem by making the substitution $x = -y^2$ and by adding additional terms to the cost function that degrade the speed of the SQP iteration so that the convergence is readily visible (without these additional terms, the SQP iteration converges to computing precision within 2 iterations). Figures 1–3 show the control iterates for successively finer meshes. The control corresponding to $k = 3$ is barely visible beneath the $k = 4$ iterate. Observe that the SQP iterations are relatively insensitive to the choice of the mesh. Specifically, $N = 10$ is already sufficiently large to obtain mesh independence. In Table 1 we give the L^∞ error in the successive iterates. In Table 2 we observe that the ratio of the error in the current iterate to the error in the prior iterate squared is slightly larger than 1.

REFERENCES

[1] E. L. Allgower, K. Böhmer, F. A. Potra, and W. C. Rheinboldt, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.

[2] W. Alt, *Discretization and mesh-independence of Newton's method for generalized equations*, in Mathematical Programming with Data Perturbations, Lecture Notes in Pure and Appl. Math. 195, Dekker, New York, 1998, pp. 1–30.

[3] W. Alt, and K. Malanowski, *The Lagrange-Newton method for state constrained optimal control problems*, Comput. Optim. Appl., 4 (1995), pp. 217–239.

[4] A. L. Dontchev, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci. 52, Springer-Verlag, Berlin, New York, 1983.

[5] A. L. Dontchev and W. W. Hager, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.

[6] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability and convergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.

[7] J. C. Dunn and T. Tian, *Variants of the Kuhn–Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1384.

[8] W. W. Hager and G. D. Ianculescu, *Dual approximations in optimal control*, SIAM J. Control Optim., 22 (1984), pp. 423–465.

[9] W. W. Hager, *Multiplier methods for nonlinear optimal control*, SIAM J. Numer. Anal., 27 (1990), pp. 1061–1080.

[10] C. T. Kelley and E. W. Sachs, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477–493.

[11] K. Kunisch and S. Volkwein, *Augmented Lagrangian-SQP Techniques and Their Approximations*, in Optimization Methods in Partial Differential Equations (South Hadley, MA, 1996), Contemp. Math. 209, AMS, Providence, RI, 1997, pp. 147–159.

[12] S. Robinson, *Generalized equations*, in Mathematical Programming: The State of the Art (Bonn, 1982), Springer-Verlag, Berlin, New York, 1983, pp. 346–367.

[13] F. Tröltzsch, *An SQP method for the optimal control of a nonlinear heat equation*, Control Cybernet., 23 (1994), pp. 267–288.

[14] S. Volkwein, *Mesh-Independence of an Augmented Lagrangian-SQP Method in Hilbert Spaces and Control Problems for the Burgers Equation*, Ph.D. thesis, TU Berlin, Berlin, Germany, 1997.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 3, pp. 981–988

INTERPOLATION THEORY AND QUANTUM COHOMOLOGY[*]

M. S. RAVI[†]

Abstract. We show that the quantum cohomology ring of the Grassmannian can be used to find the minimal degree of the solution to various interpolation problems involving matrices of rational functions. We also use computations in the quantum cohomology ring to formalize the notion of linearity in this context and distinguish between linear problems such as matrix and tangential interpolation and nonlinear problems such as pole placement.

Key words. interpolation problems, quantum cohomology

AMS subject classifications. 14N35, 65D05, 93B55, 93C35

PII. S0363012997320298

Introduction. A number of rational matrix interpolation problems were reformulated in terms of intersection theory on the space of maps from the projective line $\mathbb{P}^1$ into a Grassmannian in [2]. The general problem posed in [2] was shown to include a wide range of problems of interest, but a solution to the general problem itself was not presented.

The first goal of our paper is to show that the quantum cohomology ring of the Grassmannian provides a general solution to the intersection theory problem posed in [2]. This by itself seems to be a well-known fact among people who deal with quantum cohomology, but we aim to make the connection clear to the interpolation theory community. The second objective of our paper is to use the quantum cohomology ring of the Grassmannian to make specific computations in the cases of the matrix and tangential interpolation problems and to use these computations to show that these problems are "linear" in the sense that they have a unique solution, no solution, or infinitely many solutions. As far as the author knows, there does not seem to be any known explanation of the well-known fact that the matrix and tangential interpolation problems have linear algorithms [1], [8], whereas other problems like the pole placement problem are known to be nonlinear [10]. In the process, we also try to formalize the notion of linearity in this context. We believe that our computations provide the first convincing explanation of the difference between these problems.

1. Preliminaries.

In this section we recall the problem formulated in [2] and introduce some notation.

Let $G = \text{Grass}(p, p+m)$ denote the Grassmannian of all p-dimensional subspaces of $\mathbb{C}^{m+p}$. We recall some of the facts of interest related to the Grassmannian. Each point in G may be identified with the row space of a $p \times (m+p)$ matrix of full row rank. From this point of view, one sees that the Grassmannian has an open cover $\{U_I\}$, where for a choice of a multi-index $I = (i_1, \dots, i_p)$, the open set $U_I \subset G$ consists of subspaces represented by $p \times (m+p)$ matrices whose maximal minor corresponding to the index I is nonzero. Let us fix a flag V of subspaces $V_1 \subset \cdots \subset V_{m+p} = \mathbb{C}^{m+p}$ where $\dim V_i = i$. For any p-tuple of integers $0 < a_1 < \cdots < a_p \leq m+p$, we define

[*] Received by the editors April 28, 1997; accepted for publication (in revised form) May 26, 2000; published electronically October 20, 2000. This work was supported in part by ECU Faculty Senate Research grant 99-41.

http://www.siam.org/journals/sicon/39-3/32029.html

[†] Department of Mathematics, East Carolina University, Greenville, NC 27858 (msravi@math.ecu.edu).

the subvariety

$$\Omega_V(a_1,\dots,a_p)=\{\Lambda\in G\,|\,\dim\Lambda\cap V_{a_i}\geq i \text{ for } 1\leq i\leq p\}.$$

The varieties defined above are called Schubert varieties. The homology class of the subvariety $\Omega_V(a_1,\dots,a_p)$ in the homology ring $H_*(G,\mathbb{C})$ is independent of the flag V and depends only on the index $a=(a_1,\dots,a_p)$, since there is an invertible linear map of $\mathbb{C}^{m+p}$ that takes the flag V to any other flag W, and the set of invertible linear maps of $\mathbb{C}^{m+p}$ is connected. The homology class of $\Omega_V(a_1,\dots,a_p)$ is usually denoted by $(a_1,\dots,a_p)$. It is well known that these homology classes for the various possible choices of the index a form an additive basis for the homology ring of the Grassmannian. The intersection theory of these homology classes in the Grassmannian is usually described in terms of the dual object, namely the cohomology ring, the duality being given by the Poincaré duality map from $H_{2i}(G,\mathbb{C})\to H^{2mp-2i}(G,\mathbb{C})$. Under this duality, the intersection pairing corresponds to the product in the cohomology ring. With this in mind, we next describe a presentation of the cohomology ring of the Grassmannian.

The Grassmannian comes with two canonical vector bundles on it, the subbundle $\mathcal{S}$ that associates to each point $\Lambda\in \text{Grass}(p,p+m)$ the p-dimensional subspace corresponding to Λ, and the quotient bundle $\mathcal{Q}$ that associates to Λ the quotient vector space $\mathbb{C}^{m+p}/\Lambda$. Let $X_i\in H^{2i}(G,\mathbb{C})$ denote the ith Chern class of the subbundle $\mathcal{S}$ and let $Y_i\in H^{2i}(G,\mathbb{C})$ denote the ith Chern class of the quotient bundle $\mathcal{Q}$. Then the cohomology ring $H^*(G,\mathbb{C})$ is the polynomial ring $\mathbb{C}[X_1,\dots,X_p]$ modulo the ideal generated by $(Y_{m+1},\dots,Y_{p+m})$, where the polynomials Y_i are defined as follows: $Y_1=-X_1$ and for $i>1$,

$$Y_i=-Y_{i-1}X_1-\cdots-Y_1X_{i-1}-X_i \tag{1}$$

with the convention that $X_i=0$ for $i>p$. Alternatively, one can describe $H^*(G,\mathbb{C})$ as the polynomial ring $\mathbb{C}[Y_1,\dots,Y_m]$ modulo the ideal generated by $(X_{p+1},\dots,X_{p+m})$, where $X_1=-Y_1$ and for $i>1$,

$$X_i=-X_{i-1}Y_1-\cdots-X_1Y_{i-1}-Y_i$$

with the convention that $Y_i=0$ for $i>m$. Thus, one can switch from a presentation in terms of X_i or Y_i, depending on which is more convenient. The homology class $(a_1,\dots,a_p)$ is Poincaré dual to the cohomology class denoted by $\{m+1-a_1,\dots,m+p-a_p\}$, where

$$\{\lambda_1,\dots,\lambda_p\}=\det(Y_{\lambda_i+j-i})_{1\leq i,j\leq p}.$$

(As before, $Y_j=0$ for $j\notin\{0,1,\dots,m\}$.) For more details on cohomology of Grasmannians, see [7, pp. 193–206].

There are two specific classes of Schubert varieties that will be of interest to us. Let $R_i=\{\Lambda\in G\,|\,\Lambda\subset M_i\}$, where M_i is a fixed subspace of $\mathbb{C}^{m+p}$ of codimension i. The homology class of R_i is $(m{+}1{-}i,m{+}2{-}i,\dots,m{+}p{-}i)$, so that its Poincaré dual is the cohomology class $\{i,\dots,i\}$. In terms of the X_i, one sees that the Poincaré dual of the homology class of $[R_1]$ is the class $(-1)^pX_p$ and the class of $[R_i]$ is Poincaré dual to $(-1)^{pi}X_p^i$. Similarly, let $L_i=\{\Lambda\in G\,|\,N_i\subset\Lambda\}$, where N_i is a fixed subspace of $\mathbb{C}^{m+p}$ of dimension i. The homology class of L_i is $(1,2,\dots,i,m+i+1,\dots,m+p)$, so that its Poincaré dual is the cohomology class $\{m,\dots,m,0,\dots,0\}$, where the m occurs i

times. In terms of the Y_i, the Poincaré dual of the homology class of $[L_1]$ is given by Y_m, and that of $[L_i]$ is given by Y_m^i. The whole Grassmannian can be thought of as the homology class $(m+1,\dots,m+p)$ and the homology class of a point is $(1,\dots,p)$. So the class of a point in the Grassmannian is dual to $(-1)^{mp}X_p^m = Y_m^p$. The Schubert varieties L_i and R_i are called sub-Grassmannians in the algebraic geometry literature.

Given two arbitrary Schubert varieties Ω_1 and Ω_2, in order to determine if their intersection is nonempty, one follows the following procedure. Let ξ_i be the Poincaré duals of the homology classes of Ω_i. If the product $\xi_1 \cdot \xi_2 \in H^*(G,\mathbb{C})$ is nonzero, then one knows that $\Omega_1 \cap \Omega_2$ is nonempty and further the homology class of $\Omega_1 \cap \Omega_2$ is the Poincaré dual of $\xi_1 \cdot \xi_2$, if the intersection is proper. On the other hand, if $\xi_1 \cdot \xi_2 = 0$, and the two Schubert varieties are in "general" position, then their intersection is empty.

We try to develop this intersection theoretic approach to solving interpolation problems involving rational matrix functions in this paper, and show that using the quantum cohomology ring of the Grassmannian, all necessary computations can be made quite easily. Let us first recall the following problem formulated in [2].

Problem 1.1. *Given ℓ points $x_1,\dots,x_\ell$ on the projective line $\mathbb{P}^1$, and ℓ Schubert subvarieties $W_1,\dots,W_\ell$ of Grass$(p,p+m)$, find the lowest integer n for which there is a map of degree n from $\mathbb{P}^1$ into Grass$(p,p+m)$ that maps the point x_i into W_i for $i=1,\dots,\ell$. Further, parameterize the space of all solutions of lowest degree.*

We wish to rephrase this as a problem in intersection theory, and eventually in terms of multiplying cohomology classes. The first step is to identify the space in which the solutions are sought. Let $\mathcal{M}_n$ be the space of all maps ϕ of degree n from $\mathbb{P}^1$ into G and let $\overline{\mathcal{M}_n}$ be the compactification of this space, constructed in [13] and [9]. The space $\overline{\mathcal{M}_n}$ is a smooth, connected, compact manifold and therefore one can use intersection theory on this space. The next step is to identify the subsets of $\overline{\mathcal{M}_n}$ which are being intersected. We observe that for each of the points $x_i \in \mathbb{P}^1$, there is an evaluation map $\mathrm{ev}_i : \mathcal{M}_n \to G$ that takes a map $\phi \in \mathcal{M}_n$ to its value at x_i. Now we want to decide if $\cap_{i=1}^{\ell}\mathrm{ev}_i^{-1}(W_i)$ is empty and more specifically, we want to parameterize the intersection. We also note that by [4, Lemma 2.2A], if $x_1,\dots,x_\ell$ are distinct points and the W_i are in general position, then $\cap_{i=1}^{\ell}\mathrm{ev}_i^{-1}(W_i)$ is Zariski dense in $\cap_{i=1}^{\ell}\overline{\mathrm{ev}_i^{-1}(W_i)}$.

The next step is to identify the cohomology ring of $\overline{\mathcal{M}_n}$, identify the Poincaré duals of the subvarieties $\overline{\mathrm{ev}_i^{-1}(W_i)}$, and compute their product. We note that we are not interested in arbitrary products in the cohomology ring $H^*(\overline{\mathcal{M}_n},\mathbb{C})$ but only of those classes that arise from pullbacks of Schubert varieties through evaluation maps. It turns out that the quantum cohomology ring is set up to do precisely such calculations. Now we need to identify the cohomology classes involved. It would be most convenient to express the cohomology classes of $\mathrm{ev}_i^{-1}(W_i)$ as pullbacks of the Poincaré duals of W_i through the map on cohomology induced by the evaluation maps ev_i. While this description is quite simple, there are some technical problems. The evaluation map is not defined on all of $\overline{\mathcal{M}_n}$, but only on an open set and does not directly induce a map on cohomology rings. We will specify one way of getting around these difficulties in what follows.

The space $\overline{\mathcal{M}_n}$ has a cellular decomposition [13, Theorem 1.3], so its cohomology ring is isomorphic to its Chow ring. Specifically, $H^{2k+1}(\overline{\mathcal{M}_n},\mathbb{C}) = 0$ and $H^{2k}(\overline{\mathcal{M}_n},\mathbb{C}) \simeq A^k(\overline{\mathcal{M}_n},\mathbb{C})$, the group of codimension k cycles on $\overline{\mathcal{M}_n}$ (with complex coefficients) modulo rational equivalence. The evaluation maps can be thought of as rational maps from the compactification $\overline{\mathcal{M}_n}$ to G that are defined on at least the open set $\mathcal{M}_n \subset \overline{\mathcal{M}_n}$. The closure of the graph of this map in $\overline{\mathcal{M}_n} \times G$ is a correspondence, which gives a map $\mathrm{ev}_i^* : A^*(G, \mathbb{C}) \to A^*(\overline{\mathcal{M}_n}, \mathbb{C})$ [6, Chapter 16]. Since we can identify the Chow ring with the cohomology ring, for both $\overline{\mathcal{M}_n}$ and G, we get a map on cohomology: $\mathrm{ev}_i^* : H^*(G, \mathbb{C}) \to H^*(\overline{\mathcal{M}_n}, \mathbb{C})$. We should point out that the maps ev_i^* are *not* ring homomorphisms but just linear maps between vector spaces. The map induced on cohomology by the evaluation map is independent of the point x_i, namely, the maps $\mathrm{ev}_i^* = \mathrm{ev}^*$ for all i [4, Cor. 2.3]. Also, if ξ_i is Poincaré dual to the class of the Schubert variety W_i in G, then $\mathrm{ev}_i^*(\xi_i) = \mathrm{ev}^*(\xi_i)$ is Poincaré dual to $\overline{\mathrm{ev}_i^{-1}(W_i)}$ in $\overline{\mathcal{M}_n}$.

With all of these preliminaries out of the way, Problem 1.1 can be reformulated as follows.

PROBLEM 1.2. *Let ξ_i be the Poincaré dual of the class of the subvariety W_i in $H^*(G, \mathbb{C})$. Find the lowest integer n, such that the product $ev^*(\xi_1) \cdots ev^*(\xi_\ell) \in H^*(\overline{\mathcal{M}_n}, \mathbb{C})$ is not zero and find the product for such an n.*

The answer to the question posed above will be given in terms of the quantum cohomology ring of the Grassmannian.

We will first recall the definition of the quantum cohomology ring, as given in [3]. Recall that there is a nondegenerate inner product $(\cdot \mid \cdot)$ on $H^*(G, \mathbb{C})$ defined as follows. For ξ and $\eta \in H^*(G, \mathbb{C})$, define

$$(\xi \,|\, \eta) = \int_G \xi\eta.$$

This inner product is dual, via Poincaré duality, to the intersection pairing on homology. Note that if ξ is dual to η with respect to the inner product $(\cdot \mid \cdot)$, then $\deg \eta + \deg \xi = \dim G = mp$ (all references to dimension and codimension in this paper will be to complex dimension and complex codimension). For future reference we note that the dual of the cohomology class Y_m^i with respect to the inner product $(\cdot \,|\, \cdot)$ is Y_m^{p-i} and the dual of $(-1)^i X_p^i$ is $(-1)^{m-i} X_p^{m-i}$.

Now, given two cohomology classes ξ and η as above, their quantum product $\xi \star \eta$ is defined as follows:

$$\xi \star \eta = (\xi \star \eta)_0 + q(\xi \star \eta)_1 + \cdots q^j (\xi \star \eta)_j + \cdots ,$$

where

$$((\xi \star \eta)_j \,|\, \alpha) = \int_{\overline{\mathcal{M}_j}} \mathrm{ev}^*(\xi)\mathrm{ev}^*(\eta)\mathrm{ev}^*(\alpha) \tag{2}$$

In particular, $(\xi \star \eta)_0 = \xi \cdot \eta$ is the usual cohomology product. Further, since the product in (2) is zero unless $\deg \xi + \deg \eta + \deg \alpha = \dim \overline{\mathcal{M}_j} = mp + (m+p)j$, the series in (2) is a polynomial with only finitely many nonzero terms and the degree of $(\xi \star \eta)_j$ is $\deg \xi + \deg \eta - (m+p)j$.

Let $W_1, \ldots, W_{\ell+1}$ be subvarieties of the Grassmannian in general position, and let ξ_i be the cohomology class in $H^*(G, \mathbb{C})$ that is Poincaré dual to W_i. Suppose $\xi_1 \star \cdots \star \xi_\ell = \sum_{d \geq 0} q^d \alpha_d$ and $(\alpha_d \,|\, \xi_{\ell+1}) \neq 0$. Then if the points x_i are distinct, the intersection $\cap_{i=1}^{\ell+1} \overline{\mathrm{ev}_i^{-1}(W_i)}$ in $\overline{\mathcal{M}_d}$ is nonempty. Further, $\cap_{i=1}^{\ell} \mathrm{ev}_i^{-1}(W_i) \subset \mathcal{M}_d$ is also nonempty.

PROPOSITION 1.3. *Let ξ_i be the Poincaré dual of the class of the subvarieties $W_i \subset G$. Let the quantum product*

$$\xi_1 \star \cdots \star \xi_\ell = \sum_{d \geq 0} q^d \alpha_d,$$

where $\alpha_d \in H^*(G,\mathbb{C})$, and let n be the smallest integer such that $\alpha_n \neq 0$. Then, for subvarieties W_i in general position and n distinct points $x_i \in \mathbb{P}^1$, the lowest degree of a curve $\phi : \mathbb{P}^1 \to G$ such that $\phi(x_i) \in W_i$ is equal to n.

The final piece of the puzzle comes from an explicit characterization of the quantum cohomology ring of the Grassmannian.

THEOREM 1.4 (see [12]). *The quantum cohomology ring $QH^*(G)$ of the Grassmannian is the polynomial ring $\mathbb{C}[X_1,\dots,X_p,q]$ modulo the ideal generated by $(Y_{m+1},\dots,Y_{m+p}+(-1)^p q)$, where the Y_i are as defined in* (1).

We will denote the i-fold quantum product $\xi \star \cdots \star \xi$ by $\xi^{\star i}$ and the regular i-fold cohomology product $\xi \cdots \xi$ by ξ^i, as we have already used this notation in the introduction where we identified the Poincaré duals of $[R_i]$ and $[L_i]$.

LEMMA 1.5. *For $1 \le i \le p$, $Y_m^{\star i} = Y_m^i$, and for $1 \le j \le m$, $X_p^{\star i} = X_p^i$.*

Proof. We will use the quantum Pieri formula proved in [4]. The particular case of the formula that we need can be stated as follows:

$$\{m,0,\dots,0\} \star \{\lambda_1,\dots,\lambda_p\} = \{m,0,\dots,0\} \cdot \{\lambda_1,\dots,\lambda_p\} + q\left(\sum \{\mu_1,\dots,\mu_p\}\right), \tag{3}$$

where the last sum ranges over all p-tuples μ that satisfy the following conditions:

$$\sum \mu_i = \left(\sum \lambda_i\right) - p \text{ and } \lambda_1 - 1 \ge \mu_1 \ge \lambda_2 - 1 \ge \mu_2 \cdots \ge \lambda_p - 1 \ge \mu_p \ge 0. \tag{4}$$

As we have noted before $Y_m^i = \{\lambda_1,\dots,\lambda_p\}$, where $\lambda_1 = \cdots = \lambda_i = m$ and $\lambda_{i+1} = \cdots = \lambda_p = 0$. Therefore for $i \le p-1$, the quantum product $Y_m \star Y_m^i = Y_m^{i+1}$, since there are no p-tuples μ that can satisfy the condition $-1 = \lambda_p - 1 \ge \mu_p \ge 0$.

Let $G' = \text{Grass}(m, p+m)$ be the Grassmannian of m-dimensional planes in $\mathbb{C}^{p+m}$. Then $QH^*(G') = \mathbb{C}[\sigma_1,\dots,\sigma_m]/(\tau_{p+1},\dots,\tau_{p+m}+(-1)^m q)$, where the τ_i are defined analogously to the Y_i in (1), namely,

$$\tau_i = -\tau_{i-1}\sigma_1 - \cdots - \tau_1\sigma_{i-1} - \sigma_i.$$

Then, according to [5, Proposition 4.1], the map from $QH^*(G)$ to $QH^*(G')$ that takes X_i to τ_i is an isomorphism. Now, it follows from the first half of the proof that $\tau_p^{\star i} = \tau_p^i$ for $1 \le i \le m$. Thus $X_p^{\star i} = X_p^i$ for $1 \le i \le m$. □

2. Interpolation problems. In this section we use the machinery set up in the introduction to solve the matrix interpolation and the left and right tangential interpolation problems. First we briefly recall the statement of these problems, along with their reformulation in [2] as special cases of Problem 1.1, and then their reformulation in terms of quantum cohomology.

2.1. Matrix interpolation problem. Given ℓ points $x_1,\dots,x_\ell \in \mathbb{P}^1$ and $p \times m$ scalar matrices $Z_1,\dots,Z_\ell$, we wish to find a rational matrix of least McMillan degree $Z(s)$ such that $Z(x_i) = Z_i$. The solution to this problem is known; see [1] and [8].

If $Z(s) = D_L^{-1}N_L$ is a left coprime factorization of $Z(s)$, then the condition that $Z(x_i) = Z_i$ can be restated as $(D_L(s) \,|\, N_L(s))(x_i) = (I_p \,|\, Z_i)$. Thus as reformulated in [2], the solutions to this problem consist of maps ϕ from $\mathbb{P}^1$ into G, such that $\phi(x_i) = (I_p \,|\, Z_i)$ for $i = 1,\dots,\ell$ and the McMillan degree of $Z(s)$ equals the degree of the map ϕ from $\mathbb{P}^1$ to G. The conditions being imposed at each point x_i are that the evaluation at that point should be a specified point in the Grassmannian. As seen before the Poincaré dual of the class of a point is Y_m^p. Thus according to

Proposition 1.3, the quantum product that represents the solution to this problem is $(Y_m^p)^{\star\ell}$. By Lemma 1.5 this is equal to $Y_m^{\star p\ell}$.

We will now compute this product using our description of the quantum cohomology ring. First we let $p\ell = a(m+p)+b$, where $a \geq 0$ and $0 \leq b < m+p$. Then $Y_m^{\star p\ell} = (Y_m^{\star(m+p)})^{\star a} Y_m^{\star b}$. Now since $Y_m^{\star p} = Y_m^p = (-1)^{mp}(X_p)^m = (-1)^{mp} X_p^{\star m}$, we have $Y_m^{\star(m+p)} = (-1)^{mp} Y_m^{\star m} \star X_p^{\star m}$. The last relation in the ideal of the quantum cohomology ring implies that $X_p \star Y_m = (-1)^p q$. Therefore $Y_m^{\star(m+p)} = (-1)^{2mp} q^m = q^m$. Finally, if $p < b < m+p$, then we can write

$$\begin{aligned} Y_m^{\star b} &= Y_m^{\star p} Y_m^{\star(b-p)} = (-1)^{mp} X_p^{\star m} \star Y_m^{\star(b-p)} = (-1)^{mp+p(b-p)} q^{b-p} X_p^{\star(m+p-b)} \\ &= ((-1)^p X_p)^{\star(m+p-b)} q^{b-p}. \end{aligned}$$

Putting all of this together,

$$Y_m^{\star p\ell} = \begin{cases} q^{am} Y_m^b & \text{if } b \leq p, \\ q^{am+b-p}((-1)^p X_p)^{m+p-b} & \text{if } p < b < m+p. \end{cases}$$

In either of these cases, it is easy to verify that the exponent of q can also be characterized as the smallest integer n such that the dimension of $\overline{\mathcal{M}_n}$ which is $n(m+p)+mp$ is greater than or equal to $mp\ell$, which can be thought of as the "number of conditions imposed."

2.2. Right tangential interpolation. Here we are given ℓ distinct points, $x_1, \dots, x_\ell \in \mathbb{C}$ and $m \times s_i$ matrices W_i and $p \times s_i$ matrices Z_i, such that the rank of $W_i = s_i \leq m$. We wish to find the lowest degree of all $p \times m$ rational matrices $Z(s)$ such that $Z(x_i)W_i = M_i$ for $i = 1, \dots, \ell$. If $Z = D_L^{-1} N_L$ is a left coprime factorization of the matrix $Z(s)$, then one can rewrite the interpolating condition as $N_L(x_i)W_i = D_L(x_i)Z_i$. Thus in terms of curves from $\mathbb{P}^1$ into G, one is looking for a map ϕ such that $\phi(x_i) \subset M_i$, where M_i is a linear subspace of $\mathbb{C}^{m+p}$ of codimension s_i, defined by $M_i \cdot \left(\begin{smallmatrix} W_i \\ -Z_i \end{smallmatrix} \right) = 0$. Now the Poincaré dual of the Schubert variety R_i, consisting of all points in G whose span is contained in M_i, is $(-1)^{ps_i} X_p^{s_i}$. So the quantum product to be computed here is $(-1)^{ps_1} X_p^{s_1} \star \cdots \star (-1)^{ps_\ell} X_p^{s_\ell}$ which by Lemma 1.5 is $(-1)^{ps} X_p^{\star s}$, where $s = \sum s_i$.

As in the previous section, we start by writing $s = a(m+p)+b$, where $0 \leq b < m+p$. Further, $(-1)^{p(m+p)} X_p^{\star(m+p)} = (-1)^{p^2} X_p^p \star Y_m^p = q^p$ by a calculation similar to the last section. If $m < b < m+p$, we can write

$$\begin{aligned} (-1)^{pb} X_p^{\star b} &= (-1)^{pb} X_p^m \star X_p^{b-m} = (-1)^{p(b-m)} Y_m^p \star X_p^{b-m} = (-1)^{2p(b-m)} q^{b-m} Y_m^{m+p-b} \\ &= q^{b-m} Y_m^{m+p-b}. \end{aligned}$$

Thus, as in the previous section we get two cases:

$$(-1)^{ps} X_p^{\star s} = \begin{cases} q^a (-1)^{pb} X_p^b & \text{if } b \leq m, \\ q^{a+b-m} Y_m^{m+p-b} & \text{if } m < b < m+p. \end{cases} \tag{5}$$

One can again characterize the exponent of q in either of these cases as the smallest integer n such that $n(m+p)+mp \geq ps$.

2.3. Left tangential interpolation. Here we are given ℓ distinct points, $x_1, \dots, x_\ell \in k$ and $r_i \times p$ matrices V_i and $r_i \times m$ matrices Z_i, such that rank of $V_i = r_i \leq p$. We wish to find the lowest degree of a $p \times m$ rational matrix $Z(s)$ such that $V_i Z(x_i) = Z_i$ for $i = 1, \dots, \ell$. As in [2], if we let $Z(s) = N_R D_R^{-1}$ be the right coprime factorization, then the required condition can be expressed as $(V_i \mid Z_i) \cdot \left(\begin{smallmatrix} N_R \\ -D_R \end{smallmatrix}\right)(x_i) = 0$. The relation $(D_L \mid N_L) \cdot \left(\begin{smallmatrix} N_R \\ -D_R \end{smallmatrix}\right) = 0$ between the left and right coprime factorizations implies that one wants the row span N_i of $(V_i \mid Z_i)$ to be contained in the row span of $(D_L \mid N_L)(x_i)$, where N_i is a subspace of $\mathbb{C}^{m+p}$ of dimension r_i. The Poincaré dual of the Schubert variety L_i of all points in G whose span contains N_i is $Y_m^{r_i}$. The quantum product to be computed here is $Y_m^{r_1} \star \cdots \star Y_m^{r_\ell}$ which by Lemma 1.5 is $Y_m^{\star r}$, where $r = \sum r_i$.

As in section 2.1 if one writes $r = a(m+p) + b$, where $0 \leq b < m+p$, then

$$Y_m^{\star r} = \begin{cases} q^a Y_m^b & \text{if } b \leq p, \\ q^{a+b-p}((-1)^p X_p)^{m+p-b} & \text{if } m+p > b > p. \end{cases} \tag{6}$$

The calculations that we have made in each of the three cases above can all be put together in the following lemma. The lemma is of independent interest because it deals with the quantum computation needed for the bitangential interpolation problem, and also provides a justification for the definition of linear interpolation problems.

Lemma 2.1. *The product $(-1)^{pa} X_p^{\star a} \star Y_m^{\star b}$ in the quantum cohomology ring $QH^*(G)$ can be rewritten as a power of q times $(-1)^{pc} X_p^c$, where $c \leq m$, or as a power of q times Y_m^d, where $d \leq p$.*

Proof. The last relation in the presentation of the quantum cohomology ring given in Theorem 1.4 specifies that $X_p \star Y_m = (-1)^p q$ in the quantum cohomology ring. Using this relation, one can rewrite $(-1)^{pa} X_p^{\star a} \star Y_m^{\star b}$ as $(-1)^{2pa} q^a Y_m^{\star(b-a)} = q^a Y_m^{\star(b-a)}$, if $b > a$ and as $(-1)^{p(a+b)} q^b X_p^{\star(a-b)} = (-1)^{p(a-b)} q^b X_p^{\star(a-b)}$ if $a \geq b$. As we have seen in (5) and (6), any quantum power of Y_m or $(-1)^p X_p$ can be rewritten as claimed in the lemma. □

Given an interpolation problem of the form given in Problem 1.1, where the interpolating conditions can be expressed in terms of the cohomology classes Y_m and $(-1)^p X_p$ alone, the lemma above shows that there exists a unique interpolating condition, corresponding to a cohomology class which can also be given in terms of these cohomology classes, namely $(-1)^{p(m-c)} X_p^{m-c}$ or Y_m^{p-d}, such that imposing this additional condition on the original interpolation problem yields a unique solution of minimal degree. In this sense the solution set to the original interpolation problem is "weakly dual" to that of the interpolation problem corresponding to the condition given by the classes $(-1)^{p(m-c)} X_p^{m-c}$ or Y_m^{p-d}. A stronger duality statement, which we had hoped to prove at first, would be to say that the cohomology class corresponding to the product of the interpolating conditions is dual in the cohomology ring $H^*(\overline{\mathcal{M}_n}, \mathbb{C})$ to the pullback of $(-1)^{p(m-c)} X_p^{m-c}$ or Y_m^{p-d}, where n is the minimal degree of the interpolants, and the duality is taken with respect to the inner product $(\cdot \mid \cdot)$ in $H^*(\overline{\mathcal{M}_n}, \mathbb{C})$. We are unable to prove this statement. The problem arises because the product of cohomology classes of the interpolating conditions in $H^*(\overline{\mathcal{M}_n}, \mathbb{C})$ might contain some boundary components that are not in general identifiable using quantum products which only compute the products of pullbacks of classes from $H^*(G, \mathbb{C})$.

The matrix interpolation and the tangential problems form a distinguished class of interpolation conditions in the following sense: given any problem in this class the solution set is weakly dual to that of another interpolation problem from this same class. So we can call the interpolation conditions from this class as linear conditions. This should be contrasted with the pole placement problem which is equivalent to finding maps ϕ from $\mathbb{P}^1$ to G such that at the specified points x_i the span of $\phi(x_i)$ has a nontrivial intersection with a specified m-plane Λ_i. One checks easily that the quantum cohomology product that needs to be computed in this case is $(-X_1)^{\star\ell}$, where ℓ is the number of points at which this condition is imposed. The term of lowest degree in q in this product was computed explicitly in [10] and [11]. The main point of interest for the present discussion is that the product $(-X_1)^{\star\ell}$ has terms of multiplicity greater than one, and in particular for the values of ℓ for which there are only finitely many solutions of minimal degree, the solutions are not unique. Hence these interpolation conditions are *nonlinear* by any reasonable definition of the term.

An important question in this context, which we cannot answer at this point, is the following: Given that the intersection multiplicity of a class of interpolation problems is always one, does the existence of a linear algorithm necessarily follow? In other words, can one find a linear algorithm based on the intersection computations alone?

Acknowledgments. We would like to thank the anonymous referees for a thorough reading and their constructive criticisms.

REFERENCES

[1] A. C. Antoulas, J. A. Ball, J. Kang, and J. C. Willems, *On the solution of the minimal rational interpolation problem*, Linear Algebra Appl., 137/138 (1990), pp. 511–573.

[2] J. Ball and J. Rosenthal, *Pole placement, internal stabilization and interpolation conditions for rational matrix functions: A Grassmannian formulation*, in Linear Algebra for Control Theory, P. Van Dooren and B. Wyman, eds., IMA Vol. Math. Appl. 62, Springer-Verlag, New York, 1994, pp. 21–29.

[3] A. Beauville, *Quantum cohomology of complete intersections*, Mat. Fiz. Anal. Geom., 2 (1995), pp. 384–398.

[4] A. Bertram, *Quantum Schubert calculus*, Adv. Math., 128 (1997), pp. 289–305.

[5] A. Bertram, I. Ciocan-Fontanine, and W. Fulton, *Quantum multiplication of Schur polynomials*, J. Algebra, 219 (1999), pp. 728–746.

[6] W. Fulton, *Intersection Theory*, Ergeb. Math. Grenzgeb. (3), Vol. 2, 2nd ed., Springer-Verlag, Berlin, 1998.

[7] P. Griffiths and J. Harris, *Principles of Algebraic Geometry*, John Wiley, New York, 1978.

[8] M. S. Ravi, *Geometric methods in rational interpolation theory*, Linear Algebra Appl., 258 (1997), pp. 159–168.

[9] M. S. Ravi and J. Rosenthal, *A smooth compactification of the space of transfer functions with fixed McMillan degree*, Acta Appl. Math, 34 (1994), pp. 329–352.

[10] M. S. Ravi, J. Rosenthal, and X. Wang, *Dynamic pole assignment and Schubert calculus*, SIAM J. Control Optim., 34 (1996), pp. 813–832.

[11] M. S. Ravi, J. Rosenthal, and X. Wang, *Degree of the generalized Plücker embedding of a quot scheme and quantum cohomology*, Math. Ann., 311 (1998), pp. 11–26.

[12] B. Siebert and G. Tian, *On quantum cohomology rings of Fano manifolds and a formula of Vafa and Intriligator*, Asian J. Math., 1 (1997), pp. 679–695.

[13] S. A. Stromme, *On parameterized rational curves in Grassmann varieties*, in Space Curves, Lecture Notes in Math. 1266, Springer-Verlag, Berlin, New York, 1987, pp. 251–272.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 989–1007

DEGENERATE OPTIMAL CONTROL PROBLEMS WITH STATE CONSTRAINTS*

FRANCO RAMPAZZO[†] AND RICHARD VINTER[‡]

Abstract. Standard necessary conditions for optimal control problems with pathwise state constraints supply no useful information about minimizers in a number of cases of interest, e.g., when the left endpoint of state trajectories is fixed at x_0 and x_0 lies in the boundary of the state constraint set; in these cases a nonzero, but nevertheless trivial, set of multipliers exists. We give conditions for the existence of nontrivial multipliers. A feature of these conditions is that they allow nonconvex velocity sets and measurably time-dependent data. The proof techniques are based on refined estimates of the distance of a given state trajectory from the set of state trajectories satisfying the state constraint, originating in the dynamic programming literature.

Key words. optimal control, necessary conditions, nonsmooth analysis, state constraints

AMS subject classifications. 49K15, 49K24

PII. S0363012998340223

1. Introduction.

Consider the optimal control problem

$$
\text{(P)} \qquad \begin{cases} \text{Minimize } g(x(0), x(1)) \\ \text{over } x \in W^{1,1}([0,1]; R^n) \text{ satisfying} \\ \dot{x}(t) \in F(t, x(t)) \text{ almost everywhere (a.e.),} \\ (x(0), x(1)) \in C_0 \times C_1, \\ x(t) \in A \text{ for all } t \in [0,1] \end{cases}
$$

for which the data comprises a function $g : R^n \times R^n \to R$, closed sets A, C_0, and C_1 in R^n, and a multifunction $F : [0,1] \times R^n \rightsquigarrow R^n$. Here, $W^{1,1}([a,b]; R^n)$ denotes the Banach space of absolutely continuous R^n valued functions on the interval $[a,b]$, with norm

$$
||x||_{W^{1,1}} := |x(a)| + \int_a^b |\dot{x}(t)| dt.
$$

Interest centers on the presence of the state constraint $x(t) \in A$. We shall assume that A is expressible as the set of points satisfying a finite number of functional inequality constraints

$$
A = \bigcap_{j=1}^{m} \{x \in R^n \, : \, h_j(x) \le 0\}.
$$

Here, $h_j : R^n \to R$, $j = 1, \dots, m$, are given functions of class $C^{1,1}$ (functions which are continuously differentiable with locally Lipschitz continuous derivatives). For

*Received by the editors June 12, 1998; accepted for publication (in revised form) May 20, 2000; published electronically November 8, 2000. This research was supported by the Human Capital and Mobility Program of the European Community (contract CHTX-CY94-0431).

http://www.siam.org/journals/sicon/39-4/34022.html

[†] Dipartimento di Matematica Pura e Applicata, Università di Padova, 35131 Padova, Italy (rampazzo@pdmat1.unipd.it).

[‡] Centre for Process Systems Engineering and Department of Electrical and Electronic Engineering, Imperial College, Exhibition Road, London SW7 2BT, UK (r.vinter@ic.ac.uk).

simplicity of exposition in this introduction we restrict attention to the case when $m = 1$.

Let $\bar{x}$ be a $W^{1,1}$ local minimizer. This means that $\bar{x}$ is a $W^{1,1}([0,1]; R^n)$ function which satisfies the constraints of (P) and there exists $\delta > 0$ such that, for all $x \in W^{1,1}([0,1]; R^n)$ satisfying the constraints of (P) and also the condition

$$||x - \bar{x}||_{W^{1,1}([0,1];R^n)} \leq \delta,$$

we have

$$g(x(0), x(1)) \geq g(\bar{x}(0), \bar{x}(1)).$$

Necessary conditions for this "differential inclusion problem" with state constraints have been known for many years. (Many advances in this area were prefigured by the work of A. Ja. Dubovitskii and Milyutin in the early 1960s on problems with "mixed" constraints, e.g., [12]. Later publications, making extensive use of nonsmooth analysis, include [9], [26], [19], [30].) The following conditions, which can be deduced from those in [19] or [30], are typical. Under suitable hypotheses on the data for problem (P) (which include the hypothesis that F is convex valued), there exist $p \in W^{1,1}([0,1]; R^n)$, a nonnegative Borel measure $\mu \in C^*([0,1]; R)$, and $\lambda \geq 0$ such that

$$\begin{aligned}
&\lambda + \int_{[0,1]} \mu(ds) + ||p||_{L^\infty} \neq 0, \\
&\text{supp } \{\mu\} \subset \{t : h_1(\bar{x}(t)) = 0\}, \\
(1.1) \quad &-\dot{p}(t) \in \\
&\quad \text{co} \left\{ q : (q, \dot{\bar{x}}(t)) \in \partial H \left(t, \bar{x}(t), p(t) + \int_{[0,t)} \nabla h_1(\bar{x}(t))\mu(ds) \right) \right\} \quad \text{a.e.}, \\
&\left(p(0), -\left[p(1) + \int_{[0,1]} \nabla h_1(\bar{x}(t))\mu(ds) \right] \right) \\
&\quad \in \lambda \partial g(\bar{x}(0), \bar{x}(1)) + N_{C_0}(\bar{x}(0)) \times N_{C_1}(\bar{x}(1)).
\end{aligned}$$

Here, ∂H is the limiting subdifferential of the Hamiltonian

$$H(t, x, p) := \max_{v \in F(t,x)} p \cdot v$$

with respect to the (x, p) variables. $N_S(y)$ denotes the limiting normal cone of the closed set S at the point $y \in S$. $\partial g(y)$ is the limiting subdifferential of g at y. (These constructs are defined below.)

Necessary conditions for optimal control problems with state constraints, in which the dynamic constraint takes the traditional form of a parameterized family of differential equations, have an even longer history. (See, for example, [11], [22], [31].)

Now, necessary conditions such as those above convey no useful information about minimizers in certain important special cases. Consider, for example,

$$C_0 = \{x_0\} \text{ and } h_1(x_0) = 0$$

for some point $x_0 \in R^n$ such that $\nabla h_1(x_0) \neq 0$ (the case when the left endpoint is fixed at x_0 , and x_0 lies in the boundary of the state constraint set). Then the above conditions are satisfied with the nonzero multiplier set

$$(1.2) \qquad \lambda = 0, \ \mu = \delta_{\{0\}}, \ \text{ and } \ p(.) \equiv -\nabla h_1(x_0)$$

for *any* arc $\bar{x}$ which satisfies the constraints of problem (P). Here $\delta_{\{0\}}$ is the unit measure concentrated at $\{0\}$.

There is a growing literature, nondegenerate necessary conditions, which aims at supplying useful information about minimizers in cases such as that just described. (Early papers were [7], [3], [10].) Refinements of the standard necessary conditions have been proved which assert that, under a suitable constraint qualification concerning the dynamic constraint and the state constraint, the existence of a multiplier set (λ, μ, p) *distinct* from the trivial multiplier set (1.2) is guaranteed. We note, in particular, recent advances due to Arutyunov, Aseev, and Blagodat-Skikh [4], Arutyunov and Aseer [5], [6], and Aseev [8], in which the traditional conditions are supplemented with a new boundary condition on the Hamiltonian which, when coupled with a constraint qualification, excludes the trivial multiplier set.

For the most part, nondegenerate necessary conditions for problems with general endpoint constraints and nonsmooth data have been derived under hypotheses which require the following.

(i) $F(t, x)$ is Lipschitz continuous with respect to both t and x variables.

(ii) F takes values convex sets.

The proof techniques of [4], for example, depend critically on such hypotheses. This is because time is treated as a state-like variable (and the more stringent hypotheses governing x-dependence of F for purposes of deriving necessary conditions must therefore be extended to t-dependence) and because the perturbational analysis, based on weak convergence of velocities, breaks down when F is not convex valued.

The contribution of this paper is a methodology for deriving nondegenerate necessary conditions under hypotheses *which apply to problems with general endpoint constraints and which at the same time allow measurable time dependence and also nonconvex F's.* The methodology is to exploit estimates on the distance of a given state trajectory (which violates the state constraint) from the set of state trajectories which satisfy the state constraint. The original motivation for deriving estimates of this nature was their relevance to dynamic programming, specifically their role in proving regularity properties of value functions for optimal control problems with state constraints. The main contribution of this paper is to demonstrate the significance of these estimates also to the derivation of necessary conditions.

Different methods were used in [15] and [16] to derive nondegenerate necessary conditions for nonconvex problems, based on singular transformations and state constraint relaxation at the endpoints, respectively. The optimality conditions of [16] are formulated for a more general class of state constraints than that considered here. However, the present paper improves on these earlier results, in respect of the constraint qualifications invoked to validate optimality conditions. The constraint qualifications of [15] and [16] are expressed in terms of the minimizer we seek and therefore are not, in general, open to direct verification. The constraint qualification of this paper, like that of [4], is a simple hypothesis on the problem data, requiring the existence of inward pointing admissible velocities at all relevant points of the state constraint boundary.

Throughout, B will denote the closed unit ball in Euclidean space. We shall write $d_S : R^k \to R$ for the Euclidean distance function from a set $S \subset R^k$:

$$d_S(x) := \inf \{|x - y| : y \in S\}.$$

The *limiting normal cone* $N_S(x)$ of a closed set $S \subset R^k$ at $x \in S$ is defined to be

$$N_S(x) := \{\xi \,:\, \exists \text{ sequences } \{M_i\} \text{ in } (0, \infty),\, x_i \to x,\, \xi_i \to \xi \quad \text{such that}$$

$$x_i \in S \text{ and } \xi_i \cdot (y - x_i) \le M|y - x_i|^2 \quad \text{for all } y \in R^k,\ i = 1, 2, \ldots\}.$$

Given a lower semicontinuous function $f : R^k \to R \cup \{\infty\}$, the *limiting subdifferential* of f at a point $x \in R^k$ such that $f(x) < +\infty$ is the set

$$\partial f(x) := \{\xi : (\xi, -1) \in N_{\text{epi}\, f}(x, f(x))\}\,.$$

Here, epi f denotes the epigraph set of the function f. Properties of these constructs from nonsmooth analysis are reviewed in detail in [25] and [28]. See also [18], [20].

2. Nondegenerate necessary conditions. In this section are stated the main results of the paper, namely nondegenerate necessary conditions for problem (P), covering cases when the velocity set $F(t, x)$ is nonconvex valued, and measurably time-dependent.

We shall assume, as in the preceding section, that the state constraint set A takes the form

$$A = \bigcap_{j=1}^{m} \{x : h_j(x) \le 0\}$$

for some functions $h_j : R^n \to R$, $j = 1, \ldots, m$, of class $C^{1,1}$, but now we allow $m > 1$.

THEOREM 2.1. *Let $\bar{x}$ be a $W^{1,1}$ local minimizer for problem* (P). *Assume that, for some $k_F \in L^1$ and positive constants δ, ϵ, c, and r, the following hypotheses are satisfied.*

(H1) *F has values closed sets, $F(., x)$ is measurable for each $x \in R^n$,*

$$\begin{aligned} &F(t, x) \subset cB \quad \text{for all } x \in \bar{x}(t) + \delta B,\ t \in [0, 1], and \\ &F(t, x) \subset F(t, x') + k_F(t)|x - x'|B \quad \text{for all } x, x' \in \bar{x}(t) + \delta B, t \in [0, 1]. \end{aligned}$$

(H2) *g is Lipschitz continuous on $(\bar{x}(0), \bar{x}(1)) + \delta(B \times B)$.*

We impose, furthermore, the following constraint qualification.

(CQ) *For each $t \in [0, \epsilon]$ and $\xi \in \bar{x}(0) + \delta B$,*

$$\min\nolimits_{v \in F(t,\xi)} \nabla h_j(\xi) \cdot v < -r \tag{2.1}$$

for all index values j such that $h_j(\bar{x}(0)) = 0\}$.

Then there exist $p \in W^{1,1}([0, 1]; R^n)$, $\lambda \ge 0$, and nonnegative Borel measures $\mu_j \in C^*([0, 1]; R^n)$, $j = 1, \ldots, m$, such that

$$\lambda + \int_{(0,1]} \sum_j \mu_j(ds) + \left| p(0) + \sum_j \nabla h_j(\bar{x}(0))\mu_j(\{0\}) \right| \ne 0,$$

$$\begin{aligned} \dot{p}(t) \in co \Bigg\{ q : \Bigg(q, p(t) + \int_{[0,t)} \sum_j \nabla h_j(\bar{x}(s))\mu_j(ds) \Bigg) \\ \in N_{\text{Gr}F(t,.)}(\bar{x}(t), \dot{\bar{x}}(t)) \Bigg\} \quad a.e., \end{aligned} \tag{2.2}$$

$$\begin{aligned} \Bigg(p(0), -\Bigg(p(1) + \int_{[0,1]} \sum_j \nabla h_j(\bar{x}(s))\mu_j(ds) \Bigg)\Bigg) \\ \in \lambda \partial g(\bar{x}(0), \bar{x}(1)) + N_{C_0 \cap A}(\bar{x}(0)) \times N_{C_1}(\bar{x}(1)), \end{aligned}$$

$$\left(p(t)+\int_{[0,t)}\sum_j \nabla h_j(\bar{x}(s))\mu_j(ds)\right)\cdot\dot{\bar{x}}(t) \tag{2.3}$$
$$=\max_{v\in F(t,\bar{x}(t))}\left(p(t)+\int_{[0,t)}\sum_j \nabla h_j(\bar{x}(s))\mu_j(ds)\right)\cdot v \quad a.e.,$$
$$\operatorname{supp}\{\mu_j\}\subset\{t:h_j(\bar{x}(t))=0\}\quad \textit{for } j=1,\dots,m.$$

If, furthermore, it is assumed that F is convex valued, then (2.2) *implies condition* (2.3) *and also*

$$\dot{p}(t)\in co\left\{q:(-q,\dot{\bar{x}}(t))\right.$$
$$\left.\in\left(t,\bar{x}(t),p(t)+\int_{[0,t)}\sum_j\nabla h_j(\bar{x}(s))\mu_j(ds)\right)\right\}\quad a.e. \tag{2.4}$$

Remarks.

(i) The most significant feature of this theorem is that the customary condition

$$\lambda+||p||_{L^\infty}+\int_{[0,1]}\sum_j\mu_j(ds)\neq 0$$

has been replaced by the condition

$$\lambda+\int_{(0,1]}\sum_j\mu_j(ds)+\left|p(0)+\sum_j\nabla h_j(\bar{x}(0))\mu_j(\{0\})\right|\neq 0.$$

Notice that, if $C_0=\{x_0\}$ and $m=1$, then the theorem does not allow the trivial choice of multipliers

$$\lambda=0,\ p\equiv-\nabla h_1(\bar{x}(0)),\ \mu_1=\delta_{\{0\}},$$

since, for this choice,

$$\lambda=0,\ \int_{(0,1]}\mu_1(ds)=0,$$

and

$$p(0)+\nabla h_1(\bar{x}(0))\mu_1(\{0\})=-\nabla h_1(\bar{x}(0))+\nabla h_1(\bar{x}(0))=0.$$

(ii) The hypotheses of the theorem are less restrictive than those representative of earlier work, in so far as they allow the dynamic constraint multifunction F to be discontinuous as a function of the time variable and to take values nonconvex sets. On the other hand, we restrict attention to problems with a state constraint set which is independent of time and is represented by a finite collection of "smooth" inequality constraints.

(iii) The need for some kind of constraint qualification, of which (2.1) is an example, is clarified by an example devised by A. Ya. and V. A. Dubovitskii and reproduced in [6], in which there are no "inward pointing" velocities at the left endpoint and the only Lagrange multipliers are trivial ones.

(iv) The “adjoint inclusion” (2.2) appearing in these optimality conditions is a “partially convexified” version of the nonsmooth Euler Lagrange condition

$$\left(\dot{p}(t), p(t) + \int_{[0,t)} \sum_{j=1}^{m} \nabla_j h_j(\bar{x}(s))\mu(ds)\right) \in \operatorname{co} N_{\mathrm{Gr}F(t,.)}(\bar{x}(t), \dot{\bar{x}}(t)).$$

Necessary conditions involving the adjoint inclusion (2.2) have attracted attention in recent years, because of the unrestrictive nature of the hypotheses under which they can be derived. They are valid, in particular, when F is no longer assumed to be convex valued. (See [17], [21], [29]). This adjoint inclusion has been used in connection with nondegenerate necessary conditions by Aseev [8] (for “convex” state constrained problems with Lipschitz time-dependent data).

(v) The theorem will be proved by applying necessary conditions (involving the partially convexified Euler Lagrange inclusion above) along a suitably chosen sequence of auxiliary optimization problems, and passage to the limit. We stress, however, that the analysis in this paper is independent of the particular necessary conditions applied to the auxiliary problem. Indeed, we provide a methodology for supplementing a given set of necessary conditions with additional information to ensure nondegeneracy in the sense described in section 1. Other known necessary conditions of choice may be inserted into the analysis to follow. In particular, our methods can be used to prove a nondegenerate form of the Pontryagin maximum principle for state constrained problems. (See [23].)

(vi) Theorem 2.1 does not exclude the degenerate set of multipliers when there is a fixed *right* endpoint ($C_1 = \{x_1\}$) with x_1 a boundary point of A. In this respect, Theorem 2.1, when specialized to the case when F is convex valued and Lipschitz continuous with respect to t, is weaker than the corresponding necessary conditions in [6]. This pathology does not arise, however, in the cases when $\bar{x}(1)$ is interior to either A or C_1.

3. Existence of locally feasible arcs. Take a multifunction $F : [0,1] \times R^n \rightsquigarrow R^n$ and a closed subset $A \subset R^n$.

In the framework of viability theory [1], conditions have been extensively studied under which there exists an F-trajectory $\hat{x}(\cdot)$ with a specified initial value, which satisfies the state constraint

$$\hat{x}(t) \in A \qquad \text{for all } t \in [0, \epsilon']$$

for some $\epsilon' > 0$.

Now suppose

$$A = \bigcap_{j=1}^{m} \{x : h_j(x) \leq 0\}$$

for some functions $h_j : R^n \to R$, $j = 1, \ldots, m$, of class $C^{1,1}$.

Take a nominal F-trajectory x with initial value in A but which possibly fails to satisfy the state constraint on $[0, \epsilon']$. In this paper we shall make use of a refinement of basic existence theorems of viability theory, in which we not only assert (under appropriate additional hypotheses) the existence of an F-trajectory $\hat{x}$ with initial value $x(0)$ satisfying the state constraint but also estimate its "distance" from the nominal trajectory x. Specifically, we show that x can be chosen to satisfy

$$||x - \hat{x}||_{W^{1,1}} \leq K \max_{t \in [0,\epsilon']} h^+(x(t)) \tag{3.1}$$

for some K, independent of the nominal trajectory x. Here

$$h^+(x) := \max\{0, h_1(x), \dots, h_m(x)\}.$$

The right side of (3.1) is a measure of the extent to which the nominal arc x violates the state constraint. These estimates tell us, not unreasonably, that the smaller the margin by which the nominal trajectory x fails to satisfy the state constraint on $[0, \epsilon']$, the closer we can choose $\hat{x}$ to x.

The earliest results along these lines are apparently due to Soner [27], who gave conditions under which it is possible to construct an F-trajectory $\hat{x}$ obeying the state constraint on $[0, 1]$ and which satisfies the L^∞ estimate

$$||x - \hat{x}||_{L^\infty([0,1];R^n)} \leq K \max_{t \in [0,1]} h^+(x(t)). \tag{3.2}$$

Their role in this earlier application was to establish regularity properties of value functions of optimal control problems with state constraints. Soner assumed that the differential inclusion is parameterizable, i.e., it arises from a differential equation parameterized by a control variable.

The applications in this article require a local, sharpened version of Soner's estimate, which is valid when F is possibly nonconvex valued and nonparameterizable and in which the L^∞ estimate (3.2) is replaced by the $W^{1,1}([0, 1]; R^n)$ estimate

$$||x - \hat{x}||_{W^{1,1}([0,1];R^n)} \leq K \max_{t \in [0,\epsilon']} h^+(x(t)). \tag{3.3}$$

Since the $W^{1,1}([0, 1]; R^n)$ norm is stronger than the $L^\infty([0, 1]; R^n)$ norm, (3.3) conveys more information than (3.2). The extra information is precisely what is required to remove the convexity hypotheses commonly invoked in derivation of nondegenerate necessary conditions in optimal control.

The following theorem is a "local" version (we require satisfaction of the state constraint only on a neighborhood of $t = 0$) of a related "global" existence theorem (accompanied by $W^{1,1}$ estimates) proved in [14].

THEOREM 3.1 (existence of neighboring feasible trajectories). *Take an F-trajectory $\bar{x}$ satisfying*

$$\bar{x}(t) \in A \quad \textit{for all } t \in [0, 1].$$

Assume that, for some $k_F(.) \in L^1$ and positive constants c, δ, r, and ϵ, hypotheses (H1) *and* (H2) *and also the constraint qualification* (CQ) *of Theorem* 2.1 *are satisfied.*

Then there exist constants $\delta' \in (0, \delta)$, $\epsilon' \in (0, \epsilon)$, and $K > 0$ with the following properties. Corresponding to any F-trajectory x which satisfies the conditions

$$x(0) \in A \quad \textit{and} \quad ||x - \bar{x}||_{L^\infty([0,1];R^n)} \leq \delta',$$

an F-trajectory $\hat{x}$ can be found, such that $\hat{x}(0) = x(0)$,

$$\hat{x}(t) \in A \quad \textit{for all } t \in [0, \epsilon'],$$

and

$$||x - \hat{x}||_{W^{1,1}([0,1];R^n)} \;\leq\; K \max_{t \in [0,\epsilon']} h^+(x(t)). \tag{3.4}$$

Proof. We note at the outset that if

$$h_j(\bar{x}(0)) \;<\; 0 \quad \text{for some } j,$$

then, by choosing $\delta' > 0$ and $\epsilon' > 0$ sufficiently small, we can arrange that

$$h_j(x(t)) < 0 \quad \text{for all } t \in [0, \epsilon'],$$

for any F-trajectory satisfying $||x - \bar{x}||_{L^\infty} \leq \delta'$. This means that h_j does not contribute to the estimate (3.4) and can be ignored. Accordingly, we assume

$$h_j(\bar{x}(0)) \;=\; 0 \quad \text{for } j = 1, \ldots, m.$$

Notice also that, by scaling the h_j's (this has no effect on the state constraint set they define), we can also arrange that, for $j = 1, \ldots, m$,

$$|\nabla h_j(y)| \leq 1 \text{ for } y \in \bar{x}(0) + \delta B.$$

Fix $r' \in (0, r)$. Let $\omega : R^+ \to R^+$ be a modulus of uniform continuity for $t \to \int_0^t k_F(s)ds$, i.e., $\omega(s) \downarrow 0$ as $s \downarrow 0$ and $\omega(t - s) \geq \int_s^t k_F(\sigma)d\sigma$ for all $[s,t] \in [0,1]$. Let κ be a Lipschitz constant for ∇h on $\bar{x}(0) + \delta\, B$.

Choose a positive number τ which satisfies

$$\tau < \epsilon,\, \tau < (r - r')(c^2\kappa)^{-1},\, \omega(\tau) < \log\left(\frac{r - r'}{8c} + 1\right),\, 2c\tau e^{\omega(\tau)} < \delta/2. \tag{3.5}$$

Claim. For every $\xi \in A \cap (\bar{x}(0) + (\delta/2)B)$ there exists an F-trajectory $\tilde{x} \in W^{1,1}([0,\tau]; R^n)$ such that $\tilde{x}(0) = \xi$ and

$$h_j(\tilde{x}(t)) \;\leq\; -r't \;\text{ for all } t \in [0,\tau],\, j = 1, \ldots, m.$$

We verify the claim. We can select a measurable map $v : [0,\epsilon] \to R^n$ such that $v(t) \in F(t,\xi)$ a.e. $t \in [0,\epsilon]$, and, for $j = 1, \ldots, m$,

$$\nabla h_j(\xi) \cdot v(t) \;\leq\; -r \quad \text{a.e. } t \in [0,\epsilon].$$

Define

$$z(t) \;=\; \xi + \int_0^t v(s)ds.$$

Note that $z(t) \in \bar{x}(0) + \delta B$ for all $t \in [0,\tau]$. From (3.5) we deduce that, for all $t \in [0,\tau]$ and $j = 1, \ldots, m$,

$$\begin{aligned} h_j(z(t)) &= h_j(\xi) + \int_0^t \nabla h_j(z(t)) \cdot v(s)ds \\ &\leq 0 + \int_0^t \nabla h_j(z(t)) \cdot v(s)ds \\ &\leq \int_0^t \nabla h_j(\xi) \cdot v(s)ds + \int_0^t |\nabla h_j(z(t)) - \nabla h_j(\xi)| \cdot |v(s)|ds \\ &\leq -rt + \kappa c^2 t^2/2 \;\leq\; t(-r + (r - r')/2) \leq -\left(\frac{r + r'}{2}\right) t. \end{aligned}$$

By the Filippov Wazewski theorem [2, p.120], applied to the reference arc z, there exists an F-trajectory $\tilde{x} \in W^{1,1}([0,\tau];R^n)$ such that $\tilde{x}(0)=\xi$ and

$$\begin{aligned}|\tilde{x}(t)-z(t)| &\leq \int_0^t d_{F(s,z(s))}(\dot{z}(s))\exp\left(\int_s^t k_F(\sigma)d\sigma\right)ds\\ &\leq ct\int_0^t k_F(s)\exp\left(\int_s^t k_F(\sigma)d\sigma\right)ds\\ &\leq ct\left(\exp\left(\int_0^t k_F(s)ds\right)-1\right)\\ &\leq ct\left(e^{\omega(t)}-1\right) \;\leq\; \frac{r-r'}{8}t \;\text{ for all } t\in[0,\tau].\end{aligned}$$

But then, since each h_j has a Lipschitz constant not exceeding 1 on $\bar{x}(0)+\delta B$, we have, for $j=1,\ldots,m$,

$$\begin{aligned}h_j(\tilde{x}(t)) &\leq |\tilde{x}(t)-z(t)|+h_j(z(t))\\ &\leq \left(\frac{r-r'}{8}-\frac{r+r'}{2}\right)t \;\leq\; -r't \;\text{ for all } t\in[0,\tau].\end{aligned}$$

The claim is confirmed.

Choose $\epsilon' \in (0,\tau)$ and $\delta' \in (0,\delta/2)$ such that

$$(6c/r')(\kappa c e^{\omega(\epsilon')}\epsilon' + (e^{\omega(\epsilon')}-1)) < 1,\; (6c\delta'/r')\exp\left(\int_0^1 k_F(s)ds\right) \leq \delta/2 \;\text{ and }\; \delta' < \epsilon' r'/3.$$

Let x be any F-trajectory such that $x(0) \in A$ and

$$||x-\bar{x}||_{L^\infty([0,1];R^n)} \;\leq\; \delta'.$$

Define

$$\Delta \;:=\; \max_{t\in[0,\epsilon']} h^+(x(t)).$$

We can assume that $\Delta > 0$ since, otherwise, the assertions of the theorem hold good with $\hat{x}=x$.

Notice that since the h_j's, and therefore also h^+, have Lipschitz constant at most 1 on $\bar{x}(0)+\delta B$,

$$\Delta \;\leq\; \max_{t\in[0,\epsilon']} h(\bar{x}(t)) + ||x-\bar{x}||_{L^\infty([0,\epsilon'];R^n)} \;\leq\; 0+\delta'.$$

Set

$$\tilde{\tau} \;:=\; 3\Delta/r'.$$

Observe that $\tilde{\tau} \leq 3\delta'/r' \leq \epsilon' \leq \tau$. We know from our earlier analysis that there exists an F-trajectory $\hat{x}:[0,\tau]\to R^n$ such that $\hat{x}(0)=x(0)$ and $h_j(\hat{x}(s)) \leq -r's$ for all $s\in[0,\tilde{\tau}]$, $j\in\{1,\ldots,m\}$. We have, in particular,

$$h_j(\hat{x}(\tilde{\tau})) \;\leq\; -r'\tilde{\tau} \;=\; -3\Delta \quad \text{for } j=1,\ldots,m.$$

By the Filippov Wazewski theorem, there exists an F-trajectory $y : [\tilde{\tau}, 1] \to R^n$ such that

$$\begin{cases} \dot{y}(t) \;=\; F(t, y(t)) \quad \text{a.e. } t \in [\tilde{\tau}, 1], \\ y(\tilde{\tau}) = \hat{x}(\tilde{\tau}), \end{cases}$$

and, for a.e. $t \in [\tilde{\tau}, 1]$,

$$\begin{aligned} |y(t) - x(t)| &\leq \exp\left(\int_{\tilde{\tau}}^{t} k_F(s)ds\right) |\hat{x}(\tilde{\tau}) - x(\tilde{\tau})| \\ &\leq \exp\left(\int_{\tilde{\tau}}^{t} k_F(s)ds\right) 2c\tilde{\tau}, \qquad (3.6) \\ |\dot{y}(t) - \dot{x}(t)| &\leq k_F(t)\text{exp}\left(\int_{\tilde{\tau}}^{t} k_F(s)ds\right) |\hat{x}(\tilde{\tau}) - x(\tilde{\tau})| \\ &\leq k_F(t)\text{exp}\left(\int_{\tilde{\tau}}^{t} k_F(s)ds\right) 2c\tilde{\tau}. \qquad (3.7) \end{aligned}$$

Now extend $\hat{x}$ to all of $[0, 1]$ by setting

$$\hat{x}(t) \;:=\; y(t) \quad \text{for all } t \in (\tilde{\tau}, 1].$$

Since $\hat{x}(0) = x(0)$, we have from (3.7)

$$\begin{aligned} ||\hat{x} - x||_{W^{1,1}([0,1];R^n)} &= ||\dot{\hat{x}} - \dot{x}||_{L^1([0,\tilde{\tau}];R^n)} + ||\dot{\hat{x}} - \dot{x}||_{L^1([\tilde{\tau},1];R^n)} \\ &\leq 2c\tilde{\tau} + 2c\tilde{\tau}\left(\exp\left(\int_{\tilde{\tau}}^{1} (k_F(s)ds\right) - 1\right) \\ &= 2c\tilde{\tau}\text{exp}\left(\int_{\tilde{\tau}}^{1} k_F(s)ds\right) \\ &\leq (6c/r')\text{exp}\left(\int_{\tilde{\tau}}^{1} k_F(s)ds\right) \Delta. \end{aligned}$$

We have shown that

$$||\hat{x} - x||_{W^{1,1}([0,1];R^n)} \;\leq\; K \max_{t \in [0,\epsilon']} h^+(x(t)),$$

in which

$$K \;:=\; (6c/r')\text{exp}\left(\int_{\tilde{\tau}}^{1} k_F(s)ds\right).$$

It remains to show that

$$\hat{x} \in A \;\text{ for all } t \in [0, \epsilon'].$$

The condition is clearly satisfied for any $t \in [0, \tilde{\tau}]$. On the other hand, for any $t \in [\tilde{\tau}, \epsilon']$

and $j \in \{1, \dots, m\}$, we have from (3.6) and (3.7)

$$\begin{aligned} h_j(\hat{x}(t)) &= h_j(\hat{x}(\tilde{\tau})) + \int_{\tilde{\tau}}^{t} \nabla h_j(\hat{x}(s)) \cdot \dot{\hat{x}}(s) ds \\ &= h_j(\hat{x}(\tilde{\tau})) + \int_{\tilde{\tau}}^{t} \nabla h_j(x(s)) \cdot \dot{x}(s) ds + \int_{\tilde{\tau}}^{t} (\nabla h_j(\hat{x}(s)) - \nabla h_j(x(s))) \cdot \dot{x}(s) ds \\ &\qquad + \int_{\tilde{\tau}}^{t} \nabla h_j(\hat{x}(s)) \cdot (\dot{\hat{x}}(s) - \dot{x}(s)) ds \\ &\le -3\Delta + 2\Delta + (6c^2\kappa/r')\epsilon' e^{\omega(\epsilon')}\Delta + (e^{\omega(\epsilon')} - 1) 2c(3/r')\Delta \\ &\le (-1 + (6c/r')(\kappa c \epsilon' e^{\omega(\epsilon')} + (e^{\omega(\epsilon')-1})))\Delta \ \le \ 0, \end{aligned}$$

as required. We have confirmed that $\hat{x}(t) \in A$ for all $t \in [0, \epsilon']$. The proof is complete.

4. Proof of Theorem 2.1. We precede the proof by a summary of key ideas. Consider to begin with the free right endpoint problem (the case when $C_1 = R^n$). For simplicity, assume $m = 1$. It can be deduced from Theorem 3.1 that, for some $K > 0$ and $\epsilon' \in (0, 1]$, the $W^{1,1}$ local minimizer $\bar{x}$ for (P) is a $W^{1,1}$ local minimizer also for the problem

$$\begin{cases} \text{Minimize } g(x(0), x(1)) + K \max_{t \in [0,\epsilon']} \ (\max\{0, h_1(x(t))\}) \\ \text{over arcs } x \text{ satisfying} \\ \dot{x} \ \in \ F(t, x), \\ x(0) \in C_0 \cap A, \\ h_1(x(t)) \ \le \ 0 \quad \text{for all } t \in [\epsilon, 1]. \end{cases}$$

The notable feature of this problem is that the state constraint on $[0, \epsilon']$ has been replaced by a term in the cost penalizing state constraint violations on this subinterval.

The above problem can, in turn, be reformulated as a standard state-constrained optimal control problem, for which [30] provides necessary conditions of optimality. If λ is the cost multiplier and μ is the multiplier associated with the state constraint, then it can be deduced from the transversality condition for the reformulated problem that

$$\int_{[0,\epsilon']} \mu(ds) \ \le \ \lambda.$$

This condition obviously precludes degenerate multipliers sets in which $\lambda = 0$ and μ is a nonzero measure concentrated on $\{0\}$.

Of course if a right endpoint constraint is present, the analysis is more complicated. In this broader setting, we approximate the optimal control problem by a sequence of right endpoint constraint free problems (in a manner earlier employed by Clarke [9]), derive nondegenerate necessary conditions along the sequence as above and deduce nondegenerate necessary conditions for the original problem by passing to the limit.

After these preliminary comments, we are ready to initiate the proof. Allow once again $m > 1$. Define the scalar valued function h to be

$$h(x) \ = \ \max\{h_1(x), \dots, h_m(x)\}.$$

Let $\delta' \in (0, \delta)$, $\epsilon' \in (0, \epsilon)$ and K be the constants with the properties asserted in Theorem 3.1. Define

$$\begin{aligned} D := \{x \in W^{1,1}([0,1]; R^n) : \dot{x} \in F(t, x(t)) \text{ a.e., } \ x(0) \in C_0 \cap A, \\ \text{and } ||x - \bar{x}||_{L^\infty([0,1];R^n)} \le \delta'\}. \end{aligned}$$

Choose an arbitrary sequence $\epsilon_i \downarrow 0$ and define

$$\eta_i(x_0, x_1, y) := \max\left\{g(x_0, x_1) - g(\bar{x}(0), \bar{x}(1)) + \epsilon_i^2, y, d_{C_1}(x_1)\right\}.$$

The set $\{x \in D : x(t) \in A \text{ for all } t \in [0, \epsilon']\}$ is closed and the function

$$x \to \eta_i(x(0), x(1), \max_{t\in[\epsilon',1]} h^+(x(t)))$$

is continuous on D (with respect to the strong $W^{1,1}([0,1]; R^n)$ topology). Since η_i is nonnegative valued and

$$\eta_i(\bar{x}(0), \bar{x}(1), \max_{t\in[\epsilon',1]} h^+(\bar{x}(t))) = \epsilon_i^2,$$

it follows that

$$\begin{aligned}&\eta_i(\bar{x}(0), \bar{x}(1), \max_{t\in[\epsilon',1]} h^+(\bar{x}(t)))\\ &\quad\le \inf\{\eta_i(x(0), x(1), \max_{t\in[\epsilon',1]} h^+(x(t))) : x \in D,\ x(t) \in A \text{ for all } t \in [0, \epsilon']\} + \epsilon_i^2.\end{aligned}$$

By Ekeland's theorem [9, p. 265] then, there exists $x_i \in W^{1,1,}([0,1]; R^n)$ with the properties

$$||x_i - \bar{x}||_{W^{1,1}([0,1];R^n)} \le \epsilon_i \tag{4.1}$$

and x_i is a minimizer for

$$\text{Minimize } \{J_i(x) : x \in D,\ x(t) \in A \text{ for all } t \in [0, \epsilon']\},$$

where

$$J_i(x) := \eta_i(x(0), x(1), \max_{t\in[\epsilon',1]} h^+(x(t))) + \epsilon_i ||x - x_i||_{W^{1,1}([0,1];R^n)}.$$

In view of (4.1), we can arrange, by subsequence extraction, that

$$x_i \to \bar{x}_i \text{ uniformly} \quad \text{and } \dot{x}_i \to \dot{\bar{x}} \text{ a.e.}$$

Set

$$y_i \ := \ \max_{t\in[\epsilon',1]} h^+(x_i(t)).$$

LEMMA 4.1. *For each i sufficently large,*

$$(a, b, c) \in \partial\eta_i(x_i(0), x_i(1), y_i)$$

implies that $c \ge 0$ and

$$(a, b) \in \alpha\partial g(x_i(0), x_i(1)) + (0, \xi)$$

for some $\alpha \ge 0$ and some $\xi \in N_{C_1}(x_i(1))$ satisfying

$$\alpha + |\xi| + c = 1.$$

Proof. Note, to begin with, that, for all i sufficently large,

$$\eta_i(x_i(0), x_i(1), \max_{t\in[\epsilon',1]} h^+(x_i(t))) > 0.$$

Indeed, by choosing i sufficiently large we can make $||x_i - \bar{x}||_{W^{1,1}}$ arbitrarily small. If the above strict inequality is not satisfied, then x_i satisfies the constraints of (P) and has cost less than $\bar{x}$, in contradiction to our assumption that $\bar{x}$ is a $W^{1,1}$ local minimizer.

There are two cases to consider.

Case 1. $d_{C_1}(x_i(1)) = 0$. In view of the preceding strict inequality, we deduce from the max rule for limiting subdifferentials [25] that

$$a \in \alpha\partial g(x_i(0), x_i(1))$$

for some $\alpha \geq 0$ such that $\alpha + c = 1$. The assertions of the lemma are satisfied with $\xi = 0$.

Case 2. $d_{C_1}(x_i(1)) > 0$. In this case (by well-known properties of the distance function)

$$\xi \in \partial d_{C_1}(x_i(1)) \text{ implies } |\xi| = 1.$$

The max rule now asserts that $c \geq 0$ and that there exist $\alpha \geq 0$ and $\lambda \geq 0$ such that $\alpha + \lambda + c = 1$ and

$$(a, b) \in \alpha\partial g(x_i(0), x_i(1)) + (0, \lambda\xi')$$

for some $\xi' \in \partial d_{C_1}(x_i(1))$ such that $|\xi'| = 1$. Writing $\xi = \lambda\xi'$, we find that

$$(a, b) \in \alpha\partial g(x_i(0), x_i(1)) + (0, \xi)$$

and $\alpha + |\xi| + c = 1$. Since $\xi \in N_{C_1}(x_i(1))$, the lemma is proved. □

In the proof of the next lemma we shall use the easily confirmed fact that there exists some $k > 0$ (independent of i) such that, for any $x, x' \in D$,

$$|J_i(x) - J_i(x')| \leq k||x - x'||_{W^{1,1}([0,1];R^n)}.$$

LEMMA 4.2. *For all i sufficiently large, x_i is a $W^{1,1}$ local minimizer for the optimization problem*

$$(\mathrm{Q_i}) \qquad \textit{Minimize}\left\{ J_i(x) + kK \max_{t\in[0,\epsilon']} h^+(x(t)) : x \in D \right\}.$$

Proof. Define

$$\tilde{J}_i(x) := J_i(x) + kK \max_{t\in[0,\epsilon']} h^+(x(t)).$$

Suppose the assertions of the lemma are false. Take any $\delta'' \in (0, \delta')$. Then for any index value i_0 there exists $i \geq i_0$ and $x \in D$ such that $||x - x_i||_{W^{1,1}} \leq \delta''$ and

$$\tilde{J}_i(x) \, < \, \tilde{J}_i(x_i).$$

According to Theorem 3.1, there exists an F-trajectory $\hat{x}$ such that $\hat{x}(0) = x(0)$,

$$h(\hat{x}(t)) \, \leq \, 0 \quad \text{for } t \in [0, \epsilon'],$$

and

$$||\hat{x} - x||_{W^{1,1}([0,1];R^n)} \leq K \max_{t\in[0,\epsilon']} h^+(x(t)).$$

By choosing i_0 sufficiently large and δ'' sufficiently small we can arrange that $\hat{x}$ lies in D. It then follows from the minimizing properties of x_i that

$$\begin{aligned}\tilde{J}_i(x_i) \leq \tilde{J}_i(\hat{x}) &= J_i(\hat{x}) \\ &\leq J_i(x) + k||\hat{x} - x||_{W^{1,1}([0,1];R^n)} \\ &\leq J_i(x) + kK \max_{t\in[0,\epsilon']} h^+(x(t)) = \tilde{J}_i(x) < \tilde{J}_i(x_i),\end{aligned}$$

which is not possible. The lemma is proved. □

We deduce from the preceding lemma that, for i sufficiently large,

$$\left(x_i, y_i \equiv \max_{t\in[\epsilon',1]} h^+(x_i(t)), z_i \equiv 0\right)$$

is a strong local minimizer for the optimal control problem

$$\left\{\begin{array}{l}\text{Minimize } \max\{g(x(0),x(1)) - g(\bar{x}(0),\bar{x}(1)) + \epsilon_i^2, y(1), d_{C_1}(x(1))\} \\ \qquad\qquad +\epsilon_i|x(0) - x_i(0)| + \epsilon_i \int_0^1 |\dot{x}(t) - \dot{x}_i(t)|dt + kK \max\{0, z(1)\} \\ \text{over } x \in W^{1,1}([0,1];R^n),\ y \in W^{1,1}([0,1];R) \text{ and } z \in W^{1,1}([0,1];R) \text{ satisfying} \\ (\dot{x}(t), \dot{y}(t), \dot{z}(t)) \in F(t, x(t)) \times \{0\} \times \{0\} \text{ a.e.,} \\ x(0) \in C_0 \cap A, \\ \tilde{h}(t, x(t), y(t), z(t)) \leq 0 \text{ for all } t \in [0,1].\end{array}\right.$$

Here, $\tilde{h} : [0,1] \times R^n \times R \times R$ is the function

$$\tilde{h}(t,x,y,z) := \left\{\begin{array}{ll} h(x) - z & \text{for } t \in [0,\epsilon), \\ h(x) + \max\{-z,-y\} & \text{for } t = \epsilon, \\ h(x) - y & \text{for } t \in (\epsilon,1]. \end{array}\right.$$

We have arrived at a problem to which the Euler Lagrange type conditions of [30] are applicable. (Note, in particular, that the state constraint functional $\tilde{h}$ has the requisite upper semicontinuity properties.)

Taking note of Lemma 4.1, we draw the following conclusions: there exist $p_i \in W^{1,1}([0,1];R^n)$, $q_i \in W^{1,1}([0,1];R)$, and $r_i \in W^{1,1}([0,1];R)$ (the costate functions associated with x_i, y_i, and z_i, respectively), constants $\lambda'_i \geq 0$, $\alpha_i \in [0,1]$, $\beta_i \in [0,1]$, $\pi_i^0 \in [0,1]$, $\pi_i^1 \in [0,1]$, $\xi_i \in R^n$, a Borel measurable function $\gamma_i : [0,1] \to R^n$, and nonnegative Borel measures $\nu_i \in C^*([0,1];R)$ such that

$$(4.2)\quad ||p_i||_{L^\infty} + ||q_i||_{L^\infty} + ||r_i||_{L^\infty} + \lambda'_i + \int_{[0,1]} \nu_i(ds) = 1,$$

$$(4.3)\quad \alpha_i + \beta_i + |\xi_i| = 1,\ \pi_i^0 + \pi_i^1 = 1,$$

$$(4.4)\quad \left.\begin{array}{l}\dot{p}_i(t) \in \text{co}\left\{\eta : \left(\eta, p_i(t) + \displaystyle\int_{[0,t)} \gamma_i(s)\nu_i(ds)\right)\right. \\ \qquad\qquad \in N_{\text{Gr}F(t,.)}(x_i(t), \dot{x}_i(t)) + \{0\} \times \lambda'_i\epsilon_i B\end{array}\right\} \text{ a.e.,}$$

$$\begin{aligned}
&\left(p_i(0), -\left(p_i(1) + \int_{[0,1]} \gamma_i(s)\nu_i(ds)\right)\right) \in \alpha_i\lambda_i'\partial g(x_i(0), x_i(1)) \\
&\qquad\qquad\qquad + \epsilon_i\lambda_i' B \times \{\lambda_i'\xi_i\} + N_{C_0\cap A}(x_i(0)) \times \{0\}, \\
&q_i(0) = 0,\ \dot{q}_i = 0,\ \int_{(\epsilon',1]} \nu_i(ds) + \pi_i^0\nu_i(\{\epsilon'\}) = \beta_i\lambda_i', \\
&r_i(0) = 0,\ \dot{r}_i = 0,\ \int_{[0,\epsilon')} \nu_i(ds) + \pi_i^1\nu_i(\{\epsilon'\})\ \le\ kK\lambda_i', \\
&\xi_i \in N_{C_1}(x_i(1)), \\
&\left(p_i(t) + \int_{[0,t)} \gamma_i(s)\nu_i(ds)\right) \cdot \dot{x}_i(t) \\
&\quad = \ \max\nolimits_{v\in F(t,x_i(t))} \left(p_i(t) + \int_{[0,t)} \gamma_i(s)\nu_i(ds)\right) \cdot v - \epsilon_i|v - \dot{x}_i(t)| \quad \text{a.e.}, \\
&\operatorname{supp}\{\nu_i\} \subset \{t \in [0,\epsilon'] : h(x_i(t)) = 0\} \cup \{t \in [\epsilon', 1] : h(x_i(t)) = y_i\}, \\
&\gamma_i(t)\ \in \operatorname{co}\partial h(x_i(t)) \quad \nu\text{-a.e.}
\end{aligned}$$

It can be deduced from (4.4) (see [20]) that, for each i,

$$|\dot{p}_i(t)|\ \le\ k_F(t)\left|p_i(t) + \int_{[0,t)} \gamma(s)\nu_i(ds)\right| \quad \text{a.e.}$$

(We remark on the significance of the "partially convexified" adjoint inclusion (4.4): if (4.4) were replaced by the coarser inclusion $(\dot{p}_i, p_i) \in \operatorname{co} N_{\mathrm{Gr}\,F}$, then this pointwise bound on the $\dot{p}_i$, which has an important role in the convergence analysis, would no longer be valid under the hypotheses of Theorem 2.1.)

From (4.2), $\{p_i\}$ is a bounded sequence. By Gronwall's lemma then, the p_i's are uniformly bounded and the $\dot{p}_i$'s are uniformly integrably bounded. It can be deduced from the Dunford Pettis theorem that, along some subsequence,

$$p_i \to p \quad \text{uniformly} \qquad \text{and} \qquad \dot{p}_i \to \dot{p} \ \text{ weakly in } L^1$$

for some $p \in W^{1,1}([0,1]; R^n)$. In view of (4.2), and (4.3), we can arrange, by further subsequence extraction, that

$$\nu_i(dt) \to \nu(dt) \text{ and } \gamma_i(t)\nu_i(dt) \to \gamma(t)\nu(dt) \text{ weakly}^*$$

for some nonnegative Borel measure $\nu \in C^*([0,1]; R)$ and some ν-integrable function γ. Also,

$$(\lambda_i', \alpha_i, \beta_i, \xi_i, \pi_i^0\pi_i^1) \to (\lambda', \alpha, \beta, \xi, \pi^0, \pi^1)$$

for some $\xi \in R^n$, $\lambda' \ge 0$, $\alpha \in [0,1]$ $\beta \in [0,1]$, $\pi^0 \in [0,1]$, and $\pi^1 \in [0,1]$.

Using an analysis similar to that of [29] and [30], we may pass to the limit in the above relationships. We thereby arrive at the following relationships:

(4.5) $\quad ||p||_{L^\infty} + \lambda' + \int_{[0,1]} \nu(ds) = 1,$

(4.6) $\quad \alpha + \beta + |\xi| = 1,\ \pi^0 + \pi^1 = 1,$

$$\dot p(t) \in \operatorname{co}\left\{\eta : \left(\eta, p(t) + \int_{[0,t)} \gamma(s)\nu(ds)\right) \in N_{\mathrm{Gr}F(t,.)}(\bar x(t), \dot{\bar x}(t))\right\} \text{ a.e.,} \tag{4.7}$$

$$\left(p(0), -\left(p(1) + \int_{[0,1]} \gamma(s)\nu(ds)\right)\right) \tag{4.8}$$

$$\in \alpha\lambda'\partial g(\bar x(0), \bar x(0)) + N_{C_0 \cap A}(\bar x(0)) \times \{\lambda'\xi\},$$

$$\int_{(\epsilon',1]} \nu(ds) + \pi^0\nu_i(\{\epsilon'\}) \;=\; \beta\lambda', \tag{4.9}$$

$$\int_{[0,\epsilon')} \nu_i(ds) + \pi_i^1\nu_i(\{\epsilon'\}) \;\le\; kK\lambda', \tag{4.10}$$

$$\xi \in N_{C_1}(\bar x(1)),$$

$$\left(p(t) + \int_{[0,t)} \gamma(s)\nu(ds)\right) \cdot \dot{\bar x}(t) \;=\; \max\nolimits_{v \in F(t, x_i(t))} \left(p(t) + \int_{[0,t)} \gamma(s)\nu(ds)\right) \cdot v$$

$$\gamma(t) \in \operatorname{co}\partial h(\bar x(t)) \ \nu\text{-a.e.}$$

$$\operatorname{supp}\{\nu\} \subset \{t : h(\bar x(t)) = 0\}.$$

The most challenging task in justifying these relationships is to confirm the adjoint inclusion (4.7); we give details of this step alone. The following conclusions can be drawn from Mazur's and Carathéodory's theorems. For each i there exist integers $0 \le k_{i0} \le \cdots \le k_{in}$ and a convex combination $\{\alpha_{i0}, \ldots, \alpha_{in}\}$ such that $\sum_{j=0}^{n} \alpha_{ij}\dot p_{i+k_{ij}}(t)$, $i = 1, 2, \ldots$, converges strongly in L^1 to $\dot p$. We can arrange by subsequence extraction that convergence is pointwise on some set $\mathcal{D}$ of full measure. By modifying the set $\mathcal{D}$, if necessary, we can also arrange that $\dot x_i \to \dot{\bar x}$ pointwise on $\mathcal{D}$. Fix $t \in \mathcal{D}$. For each j, $\alpha_{ij}, i = 1, 2, \ldots$, and $\dot p_{i+k_{ij}}(t), i = 1, 2, \ldots$, are bounded. After further subsequences have been extracted, we have that, for $j = 0, \ldots, n$, $\alpha_{ij} \to \alpha_j$ for some α_j and $\dot p_{i+k_{ij}}(t) \to p_j$ for some $\tilde p_j$. Again making use of Carathéodory's theorem, we deduce that, for each k,

$$\tilde p_j \;\in\; \operatorname{co}\left\{\eta : \left(\eta, p(t) + \int_{[0,t)} \gamma(s)\nu(ds)\right) \in N_{\mathrm{Gr}F(t,.)}(x(t), \dot x(t))\right\}.$$

Clearly, the α_j's are nonnegative and sum to one, so

$$\dot p(t) = \sum_{j=0}^{n} \alpha_j \tilde p_j(t)$$

$$\in \operatorname{co}\left\{\eta : \left(\eta, p(t) + \int_{[0,t)} \gamma(s)\nu(ds)\right) \in N_{\mathrm{Gr}F(t,.)}(x(t), \dot x(t))\right\} \quad \text{a.e.}$$

Note that (4.7) implies

$$|\dot p(t)| \;\le\; k_F(t)\left|p(t) + \int_{[0,t)} \gamma(s)\nu(ds)\right| \quad \text{a.e.} \tag{4.11}$$

We see immediately that $\lambda' \neq 0$, for otherwise (4.8), (4.9), (4.10), and (4.11) (together with Gronwall's inequality) imply that $\nu = 0$ and $p = 0$, which contradicts (4.5).

Now define $\lambda = \alpha\lambda'$. We claim that

$$\lambda + \int_{(0,1]} \nu(ds) + |p(0) + \gamma(0)\nu(\{0\})| \neq 0. \tag{4.12}$$

Suppose that this condition is violated. Then $\alpha = 0$ and (by (4.9)) $\beta = 0$. It follows from (4.6) that $|\xi| = 1$. But then, by (4.8),

$$\left| p(1) + \int_{[0,1]} \gamma(s)\nu(ds) \right| \neq 0.$$

However, we have assumed that $\int_{(0,1]} \nu(ds) = 0$ and $|p(0) + \gamma(0)\nu(\{0\})| = 0$. It follows from (4.11) and Gronwall's inequality that $|p(1) + \int_{[0,1]} \gamma(s)\nu(ds)| = 0$. This contradiction confirms (4.12).

It can be deduced from the inclusion

$$\gamma(t) \in \operatorname{co} \partial h^+(\bar{x}(t)) \quad \nu\text{-a.e.},$$

the max rule for limiting subdifferentials and a measurable selection theorem, that there exist Borel measurable functions $\alpha_1, \dots, \alpha_m : [0,1] \to [0,1]$ such that, ν-a.e.,

$$\gamma(t) = \sum_j \alpha_j(t) \nabla h_j(\bar{x}(t)), \quad \sum_j \alpha_j(t) = 1,$$

and

$$h_j(\bar{x}(t)) < h(\bar{x}(t)) \text{ implies } \alpha_j(t) = 0 \text{ for } j \in \{1, \dots, m\}.$$

Define the nonnegative Borel measures $\mu_j \in C^*([0,1]; R)$, $j = 1, \dots, m$, to be

$$\mu_j(dt) := \alpha_j(t)\nu(dt).$$

We note that, for any Borel subset $I \subset [0,1]$,

$$\int_I \gamma(s)\nu(ds) = \int_I \sum_j \nabla h_j(\bar{x}(s))\mu_j(ds),$$

$$\int_I \nu(ds) = \int_I \left(\sum_j \alpha_j(s) \right) \nu(ds) = \int_I \sum_j \mu_j(ds).$$

Since $\operatorname{supp}\{\nu\} \subset \{t : h(\bar{x}(t)) = 0\}$, we conclude that

$$\operatorname{supp}\{\mu_j\} \subset \{t : h_j(\bar{x}(t)) = 0\} \quad \text{for } j = 1, \dots, m.$$

The earlier relationships, expressed in terms of the μ_j's in place of ν, amount to the assertions of the theorem, with the exception of the additional claims when F is convex valued. Notice in particular that, from (4.12), we can deduce the nondegeneracy condition

$$\lambda + \int_{(0,1]} \sum_j \mu_j(ds) + \left| p(0) + \sum_j \nabla h_j(\bar{x}(0))\mu_j(\{0\}) \right| \neq 0.$$

Finally suppose that F is convex valued. In this case and under the hypotheses of the theorem, equivalence of the adjoint inclusion (2.2) and the generalized Hamiltonian condition (2.4) is established in [24]. (See also [17].) The fact that (in the convex case) the Hamiltonian inclusion (2.4) implies the Weierstrass condition (2.3) is well known. (See, for example, [9].)

REFERENCES

[1] J.-P. Aubin, *Viability Theory*, Birkhäuser Boston, Boston, MA, 1991.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions. Set-Valued Maps and Viability Theory*, Springer-Verlag, Berlin, New York, 1984.

[3] A. V. Arutyunov, *On necessary optimality conditions in a problem with phase constraints*, Soviet Math. Dokl., 31 (1985), pp. 174–177.

[4] A. V. Arutyunov, S. M. Aseev, and V. I. Blagodat-Skikh, *First order necessary conditions in the problem of optimal control of a differential inclusion with phase constraints*, Sb. Math., 79 (1994), pp. 117–139.

[5] A. V. Arutyunov and S. M. Aseev, *State constraints in optimal control. The degeneracy phenomenon*, Systems Control Lett., 26 (1995), pp. 267–273.

[6] A. V. Arutyunov and S. M. Aseev, *Investigation of the degeneracy phenomenon of the maximum principle for optimal control problems with state constraints*, SIAM J. Control Optim., 35 (1997), pp. 930–952.

[7] A. V. Arutyunov and N. T. Tynyanskiy, *The maximum principle in a problem with phase constraints*, Izv. Akad. Nauk SSSR Tekhn. Kibernet., 4 (1984), pp. 60–68 (in Russian).

[8] S. M. Aseev, *Method of smooth approximations in the theory of necessary conditions for differential inclusions*, Iz. Math., V61, (1997), pp. 235–258.

[9] F. H. Clarke, *Optimization and nonsmooth analysis*, John Wiley, New York, 1983.

[10] A. Ja. Dubovitskii and V. A. Dubovitskii, *Necessary conditions for a strong minimum in optimal control problems with degenerate endpoint constraints and phase constraints*, Uspekhi Mat. Nauk, 40 (1985), pp. 175–176.

[11] A. Ja. Dubovitskii and A. A. Milyutin, *Extremum problems in the presence of restrictions*, Z. Vychisl. Math. i Math. Fiz., 5 (1965), pp. 395–453.

[12] A. Ja. Dubovitskii and A. A. Milyutin, *Necessary conditions for a weak extremum in optimal control problems with mixed constraints of inequality type*, Zh. Vychisl. Math. Math. Fiz., 8 (1968), pp. 725–770 (in Russian) (English translation, *Comp. Math. Math. Phys.*, 8 (1968).

[13] H. Frankowska and F. Rampazzo, *Extensions of the Filippov Wazewski Theorem to Sets*, preprint.

[14] H. Frankowska and R. B. Vinter, *Existence of neighbouring feasible trajectories: Applications to dynamic programming for state constrained optimal control problems*, J. Optim. Theory Appl., 104 (2000), pp. 21–40.

[15] M. M. A. Ferreira and R. B. Vinter, *When is the maximum principle for state-constrained problems nondegenerate?*, J. Math. Anal. Appl., 187 (1994), pp. 438–467.

[16] M. M. A. Ferreira, F. A. C. C. Fontes, and R. B. Vinter, *Nondegenerate necessary conditions for nonconvex optimal control problems with state constraints*, J. Math. Anal. Appl., 233 (1999), pp. 116–129.

[17] A. D. Ioffe, *Euler Lagrange and Hamiltonian formalisms in dynamic optimization*, Trans. Amer. Math. Soc., 349 (1997), pp. 2871–2900.

[18] P. D. Loewen, *Optimal Control via Nonsmooth Analysis*, CRM Proceedings and Lecture Notes 2, AMS, Providence, RI, 1993.

[19] P. D. Loewen and R. T. Rockafellar, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.

[20] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–282.

[21] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.

[22] L. W. Neustadt, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[23] F. Rampazzo and R. B. Vinter, *A theorem on the existence of neighbouring feasible trajectories with applications to optimal control*, IMA J. Math. Control Systems, to appear.

[24] R. T. Rockafellar, *Equivalent subgradient versions of Hamiltonian and Euler–Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1314.

[25] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, New York, 1998.

[26] J. D. L. Rowland and R. B. Vinter, *Dynamic optimization problems with free-time and active state constraints*, SIAM J. Control Optim., 31 (1993), pp. 677–697.

[27] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[28] R. B. Vinter, *Optimal Control*, Birkhäuser Boston, Boston, MA, 2000.

[29] R. B. Vinter and H. Zheng, *The extended Euler–Lagrange condition for nonconvex variational problems*, SIAM J. Control Optim., 35 (1997), pp. 56–77.

[30] R. B. Vinter and H. Zheng, *The extended Euler–Lagrange condition for nonconvex variational problems with state constraints*, Trans. Amer. Math. Soc., 350 (1998), pp. 1181–1204.

[31] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1008–1042

ON THE SYNTHESIS PROBLEM FOR OPTIMAL CONTROL SYSTEMS*

R. GABASOV†, F. M. KIRILLOVA‡, AND N. V. BALASHEVICH‡

Abstract. A new approach to the synthesis problem of optimal controls of feedback type is considered. Algorithms of operating optimal controllers which are able to calculate values of optimal feedbacks during each particular control process in real time are described. The algorithm of operating the controller of the first type (continuous controller) is based on solving special (defining) equations in real time using the Newton method. For operating the optimal controller of the second type (discrete controller) the dual method of linear programming is used. In control problems for systems under uncertainty, a nonstochastic model is used, around which guaranteed results are obtained. Under incomplete and inexact information, problems of optimal observation and identification are introduced. Algorithms of operating the optimal estimator and identifier which are able to construct estimates of available information required for operating the optimal output controller in real time are described. Results are illustrated by optimization problems of the fourth order control system.

Key words. optimal control problem, optimal state and output feedbacks, synthesis of optimal systems, optimal controllers, optimal observation and identification problems, optimal estimators and identifiers

AMS subject classifications. 49N05, 49N35, 93B50

PII. S036301299936124X

1. Introduction.

Problems of optimal control (OC) have been intensively investigated in the world literature for over forty years. During this period, a series of fundamental results have been obtained, among which should be noted the maximum principle [53] and dynamic programming [6, 7]. For many of the problems of the optimal control theory (OCT) adequate solutions are found [1, 8, 51]. Results of the theory were taken up in various fields of science, engineering, and economics. However, it is generally recognized that one of the central problems of the OCT, which is, undoubtedly, the problem of synthesis of optimal feedback, has not yet been solved even for linear systems apart from special cases [9, 12]. This essentially retards the deeper integration of the results of the OCT to practice because controls of feedback type but not open-loop controls (which are investigated in a great deal of published works) are used for solving many practical problems. This is also evidenced by the fact that synthesis of OC in the Kalman–Letov linear-quadratic problem [46, 52] is extraordinarily widely used for solving various applied problems. The Kalman–Letov problem, by virtue of its specific character, takes a particular place in the OCT. This problem is, in fact, a problem of the classical calculus of variations because in this problem there are no geometric constraints of controls which have defined the origin of the modern trend of the calculus of variations named the OCT. It is worth noting that geometric constraints are a typical nonlinearity most often met in applied problems.

*Received by the editors September 3, 1999; accepted for publication (in revised form) June 19, 2000; published electronically November 8, 2000. This research was partially supported by grant F99R-002 from the Belarus Republic Foundation for Basic Research and the State Program for Basic Research of the National Academy of Sciences of Belarus (Mechanics 49).

http://www.siam.org/journals/sicon/39-4/36124.html

†Department of Applied Mathematics and Computer Science, Byelorussian State University, F. Scorina Avenue 4, 220050, Minsk, Belarus.

‡Institute of Mathematics, NASB, Surganov str. 11, 220072, Minsk, Belarus (kirill@nsys.minsk.by, balash@im.bas-net.by).

The purpose of the article is to present some results of investigations on the problem of optimal synthesis conducted in Minsk since the beginning of 90's. These investigations rely on a new approach to the problem [19, 41]. Analysis of the classical statement of the optimal synthesis problem showed that it could not be solved for more or less serious cases. This problem imposes extremely stringent requirements on the result, and these requirements cannot be satisfied by analytical mathematical methods. It turns out that for applications it is sufficient to have much less information on optimal synthesis, and this information does not have to be in analytical form. By virtue of this, the basis for the approach described below is the idea of reasonable combination of analytical mathematical methods and potentials of modern computing. It is obvious that such an idea could not arise at the time when the classical statement of the optimal synthesis problem appeared (the beginning of 50's). In modern times it is, in our opinion, quite natural.

As well as in the case of the Kalman–Letov problem, the obtained results on optimal synthesis were taken up for solving actual problems of both the classical and modern control theory which in initial statement does not have extremal nature. But in this work we do not dwell on these problems.

The work has the following structure. In section 2 the terminal OC problem is formulated, and the notions of a realization of optimal feedback and an optimal controller are introduced. In sections 3–5 algorithms of operating three types of optimal controllers (continuous, discrete, combined open-closed loop) are described. In section 6 the terminal OC problem under constantly acting disturbances is investigated, and four various types of feedback are discussed. In section 7 the terminal OC problem under incomplete and inexact information on a current system state is considered, and the notion of a realization of optimal output feedback is introduced. In section 8 a linear optimal observation problem (OO) is introduced, and an algorithm of operating a continuous estimator, which calculates estimates of an a posteriori distribution of the initial state, is described. In section 9 an algorithm of operating a discrete estimator is described. This estimator uses measurements of output signals only at discrete moments. In section 10 an OO problem under constantly affecting bounded disturbances is solved. The cases of different information on disturbances with identification or without identification of parameters are considered. In section 11 a realization of optimal output feedback without outside disturbances is constructed. Realization of optimal output feedback under constantly affecting disturbances is described in section 12.

2. Problem statement. In the class of piecewise continuous functions $u(t)$, $t \in T = [0, t^*]$, consider a linear OC problem

$$\begin{gathered} J(u) = c'x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu, \quad x(0) = x_0, \\ Hx(t^*) = g, \quad |u(t)| \leq 1, \quad t \in T. \end{gathered} \tag{2.1}$$

Here $x = x(t)$ is an n-vector of a state of the control system at a moment t, $u = u(t)$ is a value of a scalar (one-dimensional) control, g is an m-vector of given values of output signals, and t^* is a terminal (fixed) moment; vectors participate in operations as columns, and transposition $'$ (prime) is used for obtaining a row vector.

Suppose that rank $(b, Ab, \dots, A^{n-1}b) = n$ and rank $H = m < n$.

Problem (2.1) is the simplest OC problem in the sense that removal of any element from this problem makes it trivial. On the other hand, it contains almost all typical elements of the OC problem. Some other statements of OC problems are considered in [3, 4, 14, 38, 40]. From the qualitative point of view, problem (2.1) has been investigated very deeply. Also, there exist effective numerical methods of constructing its open-loop solutions.

By an open-loop solution to problem (2.1), it is meant a piecewise continuous function $u^0(t)$, $t \in T$, which (1) satisfies the geometric constraint $|u^0(t)| \leq 1$, $t \in T$, (2) generates the trajectory $x^0(t)$, $t \in T$, satisfying the terminal constraint $Hx^0(t^*) = g$, and (3) provides the maximum value of the control criterion: $J(u^0) = \max J(u)$.

In spite of the significance of open-loop solutions for many applied problems, in the majority of applications of the control theory preference is given to positional solutions or (in another terminology) OCs of feedback type. To define such controls let us imbed problem (2.1) into a family of problems

$$\begin{gathered} J(u) = c'x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu, \quad x(\tau) = z, \\ Hx(t^*) = g, \quad |u(t)| \leq 1, \quad t \in T(\tau) = [\tau, t^*], \end{gathered} \tag{2.2}$$

depending on a scalar $\tau \in T$ and an n-vector $z \in R^n$.

Denote by $u^0(t|\tau, z)$, $t \in T(\tau)$, an optimal open-loop control for problem (2.2) for a fixed pair (τ, z) called a position. Let X_τ be a set of all vectors $z \in R^n$ for which problem (2.2) has a solution for a fixed moment τ.

Definition. A function

$$u^0(\tau, z) = u^0(\tau|\tau, z), \quad z \in X_\tau, \quad \tau \in T, \tag{2.3}$$

is said to be an OC of feedback type (a positional solution) of problem (2.2).

Constructing function (2.3) is called the synthesis of the optimal system.

Unlike an open-loop solution, a positional solution to problem (2.2) cannot be effectively constructed either by the maximum principle or by the dynamic programming. The maximum principle, first of all, is oriented to optimal open-loop controls although time-optimal systems can be synthesized with its help on plane ($n = 2$). In due time great hopes were pinned on the dynamic programming. Really, a result obtained by this method has the form of OC of feedback type. However, to obtain this result it is necessary, first of all, to define the corresponding solution to the Bellman equation and then to find a method for its constructing. Even if one manages to overcome these difficulties, the famous "curse of dimension" arises. Due to this it is impossible to tabulate functions of many variables to the high enough precision in the course of numerical solving of the Bellman equation for problem (2.2) with $n \geq 3$.

Sources of difficulties under the classical approach to the problem are, at our glance, in the classical statement of the optimal synthesis problem itself. First of all, one should take into account that the problem was stated by engineers as early as the beginning of 50's when potentials of today's computing could not even be inferred. Engineers working with heuristic feedback and synthesizing time-optimal linear two-dimensional systems thought that invoking more powerful mathematical methods could solve the problem also in the general case (if only for linear problems of OC). Furthermore, the problem statement of optimal synthesis was also influenced by the traditional (school) concept of the solution to mathematical problems when the main purpose was to obtain an answer in terms of a formula that will allow one to solve the problem once and for all. Certainly, knowing the history of mathematics, one could foresee unsolvability of the optimal synthesis problem formulated at that time. Finally, the problem statement of optimal synthesis could have been considerably influenced by the conventional (in those years) form of realization of feedback with the help of mechanical, hydraulic, electrical, etc. devices.

To leave an impasse, let us analyze the use of optimal feedbacks in actual practice as if they have been constructed in some way. First of all, it is necessary to understand that even though optimal feedback (2.3) is introduced by determined model (2.1), engineers well know that it is needed for operating a control system under real conditions when the system is affected by unaccounted and unknown disturbances in (2.1). That alone is reason enough to say that open-loop controls are useless. In mathematical works no attention is given to this circumstance. Thus, assume that for problem (2.2) optimal feedback (2.3) is constructed. Let us close control system (2.1) by it and study the behavior of the closed system[1] at the presence of unknown disturbances $w(t)$, $t \geq 0$,

$$\dot{x} = Ax + bu^0(t,x) + w(t), \quad x(0) = x_0. \tag{2.4}$$

Denote by $w^*(t)$, $t \in T$, a disturbance realized in some particular control process. It generates a trajectory $x^*(t)$, $t \in T$, of (2.4),[2] i.e., for almost all $t \in T$ the identity

$$\dot{x}^*(t) \equiv Ax^*(t) + bu^0(t, x^*(t)) + w^*(t), \quad t \in T, \quad x^*(0) = x_0^*, \tag{2.5}$$

is fulfilled.

From (2.5) it is obvious that in the process under consideration the input of control system (2.1) is fed only by the signals

$$u^*(t) = u^0(t, x^*(t)), \quad t \in T. \tag{2.6}$$

Thus, in a particular control process, not all of the optimal feedback is used, but its values are necessary only along the isolated continuous curve $x^*(t)$, $t \in T$. Moreover, for each current moment $\tau \in T$ the value $u^*(\tau) = u^0(\tau, x^*(\tau))$ does not have to be known beforehand; it is enough to know how to calculate it at the moment τ, when the system appears in a current state $x^*(\tau)$. The purpose of the proposed approach is to show that in view of the finite rate of real processes and the high speed of up-to-date computer facilities it is possible to obtain the values of the function $u^*(t)$, $t \in T$, in real time[3] for many problems (2.1).

Definition. Function (2.6) is said to be a realization of optimal feedback (2.3), and a device able to calculate its values in real time is called an optimal controller.

Thus, the problem of synthesis of optimal systems in the new statement is reduced to constructing an algorithm of operating the optimal controller.

3. Defining equations and an algorithm of operating the optimal controller. Suppose that an algorithm of operating the optimal controller has been created and the optimal controller has been operating on the interval $[0, \tau[$. Denote by $u^*(t)$, $t \in [0, \tau[$, a control produced by the controller up to the moment τ, denote by $w^*(t)$, $t \in [0, \tau[$, a realized disturbance, by $x^*(\tau)$ a state of the control system at the moment τ corresponding to $u^*(t)$, $w^*(t)$, $t \in [0, \tau[$, and the initial state $x(0) = x_0$. To calculate the current value $u^*(\tau)$ of the control $u^0(\tau, x^*(\tau))$, according to definition (2.3), the optimal controller has to know the open-loop solution $u^0(t|\tau, x^*(\tau))$, $t \in T(\tau)$, to the OC problem

[1] To simplify notations let us consider model (2.1) to be exact.

[2] For simplicity it is assumed that (2.4) has a solution, although for the classical statement of optimal synthesis problem it is not a simple question, because function (2.3) is discontinuous with respect to x.

[3] The concept "real time mode" is explained below in section 3.

$$c'x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu, \quad x(\tau) = x^*(\tau), \tag{3.1}$$

$$Hx(t^*) = g, \quad |u(t)| \le 1, \quad t \in T(\tau).$$

From the Pontryagin maximum principle [53] it follows that the optimal open-loop control $u^0(t|\tau, x^*(\tau))$, $t \in T(\tau)$, has the form

$$u^0(t|\tau, x^*(\tau)) = \mathrm{sgn}\Delta(t|\tau, x^*(\tau)), \quad t \in T(\tau), \tag{3.2}$$

where $\Delta(t|\tau, x^*(\tau)) = \psi'(t|\tau, x^*(\tau))b$, $\dot{\psi} = -A'\psi$, $\psi(t^*) = c - H'y(\tau, x^*(\tau))$, $y(\tau, x^*(\tau))$ is the Lagrange optimal vector.

From (3.2) it is obvious that the open-loop solution to problem (3.1) is completely defined by the set

$$\text{æ}(\tau) = \big(t_1 = t_1(\tau, x^*(\tau)), \ldots, t_p = t_p(\tau, x^*(\tau)); \quad y = y(\tau, x^*(\tau))\big), \tag{3.3}$$

consisting of switching points of the OC and the Lagrange vector. Set (3.3) is said to be the defining elements of the OC. They satisfy the equations

$$F(\text{æ}|\tau) = 0 \Longleftrightarrow \begin{cases} Hx(t^*) = g, \\ \Delta(t_k|\tau, x^*(\tau)) = 0, \quad k = \overline{1,p} = p(\tau, x^*(\tau)). \end{cases} \tag{3.4}$$

Equations (3.4) are said to be the defining equations of the optimal controller. When $\dot{\Delta}(t_k) \ne 0$, $k = \overline{1,p}$, the Jacobi matrix of (3.4) is nonsingular on the OC with respect to variables (3.3). So the solution to the defining equations can be constructed by the Newton method if a good enough initial approximation is known.

Based on these properties, it is possible to propose the following algorithm of operating the optimal controller. The algorithm consists of three procedures: (1) the starting procedure, (2) the procedure of constructing the defining elements on regular intervals and on sliding intervals, and (3) the procedure of changing the structure (the set of the defining equations and variables).

The first procedure constructs the initial values

$$\text{æ}(0) = \big(t_1(0, x_0), \ldots, t_{p(0)}(0, x_0); \quad y(0, x_0)\big) \tag{3.5}$$

of defining elements (3.3). For this purpose, the optimal open-loop control $u^0(t|0, x_0)$, $t \in T$, of problem (2.1) is constructed before the beginning of the control process, based on a priori information. In doing so there are no limitations on the time of calculation. The switching points and the Lagrange vector of the constructed solution set vector (3.5).

The second procedure is used on intervals of control where the number of defining equations (3.4) does not change and on intervals where sliding modes arise ((2.4) does not have the classical solution). In the first case, the earlier constructed solution æ(τ) is used as an initial approximation for constructing a solution æ$(\tau + h)$ to the defining equations at the moment $\tau + h$. The number $h > 0$ is chosen in such a way that it would take no more than three iterations to obtain the solution æ$(\tau + h)$ by the Newton method with given accuracy. This allows us to evaluate the number of calculations. If the time the computer device takes to carry out this work does not exceed h units, one can say that the optimal controller using this device realizes the optimal feedback in real time. The appropriate estimates are easy to obtain from the properties of the Newton method [5]. The algorithm of operating the optimal

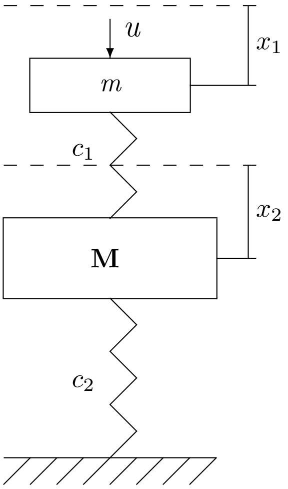


FIG. 3.1.

controller changes on sliding intervals. In [2] several alternative algorithms for sliding intervals are described.

The third procedure is intended for revealing moments for changing the structure of the defining equations and for calculating initial values for the defining elements on new intervals of constancy of the structure. Details are presented in [32, 33].

The described approach to solving the synthesis problem of optimal systems allows us, on the one hand, to select a computer device able to realize the optimal feedback for the given OC problem and, on the other hand, to point out problems in which the given computer device may be used for synthesizing the OC.

As an example of using the described approach, consider the synthesis of the optimal feedback for the problem of damping a two-mass oscillating system (Figure 3.1).

Denote by m, M masses of the objects, by x_1, x_2 their coordinates, by c_1, c_2 the coefficients of elasticity of springs, and by u the damping action. The mathematical model of the problem has the form

$$\int_0^{t^*} u(t)dt \longrightarrow \min, \quad \dot{x}_1 = x_3, \quad \dot{x}_2 = x_4,$$
$$\dot{x}_3 = (-c_1x_1 + c_1x_2 + u)/m, \quad \dot{x}_4 = (c_1x_1 - (c_1 + c_2)x_2)/M,$$
$$x(0) = \big(x_1(0), x_2(0), x_3(0), x_4(0)\big) = x_0, \quad x(t^*) = 0, \quad 0 \le u(t) \le 1, \quad t \in [0, t^*[.$$

Set the following values of parameters: $m = 1$, $M = 10$, $c_1 = 1$, $c_2 = 9.2$, $t^* = 20$. The vector $x_0 = (2, 0.5, 0, 0)$ is chosen as an initial state. Introducing a new variable $\dot{x}_5 = u$, $x_5(0) = 0$, we obtain the problem of form (2.1)

$$\begin{gathered} -x_5 \longrightarrow \max, \quad \dot{x}_1 = x_3, \quad \dot{x}_2 = x_4, \\ \dot{x}_3 = -x_1 + x_2 + u, \quad \dot{x}_4 = 0.1x_1 - 1.02x_2, \\ x_i(0) = x_{0i}, \ i = \overline{1,4}, \ x_5(0) = 0, \quad x_i(20) = 0, \ i = \overline{1,4}, \quad 0 \le u(t) \le 1, \ t \in [0, 20[. \end{gathered} \tag{3.6}$$

In Figure 3.2 dashed lines present the phase trajectories of the objects under the optimal open-loop control. The optimal value of the control criterion equals 3.040471.

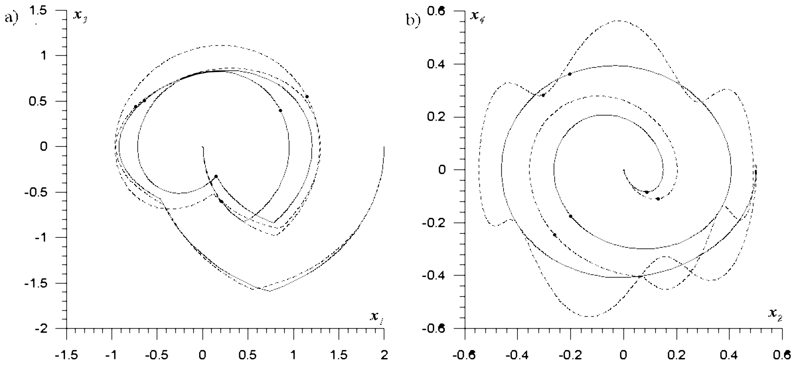


FIG. 3.2.

Let the unknown for the controller disturbance $w(t) = 0.5 \sin 5t$, $t \in [0, 9[$, $w(t) = 0$, $t \in [9, 20]$, affect the system so that the behavior of the system is described by the equations

$$\dot{x}_1 = x_3, \quad \dot{x}_2 = x_4,$$
$$\dot{x}_3 = -x_1 + x_2 + u, \quad \dot{x}_4 = 0.1x_1 - 1.02x_2 + w(t).$$

The phase trajectories corresponding to the realized disturbance and the control produced by the optimal controller are presented in Figure 3.2 (solid lines). The control criterion takes the value 3.210243. In the curves the states of the system realized at the moments $t = 5$, 10, 15 are marked.

4. Synthesis of optimal discrete controls of feedback type. When including discrete-acting devices in control channels, it is natural to use discrete controls. A control $u(t)$, $t \in T$, is said to be discrete with the quantization period $\nu > 0$ if it takes the form

$$u(t) = u_k, \quad t \in [k\nu, (k+1)\nu[, \quad k = 0, 1, 2, \dots,$$

i.e., a discrete control is described by a piecewise constant function in which discontinuities are possible only at moments $t_k = k\nu$, $k = 0, 1, 2, \dots$.

Consider problem (2.1) once again but this time in the class of discrete controls, assuming for simplicity that $t^* = N\nu$.

To define the optimal discrete control of feedback type, the family of problems (2.2) will be considered for $\tau = k\nu$, $k = 0, 1, \dots, N-1$, $z \in R^n$. The optimal discrete control of feedback type is again defined by formula (2.3), where $\tau = k\nu$, $k = \overline{0, N-1}$. Introduction of discrete time does not facilitate the search of function (2.3) considerably. Therefore, in the discrete case, by analogy with the continuous case, we introduce the concept of realization $u^*(t) = u^0(t, x^*(t))$, $t = k\nu$, $k = \overline{0, N-1}$, of the optimal discrete feedback and the concept of the optimal discrete controller. However, the algorithm of operating the optimal discrete controller differs fundamentally from the analogous one in the continuous case [43, 44, 45].

Suppose once again that the optimal discrete controller has been constructed and the control process is in the position $(\tau = k\nu, x^*(\tau))$. To calculate the value $u^*(k\nu)$ the discrete controller has to solve problem (2.2) with $\tau = k\nu$, $z = x^*(k\nu)$. This problem is equivalent to the linear programming (LP) problem. By the assumption,

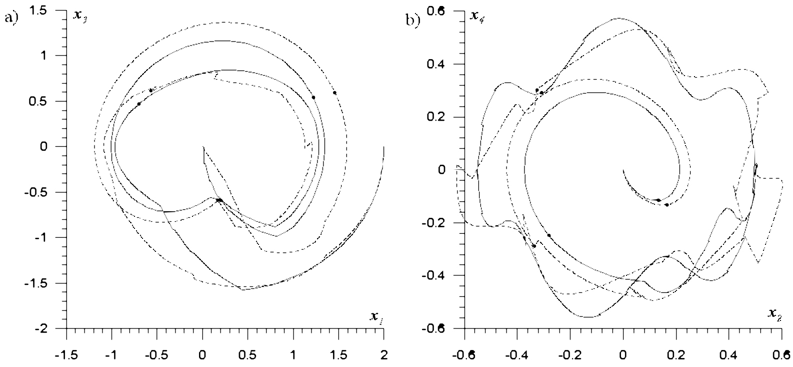


FIG. 4.1.

the controller has already faced the similar problem at the previous moment $\tau - \nu = (k-1)\nu$. Therefore instead of solving a new LP problem for τ, it is sufficient to correct the previous solution. In LP, there exists an extraordinarily effective dual method [10, 13] performing this work in a small number of iterations. If the mentioned work is performed by the available computer device in a time not exceeding ν, one can remark again about the realization of the optimal discrete feedback in real time. The initial value $u^*(0)$ is obtained after solving problem (2.2) with $\tau = 0$, $z = x_0$. This problem can be solved beforehand because it does not contain uncertain elements. In the discrete case the procedure for constructing a new structure is eliminated and the question of sliding modes does not arise. From this point of view, optimal discrete controllers are especially convenient for practical problems.

In conclusion, let us describe the starting procedure. The optimal controller begins its work at the moment $t = 0$. In calculation of the initial control signal $u^*(0) = u^0(0, x^*(0))$, two situations are possible: (1) the initial state $x^*(0)$ is known beforehand (up to the moment $t = 0$), and (2) it is only known that initial state $x^*(0)$ belongs to the bounded set G_0. In the case (1), before starting the control process, the open-loop solution $u^0(t|0, x^*(0))$, $t \in T$, is constructed based on a priori available information. In doing so, there are no restrictions on the time of solving the problem. Following definition (2.6), at the initial moment $t = 0$ the optimal controller sets $u^*(0) = u^0(0|0, x^*(0))$. In the case (2) the domain G_0 is covered with sufficiently dense finite grid. Before the beginning of the control process the optimal controller calculates optimal open-loop controls $u^0(t|0, x^{(i)})$, $t \in T$, for nodes $x^{(i)}$, $i = \overline{1, q}$, of the grid. At the initial moment $t = 0$ of the control process any initial state $x^*(0) \in G_0$ is realized. The optimal controller seeks the closest to $x^*(0)$ node $x^{(k)}$. According to the above described scheme, the controller corrects the optimal open-loop control $u^0(t|0, x^{(k)})$, $t \in T$, to the control $u^0(t|0, x^*(0))$, $t \in T$, by the dual method and sets $u^*(0) = u^0(0|0, x^*(0))$. It is clear that for a given domain G_0 one can choose a grid such that it takes no more than ν units of time to calculate $u^*(0)$. Then the whole control process can be realized in real time.

Operation of the optimal discrete controller is illustrated by the example considered earlier (section 3). Choose various values of the quantization period $\nu = 0.2$, $\nu = 1$. On the optimal open-loop control the value of the control criterion is equal to 3.046312 for $\nu = 0.2$, and 3.216836 for $\nu = 1$. As in section 3, the system is affected by unknown disturbance $w(t) = 0.5 \sin 5t$, $t \in [0, 9[$; $w(t) = 0$, $t \in [9, 20]$. The phase trajectories of the control objects are presented in Figure 4.1. Solid lines correspond to $\nu = 0.2$, and dash lines correspond to $\nu = 1$. The points identify the states of the system realized at the moments $t = 5$, 10, 15. The value of the control criterion is equal to 3.292034 for $\nu = 0.2$ and 4.223683 for $\nu = 1$.

5. Realization of combined open-closed-loop solutions to optimal control problems. Generalization of positional (section 3) and discrete (section 4) solutions to OC problems is a combined open-closed-loop solution. Consider a family of problems (2.2) in the class of piecewise continuous controls.

Definition. For a given $\nu = t^*/M > 0$ the function

$$u^0_\nu(t,x), \quad t \in [0,\nu[, \quad x \in X_\tau, \quad \tau \in T, \tag{5.1}$$

is said to be a combined open-closed-loop solution to problem (2.2) if

(1) $u^0_\nu(t,x) = u^0(\tau + t|\tau, x)$, $t \in [0,\nu[$, $x \in X_\tau$, $\tau \in T$,

(2) the trajectory $x(t)$, $t \in T$, of the system

$$\dot{x} = Ax + bu^0_\nu(t,x), \quad x(0) = x_0,$$

obtained from (2.1) with the help of closure by combined open-closed-loop feedback (5.1) presents a continuous solution to the control system

$$\dot{x} = Ax + bu(t), \quad x(0) = x_0,$$
$$u(t) = u^0_\nu(t - k\nu, x(k\nu)), \quad t \in [k\nu, (k+1)\nu[, \quad k = \overline{0, N-1}.$$

Thus, optimal combined open-closed-loop control (5.1) is corrected not at each moment $\tau \in T$ (as it was with positional solution) but only at discrete moments $\tau = k\nu$, $k = \overline{0, N-1}$, (as in the case of discrete controls). However, unlike the discrete positional control the combined open-closed-loop control does not remain constant on intervals $[k\nu, (k+1)\nu[$, $k = \overline{0, N-1}$, and can appear to be an arbitrary piecewise continuous function.

The combined open-closed-loop solution to problem (2.1) can be realized both by the continuous controller (section 3) and by its discrete analogue (section 4). For definiteness we describe only the algorithm of operating the second controller. (The continuous one is described in [27].) Restrict the class of accessible controls of problem (2.1) by discrete controls with quantization period $\mu = t^*/M$, assuming for simplicity that $\nu = K\mu$, $K \geq 1$, is an integer.

Up to the initial moment $\tau = 0$ the optimal controller solves problem (2.1) in the class of discrete controls with quantization period μ. (This problem is equivalent to the LP problem.) The controller feeds the constructed optimal open-loop control $u^0(t|0, x_0)$, $t \in T$, into the input of the system during the period of time $[0,\nu[$. Suppose that the controller has operated at the moments $0, \nu, \dots, (k-1)\nu$ and the control system at the moment $\tau = k\nu$ appears in the state $x^*(\tau)$. Using the results of the solution to LP problem at the moment $\tau - \nu$, the controller constructs the solution to the current LP problem at the moment τ using the dual method [13, 18].

The constructed optimal open-loop control $u^0(t|\tau, x^*(\tau))$, $t \in [\tau, t^*]$, is fed into the input of system (2.4) on the interval $[\tau, \tau + \nu[$.

Remark. For a large ν the correction of the OC may require a significant number of iterations of the dual method. To reduce the time of correction, one can correct the solutions to the LP problems at moments $\tau - \nu + \mu$, $\tau - \nu + 2\mu$, …, $\tau - \mu$. For small $\mu > 0$ these corrections require small work. Then the control $u^*(t)$, $t \in [\tau, \tau + \nu[$, is obtained by correcting the control $u^0(t|\tau - \mu, x^*(\tau - \mu))$, $t \in [\tau - \mu, t^*]$. Clearly, results of intermediate corrections at moments $\tau - \nu + \mu$, …, $\tau - \mu$ are not used in control of system (2.4).

6. Optimal controls of feedback type for systems under disturbances. As was mentioned in the classical statement of the optimal synthesis problem the OC of feedback type is constructed by the determined model even though it is intended for systems functioning under unknown disturbances. And in so doing, no information on disturbances is used; it is incorrect to remark on estimating the quality of the feedback with respect to disturbances. Engineers, of course, understood [12] that the reasonable use of information on disturbances can only increase the quality of the feedback. However, it was clear that in the new statement the optimal synthesis problem was considerably more difficult than the classical synthesis problem and for its solution new methods are required.

In studies of systems under disturbances it is necessary, first of all, to agree about a model of disturbances. At present time, two types of models of disturbances are distinguished: stochastic and nonstochastic. In the first case, disturbances are described in terms of the theory of stochastic processes, and in the second case only the class of possible realizations and the set of possible values of disturbances are given. Effective results on synthesis of optimal systems with the use of the first model of disturbances were obtained in stochastic analogues of the Kalman–Letov problem with the Gauss disturbances [50]. The investigations of this trend are not discussed in this work. We are dealing with OC problems based on nonstochastic models of disturbances [26, 35]. The possibility of investigation of similar problems appeared only with the creation of the OCT, with its development on differential games and nonsmooth analysis.

The inclusion of disturbances in the mathematical model of the OC problem, first of all, considerably expands the set of possible types of optimal feedback. It was, at first, noted by S. E. Dreyfus [11] in one stochastic extremal problem. In OC problems with disturbances four types of feedback are distinguished: (1) unclosed feedbacks (open-loop controls) which are closed only at the initial moment, i.e., use only a priori information; (2) unclosable feedbacks (an analogue of the classical feedback for systems without disturbances) which are closed at each current moment but do not use the fact that the closure is possible in future moments too, i.e., their action is based only on current information on the state of the system; (3) closable feedbacks which are closed at each current moment and additionally take into account that closure certainly will happen at some given future moments; (4) closed feedbacks which use information on the current states of the system and know that such information will be accessible at all future moments.

Extreme (the first and the fourth) types of feedback are not considered below because the first type does not relate to the discussed topic and the second one is so complicated that at the present moment there are not any interesting constructive results for it.

The unclosable feedback in the framework of the accepted approach in this paper was investigated in [37]. Let us briefly give only the principal moments of this work.

Instead of problem (2.1), let us introduce the problem

$$\begin{gathered} J(u) = c'x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu + dw, \quad x(0) = x_0, \\ x(t^*) \in X^* = \{x \in R^n : \ h_i'x \geq g_i, \ i = \overline{1,m}\}, \quad |u(t)| \leq 1, \quad t \in T, \end{gathered} \tag{6.1}$$

where $w(t)$, $t \in T$, is a piecewise continuous disturbance satisfying the condition

$$|w(t)| \leq 1, \quad t \in T.$$

To define the optimal unclosable feedback, problem (6.1) is imbedded into the family of problems

$$\begin{gathered} c'x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu + dw, \quad x(\tau) = z, \\ x(t^*) \in X^*, \quad |u(t)| \leq 1, \quad t \in T(\tau). \end{gathered} \tag{6.2}$$

A piecewise continuous function $u(t)$, $|u(t)| \leq 1$, $t \in T(\tau)$, is said to be an accessible (guaranteeing) open-loop control for the position (τ, z) if all solutions (for all possible $w(t)$, $t \in T(\tau)$) to the equation

$$\dot{x} = Ax + bu(t) + dw(t), \quad x(\tau) = z, \tag{6.3}$$

satisfy the condition

$$x(t^*) \in X^*. \tag{6.4}$$

Denote by $X(t|u_\tau(\cdot))$ $\big(u_\tau(\cdot) = (u(t),\, t \in T(\tau))\big)$ the set of all possible states $x(t)$ of system (6.3). Then requirement (6.4) can be written in the form $X(t^*|u_\tau(\cdot)) \subset X^*$.

The quality of the accessible control $u_\tau(\cdot)$ is evaluated by the value of the functional

$$J(u_\tau) = \min c'x, \quad x \in X(t^*|u_\tau(\cdot)).$$

An accessible control $u^0(t|\tau, z)$, $t \in T(\tau)$, is said to be an optimal (guaranteeing) open-loop control for the position (τ, z) if

$$J(u^0_\tau) = \max J(u_\tau).$$

From here at $\tau = 0$, $z = x_0$ we obtain the exact nature of problem (6.2).

Let X_τ be a set of all the vectors for which problem (6.2) with a fixed τ has an open-loop solution $u^0(t|\tau, z)$, $t \in T(\tau)$.

Definition. The function

$$u^0(\tau, z) = u^0(\tau|\tau, z), \quad z \in X_\tau, \quad \tau \in T,$$

is said to be an optimal (guaranteeing) control of feedback type (or, briefly, optimal unclosable feedback).

Realization $u^*(t) = u^0(t, x^*(t))$, $t \in T$, of the optimal unclosable feedback and the optimal controller calculating this realization are defined by analogy with section 3.

To describe an algorithm of operating the optimal controller, let us consider an arbitrary current moment τ and a current state $x^*(\tau)$. In the position $(\tau, x^*(\tau))$, in order to calculate the current value $u^*(\tau)$ of the realization of the optimal unclosable feedback, the optimal controller has to know the open-loop guaranteeing solution to the following problem:

$$\begin{gathered} c'x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu + dw, \quad x(\tau) = x^*(\tau), \\ x(t^*) \in X^*, \quad |u(t)| \leq 1, \quad t \in T(\tau). \end{gathered}$$

In [37] it is shown that the solution $u^0(t|\tau, x^*(\tau))$, $t \in T(\tau)$, to this problem coincides with the optimal open-loop control of the determined problem

$$\begin{gathered} c'x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu, \quad x(\tau) = 0, \\ h_i'x(t^*) \geq g_i(\tau), \quad i = \overline{1, m}; \quad |u(t)| \leq 1, \quad t \in T(\tau), \end{gathered} \tag{6.5}$$

where $g_i(\tau) = g_i - h_i' F(t^* - \tau)x^*(\tau) + \int_\tau^{t^*} |h_i' F(t^* - t)d|dt$, $F(t)$, $t \geq 0$, is a fundamental matrix of solutions to the homogeneous equation $\dot{x} = Ax$.

The optimal controller for problem (6.5) is constructed by the scheme of section 3. It will be an optimal controller for problem (6.1). At this point, the process of solving the synthesis problem of the optimal unclosable feedback is finished.

It is more difficult to construct a realization of an optimal closable feedback. This work was carried out in [25, 29, 30, 31], but due to massive calculations it is not possible to give the obtained results here. From the analysis of the method [29] it is obvious that the complexity of the optimal closable feedback grows rapidly as the number of moments of closure increases. However, in doing so, the quality of the optimal closable feedback approaches the optimal closed feedback which "in principle" is obtained from the Bellman–Isaacs equation. So it may be inferred that optimal closed feedback is such an ideal that cannot be achieved as yet but there are ways of steering (with appropriate efforts) to as small as one likes its neighborhood.

7. Optimal output feedbacks. So far it has been assumed that at each current moment $\tau \in T$ of the control process a complete and exact information on the current state $x^*(\tau)$ of the control system is available. Therein lies one of the features of the classical statement of the synthesis problem. It was possible for this assumption to be accepted in the 50's as the order of systems being studied was not very high. But today this assumption is considered too restrictive because available information on states of modern complex control systems is, as a rule, incomplete and inexact. Now one comes to a decision based on the readings of a measurer able to measure only some output signals of the system with limited accuracy.

That is why the output feedbacks defined not on system states but on its available output signals are urgent for the modern theory of control. Feedbacks described in previous sections are usually called state feedbacks. The problem of constructing optimal output feedbacks has been much studied in the framework of the Kalman–Letov stochastic problem [50]. Below, only nonstochastic OC problems are considered.

The essence of the new problem is illustrated by the simplest problem. More complex statements are studied in [20, 23, 24, 36]. In formal terms, the problem has the form

$$\begin{gathered} h_0' x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu, \\ x(0) = x_0 \in X_0 = \{x \in R^n : Dx = \delta, \;\; d_* \leq x \leq d^*\}, \quad x(t^*) \in X^*, \\ y(t) = c'x(t) + \xi(t), \quad \xi_* \leq \xi(t) \leq \xi^*, \quad |u(t)| \leq 1, \quad t \in T, \\ \operatorname{rank} D = l < n. \end{gathered} \tag{7.1}$$

The problem has the following sense. The initial state of the optimized system is not known exactly. A priori information on the initial state is exhausted by the inclusion $x_0 \in X_0$. By analogy with the theory of filtration we say that the set X_0 is an a priori distribution of the initial state of the control system, omitting the adjective "probable." In the control process current states are not available; only the output signal $c'x(t)$, $t \in T$, of the system is measured and the measurer does it with the error $\xi(t)$, $t \in T$, so at each moment of the control only a number $y(t)$ is at our disposal. As an error of measurement, any piecewise continuous function $\xi(t)$, $t \in T$, satisfying the inequalities $\xi_* \leq \xi(t) \leq \xi^*$, $t \in T$, may be realized. At each moment $\tau \in T$ based on the written signal $y_\tau(\cdot) = (y(t), t \in T_\tau = [0, \tau])$, the optimal controller has to produce controls $u^0(\tau, y_\tau(\cdot))$, $|u^0(\tau, y_\tau(\cdot))| \leq 1$, $\tau \in T$, which at the moment $t = t^*$ steer the system to the terminal set X^* with assurance and give the maximal guaranteed value for the control criterion $h_0'x(t^*)$.

For more precise definitions let us consider a particular control process at any current moment $\tau \in T$. Denote by $y_\tau^*(\cdot) = (y^*(t),\ t \in T_\tau)$ a signal of the measurer written by that time, and by $u_\tau^*(\cdot)$ the control used (fed into the system). As the information, available at the moment τ, contains, in the general case, additional knowledge of the initial state x_0^* realized in the process, so let us use this knowledge to decrease an a priori uncertainty. A vector $x^* \in X_0$ is called an initial state compatible with the functions $u_\tau^*(\cdot)$, $y_\tau^*(\cdot)$ if the corresponding trajectory of system (7.1) together with some possible function of errors $\xi^*(t)$, $t \in T_\tau$, generates the signal $c'x^*(t)+\xi^*(t)$, $t \in T_\tau$, coinciding with the written signal $y^*(t)$, $t \in T_\tau$: $y^*(t) \equiv c'x^*(t)+\xi^*(t)$, $t \in T_\tau$.

Denote by $\hat{X}_0(\tau)$ the set of all vectors $x^* \in X_0$ compatible with $u_\tau^*(\cdot)$, $y_\tau^*(\cdot)$, and call it the τ-a-posteriori distribution of the initial state. The set

$$\hat{X}_\tau(\tau) = \left\{ x : x = F(\tau)x_0 + \int_0^\tau F(\tau - s)bu^*(t)dt, \quad x_0 \in \hat{X}_0(\tau) \right\}$$

is said to be the τ-a-posteriori distribution of the state $x(\tau)$.

A piecewise continuous function $u_\tau(\cdot) = (u(t),\ t \in T(\tau))$, $|u(t)| \le 1$, $t \in T(\tau)$, is said to be an admissible (guaranteeing) control if at the moment $t = t^*$ it steers all states from the set $\hat{X}_\tau(\tau)$ to the terminal set X^*. Let

$$X(t^*|u_\tau(\cdot)) = \left\{ x : x = F(t^* - \tau)x_\tau + \int_\tau^{t^*} F(t^* - s)bu(s)ds, \quad x_\tau \in \hat{X}_\tau(\tau) \right\}.$$

Then the admissibility of the control $u_\tau(\cdot)$ can be written briefly: $X(t^*|u_\tau(\cdot)) \subset X^*$.

As a control criterion, the functional

$$J(u_\tau(\cdot)) = \min h_0'x, \quad x \in X(t^*|u_\tau(\cdot)),$$

is considered.

An admissible control $u^0(t|\tau, \hat{X}_\tau(\tau))$, $t \in T(\tau)$, is called a τ-a-posteriori optimal (guaranteeing) open-loop control [21] if it satisfies the equality

$$J(u_\tau^0(\cdot)) = \max J(u_\tau(\cdot)).$$

The optimal (guaranteeing) output feedback is defined by the equality

$$u^0(\tau, y_\tau(\cdot)) = u^0(\tau|\tau, \hat{X}_\tau(\tau)), \quad y_\tau(\cdot) \in Y_\tau, \quad \tau \in T,$$

where Y_τ is the set of all signals of measurer for which there exists $u^0(t|\tau, \hat{X}_\tau(\tau))$, $t \in T(\tau)$.

Following section 2, one can introduce the notion of the realization of the optimal output feedback $u^*(\tau) = u(\tau, y_\tau^*(\cdot))$, $\tau \in T$, and the optimal controller able to calculate its values in real time.

It is shown in [36] that for calculating $u^*(\tau)$ it is sufficient to solve the determined OC problem

$$\begin{gathered} h_0'x(t^*) \longrightarrow \max, \quad \dot{x} = Ax + bu, \quad x(\tau) = 0, \\ h_i'x(t^*) \ge g_i(\tau), \quad i = \overline{1,m}, \quad |u(t)| \le 1, \quad t \in T(\tau), \end{gathered} \tag{7.2}$$

in real time, where $g_i(\tau) = g_i - \alpha_i(\tau)$, $i = \overline{1,m}$, are obtained by solving m linear problems of OO

$$\alpha_i(\tau) = \min_{x \in \hat{X}_\tau(\tau)} h_i' F(t^* - \tau)x, \quad i = \overline{1,m}, \tag{7.3}$$

in real time.

Let us dwell on the latter problem.

8. OO problem, defining equations, and algorithm of operating optimal estimator. In "pure form" the OO problem used for solving the OC problem under uncertainty consists in the following.

On the interval $T = [0, t^*]$ the behavior of the dynamic system is described by the equation

$$\dot{x} = Ax \quad (x \in R^n). \tag{8.1}$$

The initial state of system (8.1) is not known exactly. It is only known that the vector x_0 belongs to the set

$$X_0 = \{x \in R^n : Dx = \delta, \quad d_* \leq x \leq d^*\}, \tag{8.2}$$

which above was called the a priori distribution of an initial state.

To refine the actual initial state x_0^* realized in some particular process the behavior of system (8.1) is observed and readings $y(t)$, $t \in T$, of the measurer

$$y = c'x + \xi, \quad \xi_* \leq \xi(t) \leq \xi^*, \quad t \in T, \tag{8.3}$$

are obtained.

According to relation (8.3) the measurer at each moment $t \in T$ can measure the given linear combination $c'x(t) = \sum_{j=1}^{n} c_j x_j(t)$ of the components $x_1(t), \ldots, x_n(t)$ of the current state $x(t)$ of system (8.1) with an error $\xi(t)$. Suppose that the function $\xi(t)$, $t \in T$, of measuring errors, an arbitrary piecewise continuous function satisfying inequalities (8.3), may be realized.

Denote by $y(t)$, $t \in T$, a signal written by the device (8.3) in some observation process. The information inherent in the signal $y(t)$, $t \in T$, allows us to decrease the a priori uncertainty of the realized vector x_0^*.

Definition. A set $\hat{X}_0$ is called the a posteriori distribution of the initial state of system (8.1) if it consists of those and only those elements of the set X_0 which together with admissible functions $\xi(t)$, $t \in T$, can produce the written signal $y(t)$, $t \in T$.

The analytical description of the set $\hat{X}_0$ has the form: a vector $x \in \hat{X}_0$ if and only if it satisfies the relations

$$\xi_* \leq y(t) - c'F(t)x \leq \xi^*, \quad t \in T; \quad Dx = \delta, \quad d_* \leq x \leq d^*. \tag{8.4}$$

The description of the set $\hat{X}_0$ contains a continuum of inequalities; thus, its structure may be very complex for effective tabulation. Luckily, in many applied problems the set $\hat{X}_0$ is not used entirely, all one has to do is to know some of its estimates (numerical characteristics). Thus, in order to solve problem (7.1), the estimates of the form

$$p'x \longrightarrow \max, \quad x \in \hat{X}_0, \tag{8.5}$$

are necessary, i.e., the extension of the set $\hat{X}_0$ in the direction of the vector $p \in R^n$.

Problem (8.5) is called a linear OO problem. In view of (8.4) this problem is a linear semi-infinite extremal problem. Similar problems occur in various applications and have a sufficiently developed theory [47]. It is interesting to note that problem (8.5) in some sense is a "dual" to the OC problem (2.1): there is a finite number of variables but an infinite number of general constraints in problem (8.5) and, conversely, there is an infinite number of variables and a finite number of general constraints in problem (2.1).

A finite method of solving problem (8.5) is described in [48].

To define positional solutions to problem (8.5) let us imbed it into the family of the problems

$$p'x \longrightarrow \max, \quad x \in \hat{X}_0(\tau), \tag{8.6}$$

where $\hat{X}_0(\tau)$ is an a posteriori distribution of an initial state of system (8.1) corresponding to the observation $y_\tau(\cdot)$.

A vector $x^0_\tau = x^0(\tau, y_\tau(\cdot)) \in \hat{X}_0(\tau)$, which gives the control criterion of problem (8.6) the maximum value, is called the optimal open-loop solution to problem (8.6).

Let $Y(\tau)$ be a set of signals $y_\tau(\cdot)$ which can be realized on the interval $[0, \tau]$ for any initial states $x_0 \in X_0$ and any measuring errors $\xi_* \le \xi(t) \le \xi^*$, $t \in T_\tau$. The functional $x^0(\tau, y_\tau(\cdot))$, $y_\tau(\cdot) \in Y(\tau)$, $\tau \in T$, is said to be a positional solution to the OO problem (8.5).

Constructing a positional solution to the OO problem in an explicit form is an extraordinarily complex problem. As in the case of the OC problem, consider a particular observation process in which an output signal $y^*(t)$, $t \in T$, is realized. It is clear that in this process values of the functional $x^0(\tau, y_\tau(\cdot))$, $y_\tau(\cdot) \in Y(\tau)$, $\tau \in T$, are necessary only along the curve $y^*(t)$, $t \in T$.

The function $x^*(\tau) = x^0(\tau, y^*_\tau(\cdot))$, $y^*_\tau(\cdot) \in Y(\tau)$, $\tau \in T$, is said to be a realization of the positional solution to the OO problem, and any device able to calculate its values in real time is called an optimal continuous estimator.

Describe an algorithm of operating the optimal estimator.

Suppose that the estimator has been constructed and operated on the interval T_τ. Using the written signal $y^*(t)$, $t \in T_\tau$, at the moment τ the estimator has to solve problem (8.6). In detailed notation, problem (8.6) has the form

$$\begin{gathered} p'x \longrightarrow \max, \quad \alpha_*(t) \le a'(t)x \le \alpha^*(t), \quad t \in T_\tau, \\ Dx = \delta, \quad d_* \le x \le d^*, \end{gathered} \tag{8.7}$$

where $\alpha_*(t) = \xi_* - y^*(t)$, $\alpha^*(t) = \xi^* - y^*(t)$, $a'(t) = (a_j(t),\ j \in J) = -c'F(t)$, $J = \{1, 2, \dots, n\}$.

Let $\hat{x}_\tau$ be an optimal feasible point of problem (8.7). Introduce the sets where constraints of problem (8.7) are active:

$$\begin{gathered} T_a(\tau) = T_a^+(\tau) \cup T_a^-(\tau), \quad T_a^+(\tau) = \{t \in T_\tau : a'(t)\hat{x}_\tau = \alpha^*(t)\}, \\ T_a^-(\tau) = \{t \in T_\tau : a'(t)\hat{x}_\tau = \alpha_*(t)\}, \\ J^+(\tau) = \{j \in J : \hat{x}_{\tau j} = d^*_j\}, \quad J^-(\tau) = \{j \in J : \hat{x}_{\tau j} = d_{*j}\}. \end{gathered}$$

The optimal feasible point $\hat{x}_\tau$ satisfies the relations

$$\begin{gathered} \Delta_j(\tau) \ge 0 \text{ if } j \in J^-(\tau), \quad \Delta_j(\tau) \le 0 \text{ if } j \in J^+(\tau), \\ \Delta_j(\tau) = 0 \text{ if } j \in J \backslash (J^+(\tau) \cup J^-(\tau)), \end{gathered}$$

where $\Delta_j(\tau) = \sum_{t\in T_a(\tau)} a_j(t)\eta(t) + \eta' D_j + p_j$, $\eta_\tau = (\eta(t),\, t \in T_a(\tau);\, \eta)$, is an optimal dual feasible point of problem (8.7) [39]:

$$\eta(t) \geq 0, \quad t \in T_a^+(\tau); \quad \eta(t) \leq 0, \quad t \in T_a^-(\tau); \quad \eta \in R^l.$$

A pair $\{\hat{x}_\tau, \eta_\tau\}$ is called a positional solution to problem (8.7).

Introduce a set $J_0(\tau) = \{j \in J : \Delta_j(\tau) = 0\}$. A solution to problem (8.7) is called nonsingular if

(1) $\operatorname{rank}(D_j';\, a_j(t), t \in T_a(\tau);\, \dot{a}_j(t), t \in T_a(\tau)\backslash\ \{0,\tau\};\, j \in J_0(\tau)) = |J_0(\tau)|$,
(2) $\operatorname{rank}(D_j';\, a_j(t), t \in T_a(\tau);\, j \in J_0(\tau)) = |T_a(\tau)| + l$,
(3) $d_*j < \hat{x}_{\tau j} < d_j^*,\ j \in J_0(\tau)$,
(4) $\eta(t) \neq 0,\ t \in T_a(\tau)$.

Here $J_0(\tau) = \{j \in J : \Delta_j(\tau) = 0\}$. Suppose that $\{0\} \notin T_a(\tau)$, $\tau > 0$.

Let $\{\hat{x}_{\tau_*}, \eta_{\tau_*}\}$ be a nonsingular solution to (8.7) at $\tau = \tau_*$. Describe rules of constructing solution to problem (8.7) for $\tau \in]\tau_*, t^*]$. Two situations are possible: (1) $\tau_* \in T_a(\tau_*)$, (2) $\tau_* \notin T_a(\tau_*)$.

Consider case (1). Assume $a'(\tau_*)\hat{x}_{\tau_*} = \alpha_*(\tau_*)$, $\dot{a}'(\tau_*)\hat{x}_{\tau_*} \neq \dot{\alpha}_*(\tau_*)$.

As is shown in [39], the set

$$\text{(8.8)} \qquad \text{æ}(\tau) = \{\hat{x}_\tau, \eta_\tau, T_a(\tau)\},$$

consisting of the optimal primal and dual feasible points of problem (8.7) and the active moments, is sought from the equations

$$\text{(8.9)}\ F(\text{æ}|\tau) = 0 \Longleftrightarrow \begin{cases} \hat{x}_{\tau j} = d_j^*, \quad j \in J^-; \quad \hat{x}_{\tau j} = d_{*j}, \quad j \in J^+; \quad D\hat{x}_\tau = \delta, \\ a'(t_i(\tau))\hat{x}_\tau = \alpha_i(t_i(\tau)), \quad i = \overline{1,q}, \\ \dot{a}'(t_i(\tau))\hat{x}_\tau = \dot{\alpha}_i(t_i(\tau)), \quad i = \overline{1,q-1}; \quad t_q(\tau) = \tau, \\ \Delta_j(t_i(\tau)) = 0, \quad j \in J_0. \end{cases}$$

Here $q = |T_a(\tau)|$, $J_0 = J_0(\tau_*)$, $I_0 = \{1,\ldots,q\}$, $I_0^+ = \{i \in I_0 :\ t_i(\tau) \in T_a^+(\tau)\}$, $I_0^- = I_0\backslash I_0^+$, $J^+ = (J\backslash J_0) \cap J^+(\tau)$, $J^- = (J\backslash J_0) \cap J^-(\tau)$, $\alpha_i(t) = \alpha^*(t)$ for $i \in I_0^+$; $\alpha_i(t) = \alpha_*(t)$ for $i \in I_0^-$.

Equations (8.9) are said to be the defining equations of the optimal estimator. The set

$$\{q,\ J_0,\ J^+,\ J^-,\ I_0^+,\ I_0^-\}$$

is called a structure of the defining equations.

A moment $\bar{\tau}$ of changing the structure of equations (8.9) is characterized by one of the following properties.

(1) At any $t_0 \in [0,\bar{\tau}]\backslash\{t_i(\bar{\tau}),\ i = \overline{1,q}\}$ one of the inequalities $\alpha_*(t_0) \leq a'(t_0)\hat{x}_{\bar{\tau}} \leq \alpha^*(t_0)$ becomes an equality.
(2) For any $j_0 \in J_0$ one of the inequalities $d_{*j_0} \leq \hat{x}_{\bar{\tau}j_0} \leq d_{j_0}^*$ becomes an equality.
(3) For any $i_0 \in I_0\backslash\{q\}$ the equality $\eta(t_{i_0}(\bar{\tau})) = 0$ takes place.
(4) For any $j_0 \in J\backslash J_0$ the equality $\Delta_{j_0}(\bar{\tau}) = 0$ takes place.
(5) The equality $\dot{a}'(\bar{\tau})\hat{x}(\bar{\tau}) = \bar{\alpha}_q(\bar{\tau})$ takes place.

Consider case (2). For $\tau \in T^+(\tau_*)$ elements (8.8) has the form

$$\text{(8.10)} \qquad \hat{x}_\tau = \hat{x}_{\tau_*},\ \eta_\tau = \eta_{\tau_*},\ T_a(\tau) = T_a(\tau_*).$$

A solution to problem (8.7) is constructed according to rule (8.10) till a moment $\bar{\tau} \in [\tau_*, t^*]$ where $a'(\bar{\tau})\hat{x}_{\bar{\tau}} = \alpha_*(\bar{\tau})$, $\dot{a}'(\bar{\tau})\hat{x}_{\bar{\tau}} \neq \dot{\alpha}_*(\bar{\tau})$ or $a'(\bar{\tau})\hat{x}_{\bar{\tau}} = \alpha^*(\bar{\tau})$, $\dot{a}'(\bar{\tau})\hat{x}_{\bar{\tau}} \neq \dot{\alpha}^*(\bar{\tau})$.

For $\tau \in T^+(\bar{\tau})$ a solution to problem (8.7) is constructed according to case (1).

As in the case of an optimal controller, the algorithm of operating the optimal estimator consists of three procedures: (1) the starting procedure, (2) the procedure of constructing defining elements (8.8) on intervals with constant structure, and (3) the procedure of changing the structure.

Details of these procedures are described in [34, 39].

9. Optimal discrete estimator. To introduce a discrete estimator let us change some conditions of observation. The object of observation remains as before (see (8.1)). Formally (8.3) of the measurer also remains. But this time, suppose that on the interval $[0, t^*]$ measurements are written at discrete moments $t_k = k\nu$, $k = 0, 1, \dots, N$, $\nu = t^*/N$. Errors of measurements are numbers ξ_k, $k = 0, 1, \dots, N$, satisfying the inequalities $\xi_* \le \xi_k \le \xi^*$, $k = 0, 1, \dots, N$. Formally, the notion of a posteriori distribution $\hat{X}_0$ of an initial state (see section 8) remains, but its analytical form is radically changed: a vector $x_0 \in \hat{X}_0$ if and only if it satisfies the finite set of inequalities

$$\begin{gathered} \xi_* \le y(k\nu) - c'F(k\nu)x_0 \le \xi^*, \quad k = \overline{0, N}, \\ Dx_0 = \delta, \quad d_* \le x_0 \le d^*. \end{gathered}$$

Of course now, the structure of the set $\hat{X}_0$ is also essentially more complex than that of the set X_0. But this structure allows one to solve OO problems considerably more easily and without any demand on analytical properties of available signals $y(t)$, $t \in T$. This is very important for solving applied problems.

Now, the OO problem takes the form

$$\begin{gathered} p'x \longrightarrow \max, \quad \xi_* \le y(k\nu) - c'F(k\nu)x \le \xi^*, \quad k = \overline{0, N}, \\ Dx = \delta, \quad d_* \le x \le d^*, \end{gathered} \tag{9.1}$$

i.e., it is an LP problem. The optimal feasible point x^0 of problem (9.1) presents an open-loop solution to the discrete OO problem.

To introduce a positional solution to the OO problem let us imbed problem (9.1) into a family of problems (8.6) where as τ discrete moments $k\nu$, $k = \overline{0, N}$, are considered, the sets $\hat{X}_0(\tau)$ are appropriately changed to $\hat{X}_0(\tau) = \{x : \xi_* \le y(i\nu) - c'F(i\nu)x \le \xi^*,\ i = \overline{0, k},\ \ Dx = \delta,\ d_* \le x \le d^*\}$, $y_\tau(\cdot) = \{y(i\nu),\ i = \overline{0, k}\}$.

Let $x^0_\tau = x^0(\tau, y_\tau(\cdot)) \in \hat{X}_0(\tau)$ be an optimal feasible point of problem (8.6) at $\tau = k\nu$. The functional

$$x^0(\tau, y_\tau(\cdot)), \quad y_\tau(\cdot) \in Y(\tau), \quad \tau \in T_\nu = [0, \nu, \dots, N\nu], \tag{9.2}$$

is said to be a positional solution to the discrete OO problem. In spite of simplification of the OO problem in the discrete case, it is very difficult to construct functional (9.2). So let us follow the approach described above for the continuous OO problem. Suppose that functional (9.2) is constructed. Use it in a particular observation process when the signal $y^*(k\nu)$, $k = \overline{0, N}$, is written. This signal corresponds to the estimates $\alpha^*(k\nu) = p'x^0(k\nu, y^*_{k\nu}(\cdot))$ of a posteriori distribution $\hat{X}_0(k\nu)$. It is clear that in order to obtain these estimates the entire functional (9.2) is not used, but it is sufficient to know its values along the sequence $y^*(k\nu)$, $k = \overline{0, N}$. The function $x^*(k\nu) = x^0(k\nu, y^*_{k\nu}(\cdot))$, $y^*_{k\nu}(\cdot) \in Y(k\nu)$, $k = \overline{0, N}$, is said to be a discrete realization of a positional solution to the OO problem, and a device able to calculate its values in real time is called an optimal discrete estimator. Algorithms of operating optimal discrete estimators are given in [15, 44].

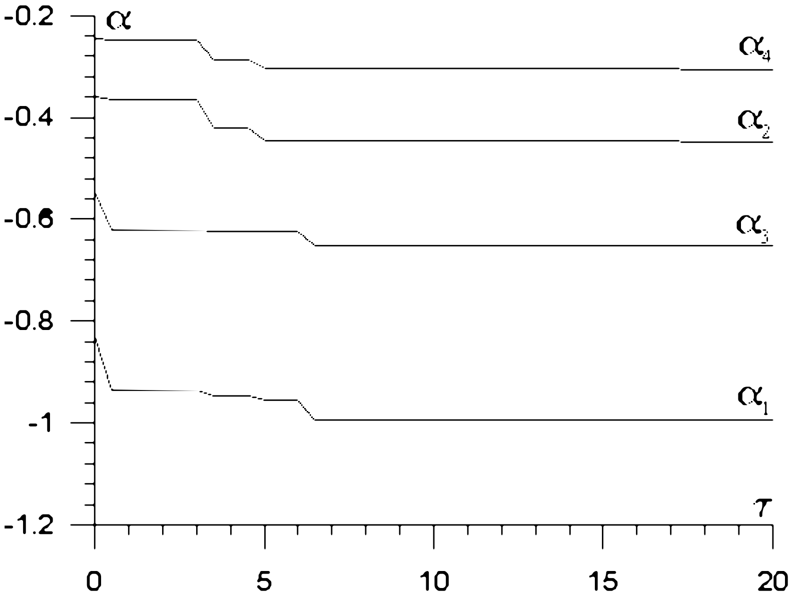


FIG. 9.1.

To describe an algorithm of operating a discrete estimator let us study a situation at an arbitrary moment $\tau = k\nu$. According to the definition, in order to calculate $\alpha^*(\tau)$, the optimal estimator needs to know a solution to the LP problem

$$\begin{gathered} p'x \longrightarrow \max, \quad \xi_* - y^*(i\nu) \le -c'F(i\nu)x \le \xi^* - y^*(i\nu), \quad i = \overline{0,k}, \\ Dx = \delta, \quad d_* \le x \le d^*. \end{gathered} \tag{9.3}$$

At the previous moment $(k-1)\nu$ the estimator dealt with the problem

$$\begin{gathered} p'x \longrightarrow \max, \quad \xi_* - y^*(i\nu) \le -c'F(i\nu)x \le \xi^* - y^*(i\nu), \quad i = \overline{0,k-1}, \\ Dx = \delta, \quad d_* \le x \le d^*. \end{gathered} \tag{9.4}$$

As problems (9.3) and (9.4) differ only by one constraint, knowing the solution to problem (9.4) makes it possible that a very few iterations would be sufficient to construct a solution $x^0(k\nu, y^*_{k\nu}(\cdot))$ to problem (9.3) with the help of the dual method of LP. If the time that it takes to do this work does not exceed ν, then one may say that the positional solution to the OO problem is realized in real time.

Return to example (3.6) and eliminate a control from the equation. The mathematical model of the system becomes of the form

$$\dot{x}_1 = x_3, \quad \dot{x}_2 = x_4, \quad \dot{x}_3 = -x_1 + x_2, \quad \dot{x}_4 = 0.1x_1 - 1.02x_2. \tag{9.5}$$

Suppose that the information on the initial state of the system is

$$x_1 = 0, \quad x_2 = 0, \quad 0.8 \le x_3 \le 1.1, \quad 0.4 \le x_4 \le 0.6,$$

and a measurer has the form

$$y = x_1 + \xi, \quad |\xi| \le 0.5.$$

Observation is conducted on the interval $T = [0, 20]$ with $\nu = 0.5$. The components of the initial state unknown for the estimator are $x_3^* = 1$, $x_4^* = 0.5$; the function of measuring errors unknown for the estimator has the form $\xi^*(t) = 0.5 \sin 2t$, $t \in T$.

In Figure 9.1 results of solving four OO problems are presented. In these problems the estimates

$$\alpha_i = x_i(t^*) \longrightarrow \max, \quad i = \overline{1, 4},$$

are calculated.

10. Observation under constantly affecting bounded disturbances. In sections 8 and 9, the OO problem, in which motion of the system takes place under ideal conditions according to equation (8.1), is considered. In this section the OO problem is studied for the case when the dynamic system is under bounded disturbances.

10.1. Observation without identification of disturbances. Suppose that on the interval $T = [0, t^*]$ the behavior of the dynamic system is described by the equation

$$\dot{x} = Ax + dw, \quad \bigl(x \in R^n, \quad w \in R\bigr), \tag{10.1}$$

where $w(t)$, $t \in T$, is an unknown piecewise continuous function of disturbances satisfying the constraint $w_* \le w(t) \le w^*$, $t \in T$.

The a priori distribution X_0 of the initial state x_0 of system (10.1) and the equation of the measurer remain in the form of (8.2), (8.3).

Reformulate the main notions of section 8 with regard to disturbances in model (10.1).

A set $\hat{X}_0$ is said to be an a posteriori distribution of the initial state of system (10.1) corresponding to the written signal $y(t)$, $t \in T$, if it consists of those and only those elements of the set X_0 which together with some admissible functions $w(t)$, $\xi(t)$, $t \in T$, can generate the signal $y(t)$, $t \in T$. The set $\hat{X}_0$ is described analytically by the following: a vector $x \in \hat{X}_0$ if and only if there exists a piecewise continuous function $w(t)$, $w_* \le w(t) \le w^*$, $t \in T$, such that the relations

$$\xi_* \le y(t) - c'F(t)x - \int_0^t c'F(t-s)dw(s)ds \le \xi^*, \quad t \in T,$$
$$Dx = \delta, \quad d_* \le x \le d^*,$$

are fulfilled.

The OO problem for system (10.1) also takes the form of (8.5), but as the description of the set $\hat{X}_0$ contains an infinite-dimensional variable $w(t)$, $t \in T$, so now problem (8.5) is an “infinite” extremal problem, i.e., it contains an unknown element of the infinite-dimensional space and a continuum of constraints.

For the purpose of numerically constructing a positional solution to problem (8.5), let us make an assumption that the function of disturbances $w(t)$, $t \in T$, is a piecewise constant function with a quantization period $\nu = t^*/N$, where N is an integer. Signals of the measurer will be correspondingly written at the moments[4] $t_k = k\nu$, $k = \overline{0, N}$.

[4] The quantization period of the disturbance $w(t)$, $t \in T$, and intervals between two written signals $y(t)$, $t \in T$, may be chosen to be different.

Then problem (8.5) takes the form

$$\begin{gathered} p'x \longrightarrow \max, \quad \xi_* - y(k\nu) \leq -c'F(k\nu)x \\ -\sum_{j=0}^{k-1} \int_{j\nu}^{(j+1)\nu} c'F(k\nu - s)ddsw_j \leq \xi^* - y(k\nu), \quad k = \overline{0,N}, \\ Dx = \delta, \quad d_* \leq x \leq d^*, \quad w_* \leq w_j \leq w^*, \quad j = \overline{0,N-1}, \end{gathered} \tag{10.2}$$

where $w_j = w(t)$, $t \in [j\nu, (j+1)\nu[$, $j = \overline{0,N-1}$.

Problem (10.2) is an LP problem with $n + N$ variables x_i, $i = \overline{1,n}$; w_j, $j = \overline{0,N-1}$. The optimal feasible point (x^0, w^0) provides the open-loop solution x^0 to the OO problem (8.5). As in the case of the OO problem without disturbances, for defining a positional solution to problem (8.5) let us imbed it into a family of problems (8.6) where

$$\begin{gathered} \hat{X}_0(\tau) = \hat{X}_0(k\nu) = \{x : \xi_* \leq y(i\nu) - c'F(i\nu)x \\ -\sum_{j=0}^{k-1} \int_{j\nu}^{(j+1)\nu} c'F(k\nu - s)ddsw_j \leq \xi^*, \ i = \overline{0,k}, \ \ Dx = \delta, \ d_* \leq x \leq d^*, \\ w_* \leq w_j \leq w^*, \ j = \overline{0,N-1}\}, \quad y_\tau(\cdot) = \{y(i\nu), \ i = \overline{0,k}\}. \end{gathered}$$

Let $x^0_\tau = x^0(\tau, y_\tau(\cdot)) \in \hat{X}_0(\tau)$ be an open-loop solution to problem (8.6) at $\tau = k\nu$. For the problem under consideration $Y(\tau)$ is the set of signals $y_\tau(\cdot)$ which can be realized on the interval $[0, \tau]$ for any initial states $x_0 \in X_0$, any disturbances $w(t)$, $|w(t)| \leq 1$, $t \in T_\tau$, and any measuring errors $\xi(t)$, $\xi_* \leq \xi(t) \leq \xi^*$, $t \in T_\tau$.

A functional

$$x^0(\tau, y_\tau(\cdot)), \quad y_\tau(\cdot) \in Y(\tau), \quad \tau \in T_\nu,$$

is said to be a positional solution to the discrete OO problem under constantly affecting disturbances.

By analogy with section 9 consider a particular observation process, in which a signal $y^*(k\nu)$, $k = \overline{0,N}$, is written, and introduce a notion of a discrete realization $x^*(k\nu) = x^0(k\nu, y^*_{k\nu}(\cdot))$, $y^*_{k\nu}(\cdot) \in Y(k\nu)$, $k = \overline{0,N}$, of a positional solution to the OO problem.

Thus, at the moment $\tau = k\nu$ in order to calculate the estimate $\alpha^*(k\nu) = p'x^0(k\nu, y^*_{k\nu}(\cdot))$, the optimal estimator solves the following LP problem:

$$\begin{gathered} p'x \longrightarrow \max, \quad \xi_* - y^*(i\nu) \leq -c'F(i\nu)x \\ -\sum_{j=0}^{i-1} \int_{j\nu}^{(j+1)\nu} c'F(i\nu - s)ddsw_j \leq \xi^* - y^*(i\nu), \quad i = \overline{0,k}, \\ Dx = \delta, \quad d_* \leq x \leq d^*, \quad w_* \leq w_j \leq w^*, \ \ j = \overline{0,k-1}. \end{gathered}$$

This problem differs from the one solved at the previous moment $(k-1)\nu$ by a new constraint and a new variable w_k. In order to solve it, the dual method of LP is used.

10.2. Optimal identification problem of parametric disturbances. Let us consider a dynamic system on the interval $T = [0, t^*]$ with a known initial state

$$\dot{x} = Ax + d\mu, \quad x(0) = x_0 \quad (x \in R^n, \quad \mu \in R). \tag{10.3}$$

Mathematical model (10.3) contains the disturbances $\mu(t)$, $t \in T$, and the following a priori information about them is known: (1) functions $\mu(t)$, $t \in T$, admit the representation

$$\mu(t) = \mu(t, w) = w_1\omega_1(t) + \cdots + w_q\omega_q(t), \quad t \in T,$$

where $\omega_1(t), \ldots, \omega_q(t)$, $t \in T$, are known piecewise continuous functions, and $w_1, \ldots, w_q$ are constant numbers; and (2) the vector $w = (w_1, \ldots, w_q)$ is not known exactly, and it can take any value from the bounded set

$$\check{W} = \{w \in R^q : f_* \leq w \leq f^*\}. \tag{10.4}$$

The set $\check{W}$ is said to be an a priori distribution of parameters of a disturbance.

The measurement of the output signal is made by measurer (8.3). In the course of observation of the output signal $y(t)$, $t \in T$, there appears additional information on realized values of parameters w of the disturbance.

A set $\hat{W}$ is called an a posteriori distribution of parameters of the disturbance of system (10.3) corresponding to the written signal $y(t)$, $t \in T$, if it consists of those and only those elements of the set $\check{W}$ which together with admissible functions $\xi(t)$, $t \in T$, can generate the signal $y(t)$, $t \in T$, for the initial state x_0.

The problem of optimal identification (OI) of parametric disturbances [42, 49] consists of calculating an estimate α of the set $\hat{W}$:

$$\alpha = \max p'w, \quad w \in \hat{W}. \tag{10.5}$$

In detailed notation, problem (10.5) takes the form

$$p'w \longrightarrow \max, \quad \xi_* - y(t) + c'F(t)x_0$$
$$\leq -\sum_{j=1}^{q} \int_0^t c'F(t-s)d\omega_j(s)dsw_j \leq \xi^* - y(t) + c'F(t)x_0, \quad f_* \leq w \leq f^*, \quad t \in T.$$

As well as problem (8.5), problem (10.5) is a semi-infinite extremal problem.

In order to define positional solutions to problem (10.5) let us imbed it into a family of problems

$$p'w \longrightarrow \max, \quad w \in \hat{W}(\tau), \tag{10.6}$$

where $\hat{W}(\tau)$ is an a posteriori distribution of parameters of the disturbance corresponding to the observation $y_\tau(\cdot)$.

Let $w^0_\tau = w^0(\tau, y_\tau(\cdot)) \in \hat{W}(\tau)$ be the optimal feasible point of problem (10.6), and let $Y(\tau)$ be the set of signals $y_\tau(\cdot)$ which can be realized on the interval $[0, \tau]$ for any parameters $w \in \check{W}$ of a disturbance and any measuring errors $\xi_* \leq \xi(t) \leq \xi^*$, $t \in T_\tau$. A functional $w^0(\tau, y_\tau(\cdot))$, $y_\tau(\cdot) \in Y(\tau)$, $\tau \in T$, is said to be a positional solution to the OI problem of parametric disturbances.

Consider some particular identification process in the course of which an output signal $y^*(t)$, $t \in T$, is realized. The function $w^*(\tau) = w^0(\tau, y^*_\tau(\cdot))$, $y^*_\tau(\cdot) \in Y(\tau)$, $\tau \in T$, is called a realization of the positional solution to the OI problem of parametric disturbances. Any device able to calculate its values in real time is said to be an optimal identifier of parametric disturbances.

In order to calculate the estimate $\alpha^*(\tau) = p'w^0(\tau, y^*_\tau(\cdot))$ the identifier solves problem (10.6) with $y(t) = y^*(t)$, $t \in T_\tau$:

$$p'w \longrightarrow \max, \quad \xi_* - y^*(t) + c'F(t)x_0$$
$$\leq -\sum_{j=1}^{q} \int_0^t c'F(t-s)d\omega_j(s)dsw_j \leq \xi^* - y^*(t) + c'F(t)x_0, \quad f_* \leq w \leq f^*, \quad t \in T_\tau.$$

In the case of constructing a discrete realization $w^*(k\nu) = w^0(k\nu, y^*_{k\nu}(\cdot))$, $y^*_{k\nu}(\cdot) \in Y(k\nu)$, $k = \overline{0, N}$, of the positional solution to the OI problem of parametric disturbances, the identifier solves the LP problem with q variables w_j, $j = \overline{1, q}$:

$$p'w \longrightarrow \max,$$
$$\xi_* - y^*(i\nu) + c'F(i\nu)x_0 \leq -\sum_{j=1}^{q} \int_0^{i\nu} c'F(i\nu - s)d\omega_j(s)dsw_j$$
$$\leq \xi^* - y^*(i\nu) + c'F(i\nu)x_0, \quad i = \overline{0, k}, \quad f_* \leq w \leq f^*,$$

using the dual method.

10.3. Observation with identification of parametric disturbances. Let us expand the above studied OI problem of parametric identification by the element of the OO problem—uncertainty in the initial state. Now we consider the situation when in system (10.3) the initial state x_0 is not known exactly; it is only known that it belongs to a given set X_0 (8.2).

Vectors $x_0 \in X_0$, $w \in \check{W}$ are said to be compatible with the written signal $y(t)$, $t \in T$, if the trajectory $x(t) = x(t|x_0, w)$, $t \in T$, corresponding to the initial state x_0 and the disturbance $\mu(t, w) = \sum_{i=1}^{q} w_i\omega_i(t)$, $t \in T$, together with any admissible error function $\xi(t)$, $t \in T$, generates the signal $y(t)$, $t \in T$.

The set of all pairs (x_0, w) of vectors $x_0 \in X_0$, $w \in \check{W}$, compatible with a $y(t)$, $t \in T$, is denoted by $\hat{X}_0 \times \hat{W}$ and is called an a posteriori distribution of the initial state and parameters of the disturbance.

The problem of optimal observation-identification (OOI) consists in calculating an estimate α of the set $\hat{X}_0 \times \hat{W}$:

$$\alpha = \max(p_1'x + p_2'w), \quad x \in \hat{X}_0, \quad w \in \hat{W}. \tag{10.7}$$

Problem (10.7) is a semi-infinite LP problem with $n+q$ variables x_i, $i = \overline{1, n}$; w_j, $j = \overline{1, q}$.

To introduce positional solutions to problem (10.7) let us imbed it into the family of problems

$$p_1'x + p_2'w \longrightarrow \max, \quad x \in \hat{X}_0(\tau), \quad w \in \hat{W}(\tau), \tag{10.8}$$

where $\hat{X}_0(\tau) \times \hat{W}(\tau)$ is an a posteriori distribution of an initial state and parameters of a disturbance corresponding to the observation $y_\tau(\cdot)$.

Let $x^0_\tau = x^0(\tau, y_\tau(\cdot)) \in \hat{X}_0(\tau)$, $w^0_\tau = w^0(\tau, y_\tau(\cdot)) \in \hat{W}(\tau)$ be an optimal feasible point of problem (10.7), and let $Y(\tau)$ be a set of signals $y_\tau(\cdot)$ which can be realized on the interval $[0, \tau]$ for any initial state $x_0 \in X_0$, any parameters $w \in \check{W}$ of a disturbance, and any measuring error $\xi_* \leq \xi(t) \leq \xi^*$, $t \in T_\tau$. A pair

$$\big(x^0(\tau, y_\tau(\cdot)), w^0(\tau, y_\tau(\cdot))\big), \quad y_\tau(\cdot) \in Y(\tau), \quad \tau \in T, \tag{10.9}$$

is said to be a positional solution to the OOI problem.

As well as in the OO problem, consider a particular process in the course of which the output signal $y^*(t)$, $t \in T$, is realized. A pair

$$(10.10)\quad \big(x^*(\tau) = x^0(\tau, y^*_\tau(\cdot)),\ w^*(\tau) = w^0(\tau, y^*_\tau(\cdot))\big), \quad y^*_\tau(\cdot) \in Y(\tau), \quad \tau \in T,$$

is said to be a realization of the positional solution to the OOI problem.

In order to calculate the estimate $\alpha^*(\tau) = p'_1 x^0(\tau, y^*_\tau(\cdot)) + p'_2 w^0(\tau, y^*_\tau(\cdot))$, problem (10.8), solved by the optimal estimator in detailed notation, takes the form

$$p'_1 x + p'_2 w \longrightarrow \max,$$
$$\xi_* - y^*(t) \le -c'F(t)x - \sum_{j=1}^{q} \int_0^t c'F(t-s)d\omega_j(s)ds w_j \le \xi^* - y^*(t),$$
$$Dx = \delta, \quad d_* \le x \le d^*, \quad f_* \le w \le f^*, \quad t \in T_\tau.$$

In the case of constructing a discrete realization $x^*(k\nu) = x^0(k\nu, y^*_{k\nu}(\cdot))$, $w^*(k\nu) = w^0(k\nu, y^*_{k\nu}(\cdot))$, $y^*_{k\nu}(\cdot) \in Y(k\nu)$, $k = \overline{0,N}$, of the positional solution to the OOI problem, the estimator solves the LP problem

$$p'_1 x + p'_2 w \longrightarrow \max,$$
$$\xi_* - y^*(i\nu) \le -c'F(i\nu)x - \sum_{j=1}^{q} \int_0^{i\nu} c'F(i\nu - s)d\omega_j(s)ds w_j$$
$$\le \xi^* - y^*(i\nu), \quad i = \overline{0,k}, \quad Dx = \delta, \quad d_* \le x \le d^*, \quad f_* \le w \le f^*,$$

by the dual method.

10.4. Observation with partial identification of parametric disturbances. Suppose that on the interval $T = [0, t^*]$ the behavior of a dynamic system is described by the equation

$$(10.11)\qquad \dot{x} = Ax + d_1\mu + d_2 v, \quad x_0 \in X_0 \quad (x \in R^n,\ \mu, v \in R).$$

System (10.11) contains the uncertainties studied in systems (10.1) and (10.3): the initial state of system (10.11) is not known exactly; it is only known that it belongs to a set X_0 (8.2); a disturbance $\mu(t)$, $t \in T$, affecting the system in the course of observation process is a linear combination of known piecewise continuous functions $\omega_j(t)$, $t \in T$, $j = \overline{1,q}$, with unknown coefficients $w = (w_j,\ j = \overline{1,q})$ from the set $\check{W}$ (10.4); furthermore, there are unknown piecewise continuous disturbances $v(t)$, $t \in T$, satisfying the constraints $v_* \le v(t) \le v^*$, $t \in T$, in the system.

Vectors $x_0 \in X_0$, $w \in \check{W}$ are said to be compatible with the written signal $y(t)$, $t \in T$, if the trajectory $x(t) = x(t|x_0, w)$, $t \in T$, of system (10.11) corresponding to the initial state x_0 and the disturbance $\mu(t, w) = \sum_{i=1}^{q} w_i\omega_i(t)$, $t \in T$, together with some admissible functions $v(t)$, $\xi(t)$, $t \in T$, generates the signal $y(t)$, $t \in T$.

A set of all vectors $x_0 \in X_0$, $w \in \check{W}$ compatible with $y(t)$, $t \in T$, is denoted by $\hat{X}_0 \times \hat{W}$ and is called an a posteriori distribution of an initial state and parameters of a disturbance.

A pair $(x, w) \in \hat{X}_0 \times \hat{W}$ if and only if there exists a piecewise continuous function $v(t)$, $v_* \le v(t) \le v^*$, $t \in T$, such that the relations

$$\xi_* \le y(t) - c'F(t)x - \sum_{j=1}^{q} \int_0^t c'F(t-s)d_1\omega_j(s)ds w_j - \int_0^t c'F(t-s)d_2 v(s)ds \le \xi^*,$$
$$Dx = \delta, \quad d_* \le x \le d^*, \quad f_* \le w \le f^*, \quad t \in T,$$

hold.

The OOI problem for system (10.11) has the form of (10.7). As there is the infinite-dimensional variable $v(t)$, $t \in T$, in description of the set $\hat{X}_0 \times \hat{W}$, so let us move on to the discrete form of the OOI problem; namely, consider that the function $v(t)$, $t \in T$, is piecewise continuous with a quantization period $\nu = t^*/N$ and measurements are conducted at the moments $t_k = k\nu$, $k = \overline{0, N}$. Problem (10.7) becomes of the form

$$
\begin{gathered}
p_1'x + p_2'w \longrightarrow \max, \\
\xi_* - y(i\nu) \le -c'F(i\nu)x - \sum_{j=1}^{q} \int_0^{i\nu} c'F(i\nu - s)d_1\omega_j(s)ds w_j \\
-\sum_{j=0}^{i-1} \int_{j\nu}^{(j+1)\nu} c'F(i\nu - s)d_2 ds v_j \le \xi^* - y(i\nu), \quad i = \overline{0, N}, \\
Dx = \delta, \quad d_* \le x \le d^*, \quad f_* \le w \le f^*, \quad v_* \le v_j \le v^*, \quad j = \overline{0, N-1},
\end{gathered}
\tag{10.12}
$$

where $v(t) = v_j$, $t \in [j\nu, (j+1)\nu[$, $j = \overline{0, N-1}$.

Problem (10.12) is an LP problem with $n+q+N$ variables x_i, $i = \overline{1, n}$; w_i, $i = \overline{1, q}$; v_i, $i = \overline{0, N-1}$. As before, in order to define positional solutions, problem (10.12) is imbedded into a family of problems (10.8), the positional solution to problem (10.12) is defined by relation (10.9), and the notion of realization of the positional solution (10.10) to the OOI problem is introduced.

To calculate the estimate $\alpha^*(k\nu) = p_1'x^0(k\nu, y_\tau^*(\cdot)) + p_2'w^0(k\nu, y_\tau^*(\cdot))$, the optimal estimator at the moment $\tau = k\nu$ solves the LP problem

$$
\begin{gathered}
p_1'x + p_2'w \longrightarrow \max, \\
\xi_* - y^*(i\nu) \le -c'F(i\nu)x - \sum_{j=1}^{q} \int_0^{i\nu} c'F(i\nu - s)d\omega_j(s)ds w_j \\
-\sum_{j=0}^{i-1} \int_{j\nu}^{(j+1)\nu} c'F(i\nu - s)d_2 ds v_j \le \xi^* - y^*(i\nu), \quad i = \overline{0, k}, \\
Dx = \delta, \quad d_* \le x \le d^*, \quad f_* \le w \le f^*, \quad v_* \le v_j \le v^*, \quad j = \overline{0, k-1},
\end{gathered}
$$

by the dual method.

10.5. Example. Consider system (9.5) under unknown bounded disturbances $w(t)$, $t \in T = [0, 20]$:

$$\dot{x}_1 = x_3, \quad \dot{x}_2 = x_4, \quad \dot{x}_3 = -x_1 + x_2, \quad \dot{x}_4 = 0.1x_1 - 1.02x_2 + w.$$

There is the following information on an initial state of the system:

$$x_1 = 0, \quad x_2 = 0, \quad 0.8 \le x_3 \le 1.1, \quad 0.4 \le x_4 \le 0.6.$$

In the course of the observation process, the output signal

$$y = x_1 + \xi, \quad |\xi| \le 0.2,$$

is available.

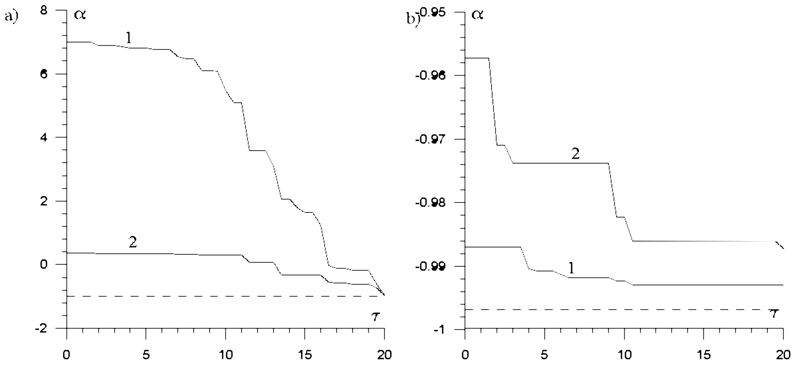


FIG. 10.1.

The parameter of the method ν is put to be equal to 0.5, the components of the initial state unknown for the estimator take the values $x_2^* = 1$, $x_4^* = 0.5$, and the function of measuring errors has the form $\xi^*(t) = 0.2\sin 3t$, $t \in T$. As an unknown disturbance, the function of the form $w(t) = w_1\omega_1(t) + w_2\omega_2(t) + w_3\omega_3(t)$, $t \in T$, is taken: $w^*(t) = 0.15\sin 0.5t + 0.1\sin 2t + 0.05\sin 5t$, $t \in T$.

In the course of the observation process, the value of the estimate $\alpha = \max x_1(20)$ is calculated. With the help of the Cauchy formula it can be written in the form

$$\alpha = \max\left(h_1' F(20)x_0 + \int_0^{20} h_1' F(20-s)w(s)ds\right), \tag{10.13}$$

where $h_1 = (1, 0, 0, 0)$.

Curve 1 in Figure 10.1a presents the behavior of the estimates $\alpha^*(\tau)$, $\tau \in T$, calculated by the optimal estimator without identification of disturbances with regard to the constraints $|w_j| \leq 0.3$, $j = \overline{0, 39}$. Curve 2 is constructed as a result of observation with partial identification of disturbances: the functions $\omega_1(t) = \sin 0.5t$, $\omega_2(t) = \sin 2t$, $t \in T$, were considered to be known for the estimator, the coefficients w_1, w_2 were identified during observation process, and the inequalities $0 \leq w_j \leq 0.2$, $j = 1, 2$, were set for them. The addend $v(t) = 0.05\sin 5t$, $t \in T$, was unknown for the estimator. The estimator knew only that this addend had satisfied the constraint $|v_j| \leq 0.05$, $j = \overline{0, 39}$.

In Figure 10.1b there are the values of the estimate $\alpha^*(\tau)$, $\tau \in T$, obtained by solving the OI problem of parametric disturbances with the known initial state (curve 1) and solving the OOI problem (curve 2). As the identified parameters, the coefficients w_1, w_2, w_3 satisfying the inequalities $0 \leq w_j \leq 0.2$, $j = 1, 2$, $0 \leq w_3 \leq 0.1$ were considered. The functions $\omega_1(t) = \sin 0.5t$, $\omega_2(t) = \sin 2t$, $\omega_3(t) = \sin 5t$, $t \in T$, were considered to be known for the estimator.

The actual value of the estimate calculated according to (10.13) on the base of the exact values x_0 and $w(t)$, $t \in T$, is equal to $\alpha^0 = -0.996845$, and it is presented by dashed lines in the figures.

11. Realization of optimal output feedback. As was noted in section 7, the calculation of the realization of the optimal output feedback $u^*(\tau)$, $\tau \in T$, is performed by solving problems (7.2), (7.3) in real time [24, 36].

Problem (7.2) differs form the previously studied OC problem (2.1) by the presence of terminal constraints-inequalities.

According to the Pontryagin maximum principle [53] the optimal open-loop control $u^0(t|\tau)$, $t \in T(\tau)$, of problem (7.2) has the form of (3.2), where the adjoint system becomes of the form

$$\dot{\psi} = -A'\psi, \quad \psi(t^*) = h_0 - \sum_{i \in I(\tau)} h_i \nu_i(\tau),$$

where $\nu(\tau) = (\nu_i(\tau),\ \ i \in I(\tau))$ is the Lagrange optimal vector, and $I(\tau) = \{i \in I : h_i' x_\tau^0(t^*) = g_i(\tau)\}$, $I = \{1, 2, \dots, m\}$, $x_\tau^0(t)$, $t \in T(\tau)$, is the optimal trajectory of problem (7.2).

Thus, the defining elements of problem (7.2) are the switching points of OC and the Lagrange vector

$$\text{æ}(\tau) = \big(t_1(\tau), \dots, t_p(\tau);\ \ \nu(\tau))\big). \tag{11.1}$$

Elements (11.1) satisfy the equations

$$F(\text{æ}|\tau) = 0 \Longleftrightarrow \begin{cases} h_i' x(t^*) = g_i(\tau), & i \in I(\tau), \\ \Delta(t_k|\tau) = 0, & k = \overline{1, p(\tau)}. \end{cases} \tag{11.2}$$

The algorithm of operating the optimal controller is analogous to the one discussed in section 3.

The values $g_i(\tau) = g_i - \alpha_i(\tau)$, $i = \overline{1, m}$, used in problem (7.2), are calculated by solving problems (7.3).

Consider one (ith) of problems (7.3). With the help of the Cauchy formula it can be written in the form

$$\begin{gathered} \alpha_i(\tau) = h_i' F(t^*) x + \int_0^\tau h_i' F(t^* - t) b u^*(t) dt \longrightarrow \min_x, \\ \alpha_*(t) \le a'(t) x \le \alpha^*(t), \quad t \in T_\tau, \quad Dx = \delta, \quad d_* \le x \le d^*, \end{gathered} \tag{11.3}$$

where $\alpha_*(t) = \xi_* - y^*(t) + \int_0^t h_i' F(t - s) b u^*(s) ds$, $\alpha^*(t) = \xi^* - y^*(t) + \int_0^t h_i' F(t - s) b u^*(s) ds$, $t \in T_\tau$, and $a'(t) = -h_i' F(t)$.

According to the scheme described in section 8, the optimal estimator at a current moment $\tau \in T$ constructs a solution $x^0(\tau, y_\tau^*(\cdot))$ to problem (11.3) and calculates an estimate $\alpha_i(\tau) = h_i' F(t^*) x^0(\tau, y_\tau^*(\cdot)) + \int_0^\tau h_i' F(t^* - t) b u^*(t) dt$.

All m problems (11.3) are solved in parallel based on the known values of the output signal $y_\tau^*(\cdot)$ and the realization of the optimal feedback $u^*(t)$, $t \in T_\tau$. Constructed values $\alpha_i(\tau)$, $i = \overline{1, m}$, are used by the optimal controller for calculating the realization $u^*(\tau)$.

In the case of constructing a discrete realization of the optimal output feedback $u^*(k\nu) = u^0(k\nu, y_{k\nu}^*(\cdot))$, $k = \overline{0, N-1}$, the optimal controller at each current moment $\tau = k\nu$ uses a solution to the LP problem

$$\begin{gathered} h_0'\sum_{j=1}^{N-k}\int_{(k+j-1)\nu}^{(k+j)\nu}F(t^*-t)bdtu_j \longrightarrow \max, \\ h_i'\sum_{j=1}^{N-k}\int_{(k+j-1)\nu}^{(k+j)\nu}F(t^*-t)bdtu_j \geq g_i(k\nu), \quad i=\overline{1,m}, \\ |u_j|\leq 1, \quad j=\overline{1,N-k}, \end{gathered} \tag{11.4}$$

which is equivalent to problem (7.2).

The solution (u_j^0, $j=\overline{1,N-k}$) to problem (11.4) is constructed by the dual method according to section 4. The value $u^*(k\nu)=u_1^0$ is fed into the input of system (7.1).

The OO problems (7.3) for the moment $\tau=k\nu$ in the discrete case take the form

$$\begin{gathered} \alpha_i(k\nu)=h_i'F(t^*)x+\sum_{j=0}^{k-1}\int_{j\nu}^{(j+1)\nu}h_i'F(t^*-t)bdtu^*(j\nu)\longrightarrow \min_x, \\ \alpha_*(j\nu)\leq a'(j\nu)x\leq \alpha^*(j\nu), \quad j=\overline{0,k}, \quad Dx=\delta, \quad d_*\leq x\leq d^*, \end{gathered} \tag{11.5}$$

where $\alpha_*(j\nu)=\xi_*-y^*(j\nu)+\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)bdsu^*(r\nu)$, $\alpha^*(j\nu)=\xi^*-y^*(j\nu)+\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)bdsu^*(r\nu)$, $a'(j\nu)=-h_i'F(j\nu)$.

According to the scheme described in section 9, the optimal estimator at a current moment $\tau\in T$ constructs a solution $x^0(k\nu,y^*_{k\nu}(\cdot))$ to problem (11.5) and calculates an estimate $\alpha_i(k\nu)=h_i'F(t^*)x^0(k\nu,y^*_{k\nu}(\cdot))+\sum_{j=0}^{k-1}\int_{j\nu}^{(j+1)\nu}h_i'F(t^*-t)bdtu^*(j\nu)$.

As in the case of continuous realization, problems (11.5) are solved in parallel by m estimators, and constructed values $\alpha_i(k\nu)$, $k=\overline{0,N-1}$, are used by the optimal controller for calculating the realization $u^*(k\nu)$.

Discrete realization of optimal output feedback is considered in detail in [28, 44].

12. Optimal output feedback under constantly affecting disturbances.

In section 11 the optimization problem of control systems with uncertainty in the initial state is considered. In this section the case in point is the optimization problem of a dynamic system under uncertainty both in an initial state and in an equation of motion [16, 17, 22, 37].

12.1. Optimization of a control system under uncertainty without identification of disturbances.

In the class of piecewise continuous controls $u(t)$, $t\in T$, consider the problem

$$\begin{gathered} J(u)=h_0'x(t^*)\longrightarrow \max, \quad \dot{x}=Ax+bu+d\mu, \\ x(0)=x_0\in X_0=\{x\in R^n : Dx=\delta, \quad d_*\leq x\leq d^*\}, \\ x(t^*)\in X^*=\{x\in R^n : \quad h_i'x\geq g_i, \ i=\overline{1,m}\}, \quad |u(t)|\leq 1, \quad t\in T, \end{gathered} \tag{12.1}$$

where $\mu(t)$, $t\in T$, is an unknown piecewise continuous function of disturbances satisfying the inequality $w_*\leq\mu(t)\leq w^*$, $t\in T$. In the course of the control process the measurement of the output signal $y=c'x+\xi$, $\xi_*\leq\xi(t)\leq\xi^*$, $t\in T$, is conducted.

Give the necessary definitions for problem (12.1).

Consider some particular control process at an arbitrary current moment $\tau\in T$. Denote by $y^*_\tau(\cdot)=(y^*(t),\ t\in T_\tau)$ a signal of the measurer written up until this moment, and denote by $u^*_\tau(\cdot)$ a control fed into the system.

A vector $x^* \in X_0$ is said to be compatible with the functions $u^*_\tau(\cdot)$, $y^*_\tau(\cdot)$ if the trajectory $x^*(t) = x^*(t|x^*)$ corresponding to the initial state x^*, and the control $u^*_\tau(\cdot)$ together with some admissible functions of disturbances $\mu^*(t)$, $t \in T_\tau$, and errors $\xi^*(t)$, $t \in T_\tau$, generates the written signal $y^*(t)$, $t \in T_\tau$. The set of all vectors $x^* \in X_0$ compatible with $u^*_\tau(\cdot)$, $y^*_\tau(\cdot)$ is denoted by $\hat{X}_0(\tau)$ and is called a τ-a-posteriori distribution of the initial state. The set

$$\hat{X}_\tau(\tau) = \left\{ x : x = F(\tau)x_0 + \int_0^\tau F(\tau - t)bu^*(t)dt + \int_0^\tau F(\tau - t)d\mu^*(t)dt, \right.$$
$$\left. x_0 \in \hat{X}_0(\tau), w_* \le \mu^*(t) \le w^*, t \in T_\tau \right\}$$

is called a τ-a-posteriori distribution of a state $x(\tau)$.

As in section 7, the notion of a τ-a-posteriori optimal open-loop control is introduced. The optimal output feedback is defined by the equation

$$u^0(\tau, y_\tau(\cdot)) = u^0(\tau|\tau, \hat{X}_\tau(\tau)), \quad y_\tau \in Y_\tau, \quad \tau \in T,$$

where Y_τ is a set of all signals of the measurer for which there exists $u^0(t|\tau, \hat{X}_\tau(\tau))$, $t \in T(\tau)$.

As there is the infinite-dimensional variable $\mu^*(t)$, $t \in T_\tau$, in description of the set $\hat{X}_\tau(\tau)$, so let us describe a way of constructing a discrete realization $u^*(k\nu) = u^0(k\nu, y^*_{k\nu}(\cdot))$, $k = \overline{0, N-1}$, of the optimal feedback. In order to calculate the value $u^*(k\nu)$ at the current moment $k\nu$, the optimal controller solves the determined LP problem (11.4) in real time. In the problem the values $g_i(k\nu) = g_i - \alpha_i(k\nu)$, $i = \overline{1, m}$, are necessary. They are calculated as a result of solving in real time m problems

$$\begin{aligned} \alpha_i(k\nu) = \min \Bigg[h_i' F(t^*)x + \sum_{j=0}^{k-1} \int_{j\nu}^{(j+1)\nu} h_i' F(t^* - s) dds\mu_j \Bigg] \\ + \sum_{j=0}^{k-1} \int_{j\nu}^{(j+1)\nu} h_i' F(t^* - s) bdsu^*(j\nu), \\ \alpha_*(j\nu) \le -h_i' F(j\nu)x - \sum_{r=0}^{j-1} \int_{r\nu}^{(r+1)\nu} h_i' F(j\nu - s) dds\mu_j \le \alpha^*(j\nu), \quad j = \overline{0, k}, \\ Dx = \delta, \quad d_* \le x \le d^*, \quad w_* \le \mu_j \le w^*, \quad j = \overline{0, N-1}, \end{aligned} \tag{12.2}$$

where $\alpha_*(j\nu) = \xi_* - y^*(j\nu) + \sum_{r=0}^{j-1} \int_{r\nu}^{(r+1)\nu} h_i' F(j\nu - s) bdsu^*(r\nu)$ and $\alpha^*(j\nu) = \xi^* - y^*(j\nu) + \sum_{r=0}^{j-1} \int_{r\nu}^{(r+1)\nu} h_i' F(j\nu - s) bdsu^*(r\nu)$.

The optimal estimator at a current moment $k\nu$ constructs a solution $(x^0; \mu_j, j = \overline{0, N-1})$ to problem (12.2) and calculates an estimate $\alpha_i(k\nu)$. Problems (12.2) are solved in parallel mode by m estimators. Calculated values $\alpha_i(k\nu)$, $i = \overline{1, m}$, are given to the optimal controller. The optimal controller constructs the solution $(u^0_j, j = \overline{1, N-k})$ to problem (11.4) by the dual method and feeds the value $u^*(k\nu) = u^0_1$ into the system input.

12.2. Optimization of control systems under uncertainty with identification of parametric disturbances. Let a disturbance $\mu(t)$, $t \in T$, in problem (12.1) contain known piecewise continuous functions $\omega_j(t)$, $t \in T$, $j = \overline{1,q}$, and unknown parameters $w = (w_j,\ j = \overline{1,q})$ from the set $\check{W}$:

$$\begin{aligned} &\mu(t) = \mu(t,w) = w_1\omega_1(t) + \cdots + w_q\omega_q(t), \quad t \in T,\\ &w = (w_1, \ldots, w_q) \in \check{W} = \{w \in R^q : f_* \leq w \leq f^*\}. \end{aligned}$$

Suppose that in some particular process by the time $\tau \in T$ an output signal $y^*_\tau(\cdot)$ and a control $u^*_\tau(\cdot)$ have been realized.

Vectors $x^* \in X_0$, $w^* \in \check{W}$ are said to be compatible with the functions $u^*_\tau(\cdot)$, $y^*_\tau(\cdot)$ if the trajectory $x^*(t) = x^*(t|x^*, w^*)$, $t \in T_\tau$, corresponding to the initial state x^*, the disturbance $\mu(t, w^*)$, $t \in T_\tau$, and the control $u^*_\tau(\cdot)$ together with some admissible error function $\xi^*(t)$, $t \in T_\tau$, generates the signal $y^*(t)$, $t \in T_\tau$. The set of all pairs (x^*, w^*) of vectors $x^* \in X_0$, $w^* \in \check{W}$ compatible with $u^*_\tau(\cdot)$, $y^*_\tau(\cdot)$ is denoted by $\hat{X}_0(\tau) \times \hat{W}(\tau)$ and is called a τ-a-posteriori distribution of an initial state and parameters of a disturbance. The set

$$\hat{X}_\tau(\tau) = \left\{ x : x = F(\tau)x_0 + \int_0^\tau F(\tau - t)bu^*(t)dt + \int_0^\tau F(\tau - t)d\mu(t, w^*)dt, \right.$$
$$\left. x_0 \in \hat{X}_0(\tau), \quad w^* \in \hat{W}(\tau) \right\}$$

is called a τ-a-posteriori distribution of a state $x(\tau)$. As in section 7, introduce the notion of an optimal output feedback

$$u^0(\tau, y_\tau(\cdot)) = u^0(\tau|\tau, \hat{X}_\tau(\tau)), \quad y_\tau \in Y_\tau, \quad \tau \in T, \tag{12.3}$$

where Y_τ is a set of all admissible signals of the measurer for which there exists a τ-a-posteriori optimal open-loop control $u^0(t|\tau, \hat{X}_\tau(\tau))$, $t \in T(\tau)$.

Similarly, the notion of a realization of the optimal output feedback $u^*(\tau) = u^0(\tau, y^*_\tau(\cdot))$, $\tau \in T$, and an optimal controller able to construct its values in real time are introduced.

It can be shown that values $u^*(\tau)$, $\tau \in T$, are calculated by solving the determined OC problem (7.2) in real time, where

$$g_i(\tau) = g_i - \alpha_i(\tau), \quad i = \overline{1,m},$$

$\alpha_i(\tau)$, $i = \overline{1,m}$, are estimates calculated as a result of solving in real time m OOI problems

$$\alpha_i(\tau) = \min \left[h_i' F(t^*)x + \sum_{j=1}^q \int_0^\tau h_i' F(t^* - t)d\omega_j(t)dtw_j \right] + \int_0^\tau h_i' F(t^* - t)bu^*(t)dt,$$

$$\begin{aligned} &\alpha_*(t) \leq -h_i' F(t)x - \sum_{j=1}^q \int_0^t h_i' F(t - s)d\omega_j(s)dsw_j \leq \alpha^*(t), \quad t \in T_\tau, \\ &Dx = \delta, \quad d_* \leq x \leq d^*, \quad f_* \leq w \leq f^*, \end{aligned} \tag{12.4}$$

where $\alpha_*(t) = \xi_* - y^*(t) + \int_0^t h_i' F(t - s)bu^*(s)ds$ and $\alpha^*(t) = \xi^* - y^*(t) + \int_0^t h_i' F(t - s)bu^*(s)ds$, $t \in T_\tau$.

Problems (7.2), (12.4) are solved in real time according to the scheme of section 11.

In the case of constructing a discrete realization $u^*(k\nu)$, $k=\overline{0,N-1}$, of optimal feedback the controller in real time solves the determined LP problem (11.4) in which the values $g_i(k\nu)=g_i-\alpha_i(k\nu)$, $i=\overline{1,m}$, are calculated as a result of solving m discrete OOI problems in real time

$$\begin{gathered}\alpha_i(k\nu)=\min\left[h_i'F(t^*)x+\sum_{j=1}^{q}\sum_{r=0}^{k-1}\int_{r\nu}^{(r+1)\nu}h_i'F(t^*-s)d\omega_j(s)dsw_j\right]\\+\sum_{j=0}^{k-1}\int_{j\nu}^{(j+1)\nu}h_i'F(t^*-s)bdsu^*(j\nu),\\\alpha_*(j\nu)\leq -h_i'F(j\nu)x-\sum_{l=1}^{q}\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)ddds\mu_j\leq\alpha^*(j\nu),\quad j=\overline{0,k},\\Dx=\delta,\quad d_*\leq x\leq d^*,\quad w_*\leq\mu_j\leq w^*,\quad j=\overline{0,N-1},\end{gathered}\tag{12.5}$$

where $\alpha_*(j\nu)=\xi_*-y^*(j\nu)+\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)bdsu^*(r\nu)$ and $\alpha^*(j\nu)=\xi^*-y^*(j\nu)+\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)bdsu^*(r\nu)$.

The scheme of solving problems (7.2), (12.5) is analogous to the one given in section 11.

12.3. Optimization of control systems under uncertainty with partial identification of parametric disturbances. In the class of piecewise continuous controls $u(t)$, $t\in T$, consider the problem

$$\begin{gathered}J(u)=h_0'x(t^*)\longrightarrow\max,\quad\dot{x}=Ax+bu+d_1\mu+d_2v,\\x(0)=x_0\in X_0,\quad x(t^*)\in X^*,\quad|u(t)|\leq 1,\quad t\in T,\end{gathered}\tag{12.6}$$

where $\mu(t)=\mu(t,w)=w_1\omega_1(t)+\cdots+w_q\omega_q(t)$, $t\in T$, $w=(w_1,\ldots,w_q)\in\check{W}$; $\omega_j(t)$, $j=\overline{1,q}$, $t\in T$, are known piecewise continuous functions and $v(t)$, $t\in T$, is an unknown piecewise continuous function of disturbances satisfying the inequality $v_*\leq v(t)\leq v^*$, $t\in T$. The equation of the measurer has form (8.3).

Vectors $x^*\in X_0$, $w^*\in\check{W}$ are said to be compatible with a written signal $y_\tau^*(\cdot)$ and the control $u_\tau^*(\cdot)$ if the trajectory $x^*(t)=x^*(t|x^*,w^*)$, $t\in T_\tau$, corresponding to the initial state x^*, the disturbance $\mu(t,w^*)$, $t\in T_\tau$, and the control $u_\tau^*(\cdot)$ together with some admissible functions $v^*(t)$, $\xi^*(t)$, $t\in T_\tau$, generates the signal $y^*(t)$, $t\in T_\tau$. The set of all pairs (x^*,w^*) of vectors $x^*\in X_0$, $w^*\in\check{W}$ compatible with $u_\tau^*(\cdot)$, $y_\tau^*(\cdot)$ is denoted by $\hat{X}_0(\tau)\times\hat{W}(\tau)$ and is called a τ-a-posteriori distribution of an initial state and parameters of a disturbance. The set

$$\begin{gathered}\hat{X}_\tau(\tau)=\left\{x:x=F(\tau)x_0+\int_0^\tau F(\tau-t)bu^*(t)dt+\int_0^\tau F(\tau-t)d_1\mu(t,w^*)dt\right.\\\left.+\int_0^\tau F(\tau-t)d_2v(t)dt,\quad x_0\in\hat{X}_0(\tau),\quad w^*\in\hat{W}(\tau),\quad v_*\leq v(t)\leq v^*\right\}\end{gathered}$$

is called a τ-a-posteriori distribution of a state $x(\tau)$. The optimal output feedback is defined by equality (12.3).

As in the case of optimal feedback without identification of disturbances, move on to a discrete realization $u^*(k\nu)=u^0(k\nu,y_{k\nu}^*(\cdot))$, $k=\overline{0,N-1}$.

Values $u^*(k\nu)$, $k=\overline{0,N-1}$ are calculated by the optimal controller in real time as a result of solving the LP problem (11.4). In order to calculate the values $g_i(k\nu)=g_i-\alpha_i(k\nu)$, $i=\overline{1,m}$, in real time m, discrete OOI problems

$$\alpha_i(k\nu)=\min\left[h_i'F(t^*)x+\sum_{j=1}^{q}\sum_{r=0}^{k-1}\int_{r\nu}^{(r+1)\nu}h_i'F(t^*-s)d_1\omega_j(s)dsw_j\right.$$

$$\left.+\sum_{r=0}^{k-1}\int_{r\nu}^{(r+1)\nu}h_i'F(t^*-s)d_2dsv_r\right]+\sum_{r=0}^{k-1}\int_{r\nu}^{(r+1)\nu}h_i'F(t^*-s)bdsu^*(r\nu),$$

$$(12.7)\qquad \alpha_*(j\nu)\le -h_i'F(j\nu)x-\sum_{l=1}^{q}\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)d_1\omega_l(s)dsw_l$$

$$-\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)d_2dsv_r\le\alpha^*(j\nu),\quad j=\overline{0,k},$$

$$Dx=\delta,\quad d_*\le x\le d^*,\quad w_*\le w_j\le w^*,\quad v_*\le v_j\le v^*,\quad j=\overline{0,N-1},$$

are solved where $\alpha_*(j\nu)=\xi_*-y^*(j\nu)+\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)bdsu^*(r\nu)$, and $\alpha^*(j\nu)=\xi^*-y^*(j\nu)+\sum_{r=0}^{j-1}\int_{r\nu}^{(r+1)\nu}h_i'F(j\nu-s)bdsu^*(r\nu)$.

Problem (12.7) contains $n+2N$ variables x_j, $j=\overline{1,n}$; w_j, v_j, $j=\overline{0,N-1}$.

Problems (7.2), (12.7) are solved in real time by the scheme of section 11.

12.4. Numerical modelling. The operation of the proposed optimal feedbacks is illustrated by the problem of minimization of fuel consumption in steering a two-mass mechanical system (Figure 3.1) in the vicinity of the equilibrium state. The mathematical model of the problem has the form

$$\int_0^{20}u(t)dt\longrightarrow\min,$$

$$\dot{x}_1=x_3,\quad \dot{x}_2=x_4,\quad \dot{x}_3=-x_1+x_2,\quad \dot{x}_4=0.1x_1-1.02x_2+w,$$

$$x(0)\in X_0,\quad x(20)\le 0.$$

The set X_0 is described by the relations

$$x_1=0,\quad x_2=0,\quad 0.8\le x_3\le 1.1,\quad 0.4\le x_4\le 0.6.$$

Three types of feedback are studied: (1) classical state feedback (section 2), (2) guaranteeing state feedback (section 6), and (3) guaranteeing output feedback.

In all cases the initial state is the vector $x(0)=(0,0,1,0.5)$.

The following information on disturbances is available: (a) a controller knows that a disturbance does not affect the system (classical determined case); (b) a disturbance affecting the system is known; (c) it is known that the system is under disturbances but there is no information on disturbances; (d) the system is under a bounded disturbance of general form (a piecewise continuous disturbance with a known set of values $|w(t)|\le 0.3$, $t\in T$); (e) a disturbance is a sum of a bounded disturbance of general form and a disturbance is known with a precision of parameters (section 12.3); (f) a disturbance is parametric (section 12.2).

When constructing the optimal output feedback, the measurer $y=x_1+\xi$, $|\xi(t)|\le 0.2$ is used.

Table 12.1
Results of numerical modelling.

		Type of feedback			
		State 1		Output 2	
N	Information on disturbances	J	Terminal state	J	Terminal state
		a) Classical feedback			
1	Disturbances are absent	2.6014	0 0 0 0	2.6082	0.00377 0.00069 0.00269 0.00050
2	Controller knows disturbances	2.9232	0 0 0 0	3.1324	0.00377 0.00070 0.00270 0.00050
3	There is not any information on disturbances	8.4457	3.9889 0.02551 2.17137 0.12837	9.36812	2.82342 0.31323 2.43286 0.39572
		b) Unclosable feedback			
				Disturbance is not identified	
4	Disturbances of general form with known set of values	8.8744	3.98291 0.06223 2.17727 0.27214	13.4163	2.27723 1.25260 4.85734 1.01632
				Parameters of disturbance are identified	
5	Disturbance: Sum of bounded disturbance of general form with known set of values and parametric disturbance	3.8132	0.24984 0.10335 0.14617 0.13778	9.7320	0.69001 1.09105 1.38797 1.05552
6	Parametric disturbance	3.7163	0.15076 0.09622 0.15070 0.11869	3.1619	0.00659 0.01533 0.05437 0.10126

In each case a discrete realization of the optimal feedback is constructed. The parameter ν is put to be equal to 0.5, and the realized error function has the form $\xi^*(t) = 0.2\sin 3t$, $t \in T$. The system is under the disturbance of the form $w(t) = w_1\omega_1(t) + w_2\omega_2(t) + w_3\omega_3(t)$, $t \in T$: $w^*(t) = 0.15\sin 0.5t + 0.1\sin 2t + 0.05\sin 5t$, $t \in T$. OO problems are solved with the same data as in section 10.5.

Results of numerical experiments are listed in Table 12.1. In the table, for each situation under consideration, fuel consumption and coordinates of the terminal state are presented. The ideal situation (determined system, optimal state feedback) is identified by a box (1, 1) placed in the first row and in the first column. It is seen from the table that results are impaired as uncertainty increases: fuel consumption grows, and the terminal state moves away from the equilibrium state; additional efforts on identification of parametric disturbances improve efficiency of control.

In Figure 12.1a the phase trajectories of the optimized control system closed by the optimal output feedback are presented. The dashed line corresponds to the case when a disturbance does not affect the system. The solid line is constructed

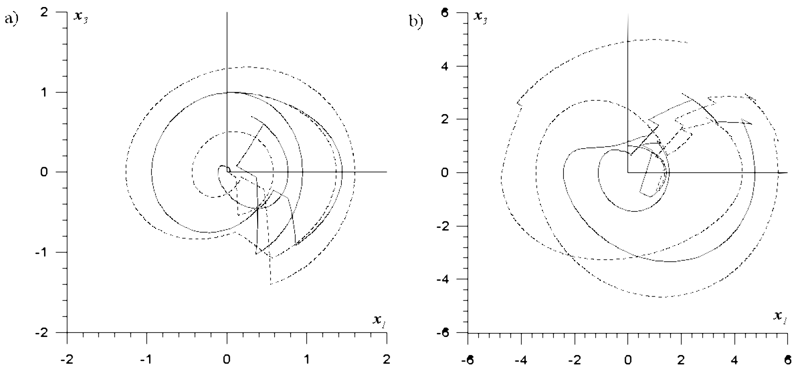


FIG. 12.1.

for the case of identification of parametric disturbances. In Figure 12.1b the dashed line presents the phase trajectory constructed without identification of disturbances, and the solid line is constructed for the case of partial identification of parametric disturbances.

Acknowledgments. The authors thank Professors R. Bulirsch, M. Hautus, K. Malanowski, K. Malinowski, B. Mordukhovich, and M. Thoma for favorable discussions of results on the first, especially difficult, stages of our investigations in the optimal synthesis problem. We are also very grateful to B. Shulkin for his help in preparing the English version of the paper.

REFERENCES

[1] M. Athans and P.L. Falb, *Optimal Control*, McGraw-Hill, New York, 1965.

[2] N.V. Balashevich, R. Gabasov, and F.M. Kirillova, *Suboptimal controller smoothing controls and filtering pertubations on sliding intervals*, Izv. Akad. Nauk Tekhn. Kibernetika, 6 (1993), pp. 25–32 (in Russian).

[3] N.V. Balashevich, R. Gabasov, and F.M. Kirillova, *A controller for the sliding optimization of dynamic control systems*, Comput. Math. Math. Phys., 34 (1994), pp. 529–533.

[4] N.V. Balashevich, R. Gabasov, and F.M. Kirillova, *Optimal positional control of a group of systems*, Automat. Remote Control, 55 (1994), pp. 164–171.

[5] N.V. Balashevich, R. Gabasov, and F.M. Kirillova, *Optimal positional mobile control of dynamic plants*, J. Comput. System Sci., 37 (1998), pp. 383–390.

[6] R.E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1963.

[7] R.E. Bellman, I. Glicksberg, and O.A. Gross, *Some Aspects of the Mathematical Theory of Control Processes*, Report R-313, Rand Corporation, Santa Monica, CA, 1958.

[8] A.E. Bryson and Yu-Chi Ho, *Applied Optimal Control*, Blaisdell, Toronto, Canada, 1969.

[9] D.W. Bushaw, *Experimental Towing Tank*, Report 469, Stevens Institute of Technology, Hoboken, NJ, 1953.

[10] G.B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.

[11] S.E. Dreyfus, *Some types of optimal control of stochastic systems*, J. SIAM Ser. A Control, 2 (1964), pp. 120–134.

[12] A.A. Feldbaum, *Fundamentals of the Theory of Optimal Automatic Systems*, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1963 (in Russian).

[13] R. Gabasov and F.M. Kirillova, *New algorithms and results of numerical experiments for solution of mathematical programming and optimal control problems*, in Selected Topics in Operation Research and Mathematical Economics, Lecture Notes in Econom. and Math. Systems 226, G. Hammer and D. Pallaschke, eds., Springer-Verlag, Berlin, 1984, pp. 440–456.

[14] R. Gabasov, N.V. Balashevich, and F.M. Kirillova, *Synthesis of adaptive optimal controls for linear dynamic systems*, in Computational Optimal Control (Munich, 1992), Internat. Ser. Numer. Math. 115, Birkhäuser, Basel, 1994, pp. 83–88.

[15] R. Gabasov, P.V. Gaishun, and F.M. Kirillova, *Optimal estimator for linear systems*, Informatica (Vilnius),2 (1991), pp. 195–220.

[16] R. Gabasov, P.V. Gaishun, F.M. Kirillova, and S.V. Prischepova, *Optimal feedback for discrete systems with perturbations.* I. *Design of optimal estimator*, Automat. Remote Control, 53 (1992), pp. 44–51.

[17] R. Gabasov, P.V. Gaishun, F.M. Kirillova, and S.V. Prischepova, *Optimal feedback for discrete systems with perturbations.* II. *Optimal control synthesis*, Automat. Remote control, 53 (1992), pp. 500–505.

[18] R. Gabasov and F.M. Kirillova, *Consideration of optimal control problems specificity on generalizing mathematical programming algorithms*, in the Preprints of the 9th IFAC World Congress on Automatic Control, Budapest, Hungary, 5 (1984), pp. 264–269.

[19] R. Gabasov and F.M. Kirillova, *Synthesis of optimal control in flexible feedback form*, in Abstracts of Xth Union (USSR) Conference on Control Problems, Alma-Ata, 1986, pp. 137–138 (in Russian).

[20] R. Gabasov and F.M. Kirillova, *Adjoint problems of control, observation and identification*, Dokl. Akad. Nauk Belarusi, 34 (1990), pp. 777–780 (in Russian).

[21] R. Gabasov and F.M. Kirillova, *Finite algorithm for construction of programs for incompletely determined linear control problems*, Automat. Remote Control, 52 (1991), pp. 912–918.

[22] R. Gabasov and F.M. Kirillova, *Optimization of dynamical systems with identification of input perturbations*, Problems Control Inform. Theory, 20 (1991), pp. 233–246.

[23] R. Gabasov and F.M. Kirillova, *Dual optimization of dynamic systems*, in Simulation and Optimization, Lecture Notes in Econom. and Math. Systems 374, Q. Pflug and U. Dieter, eds., Springer-Verlag, New York, 1992, pp. 109–118.

[24] R. Gabasov and F.M. Kirillova, *Optimal output feedback*, Vestn. Beloruss. Gos. Univ. Ser. 1 Fiz. Mat. Inform., 2 (1994), pp. 64–68 (in Russian).

[25] R. Gabasov and F.M. Kirillova, *Real time construction of optimal closable feedbacks*, in the Proceedings of IFAC 13th Triennial World Congress, San Francisco, CA, 1996, pp. 231–236.

[26] R. Gabasov and F.M. Kirillova, *An optimal feedback on the basis of a mathematical model with test perturbations*, Dokl. Math., 54 (1996), pp. 637–639.

[27] R. Gabasov, F.M. Kirillova, and N.V. Balashevich, *Program-positional optimization for dynamic systems*, in Optimal Control. Calculus of Variations, Optimal Control Theory and Numerical Methods, Internat. Ser. Numer. Math. 111, Birkhäuser, Basel, 1993, pp. 195–205.

[28] R. Gabasov, F.M. Kirillova, P.V. Gaishun, and S.V. Prischepova, *Synthesis of optimal controls on nonexact measurements of output signals*, Problems Control Inform. Theory, 20 (1992), pp. 406–427.

[29] R. Gabasov, F.M. Kirillova, and E.A. Kostina, *Closable feedbacks for guaranteed optimization of undeterminate control systems*, Dokl. Math., 54 (1996), pp. 637–639.

[30] R. Gabasov, F.M. Kirillova, and E.A. Kostina, *Closed state feedback for optimization of uncertain control systems. Part* I: *Multiple contact*, Automat. Remote Control, 57 (1996), pp. 1008–1015.

[31] R. Gabasov, F.M. Kirillova, and E.A. Kostina, *Closed state feedback for optimization of uncertain control systems. Part* II: *Multi-closable feedback*, Automat. Remote Control, 57 (1996), pp. 1137–1145.

[32] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova, *Construction of optimal controls of feedback type in a linear problem*, Soviet Math. Dokl., 44 (1992), pp. 608–613.

[33] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova, *Optimization of linear control system in real-time mode*, Izv. Akad. Nauk. Tekhn. Kibernetika, 4 (1992), pp. 3–19 (in Russian).

[34] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova, *Construction of optimal estimators for linear dynamic systems*, Vestn. Beloruss. Gos. Univ. Ser. 1 Fiz. Mat. Inform., 3 (1992), pp. 45–49 (in Russian).

[35] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova, *A method of the optimal control of the motion of a dynamic system in the presence of constantly operating perturbations*, Appl. Math. Mech. (English Ed.), 56 (1992), pp. 755–764.

[36] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova, *Algorithm for real-time optimization of incompletely determined linear control systems.* I. *Indefinite initial state*, Automat. Remote Control, 54 (1993), pp. 555–563.

[37] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova, *Algorithm for real-time optimization*

of incompletely determined linear control systems. II. *Constantly acting perturbations*, Automat. Remote Control, 54 (1993), pp. 715–722.

[38] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova, *Optimally fast position control of linear nonsteady systems*, Physics–Doklady, 39 (1994), pp. 680–682.

[39] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova, *Optimal positional observation for linear systems*, Physics–Doklady, 39 (1994), pp. 846–849.

[40] R. Gabasov, F.M. Kirillova, O.I. Kostyukova, and N.V. Balashevich, *An algorithm of the solution of optimal control problems in real time for linear systems with many inputs*, Systems Sci., 20 (1994), pp. 25–38.

[41] R. Gabasov, F.M. Kirillova, O.I. Kostyukova, and A.V. Pokatayev, *Optimal program controls and flexible feedback*, in Preprints of 10th IFAC Congress on Automatic Control, Munich, Germany, 1987, pp. 119–124.

[42] R. Gabasov, F.M. Kirillova, and S.V. Prischepova, *Optimal identificator for linear systems*, Informatica (Vilnius), 3 (1991), pp. 367–377.

[43] R. Gabasov, F.M. Kirillova, and S.V. Prischepova, *Synthesis of time optimal discrete system*, Automat. Remote Control, 52 (1991), pp. 1716–1722.

[44] R. Gabasov, F.M. Kirillova, and S.V. Prischepova, *Optimal Feedback Control*, Lecture Notes in Control and Inform. Sci. 207, M. Thoma, ed., Springer-Verlag, London, 1995.

[45] R. Gabasov, F.M. Kirillova, and S.V. Prischepova, *Optimal controller for discrete system with indeterminate perturbations*, Control Cybernet., 26 (1997), pp. 227–239.

[46] R. Kalman, *Control systems general theory*, in the Proceedings of the 1st IFAC Congress, Moscow, USSR, 1961, pp. 521–547 (in Russian).

[47] K. Kortanek and R. Hettich, *Semi-infinite programming: Theory, methods and applications*, SIAM Rev., 35 (1993), pp. 380–429.

[48] O.I. Kostyukova, *Superbasis feasible points of linear extremal problem with continuum of constraints*, Dokl. Akad. Nauk Belarusi, 33 (1989), pp. 687–689 (in Russian).

[49] A.B. Kurzhanski, *Identification—a theory of guaranteed estimates*, in From Data to Model, J.C. Willems, ed., Springer-Verlag, Berlin, 1989, pp. 135–214.

[50] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley and Sons, New York, 1972.

[51] E.B. Lee and L. Markus, *Foundations of Optimal Control Theory*, Wiley and Sons, New York, 1967.

[52] A.M. Letov, *Analytic construction of regulators. Parts* I–III, Avtomat. i Telemekh., 21 (1960), pp. 436–441, 561–568, 661–665 (in Russian).

[53] L.S. Pontryagin, V.G. Boltyanski, R.V. Gamkrelidze, and E.F. Mischenko, *The Mathematical Theory of Optimal Processes*, Interscience Publishers, New York, 1962.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1043–1064

EXISTENCE OF LIPSCHITZ AND SEMICONCAVE CONTROL-LYAPUNOV FUNCTIONS*

LUDOVIC RIFFORD†

Abstract. Given a locally Lipschitz control system which is globally asymptotically controllable at the origin, we construct a control-Lyapunov function for the system which is Lipschitz on bounded sets, and we deduce the existence of another one which is semiconcave (and so locally Lipschitz) outside the origin. The proof relies on value functions and nonsmooth calculus.

Key words. asymptotic stabilizability, semiconcave Lyapunov function, nonsmooth analysis, viscosity solutions

AMS subject classifications. 93D05, 93D20, 93B05, 34D20, 49J52, 49L25, 70K15

PII. S0363012999356039

1. Introduction.

This paper is concerned with the stabilization problem for a standard control system of the form $\dot{x}(t) = f(x(t), u(t))$. Lyapunov-like techniques have been successfully used in many problems in control theory, such as stabilizability, asymptotic controllability, and stability. Stabilization by smooth feedback has been a subject of research by many authors. Among them, Artstein provided an important contribution (see [3]), proving that a control system admits a smooth Lyapunov function if and only if there is a stabilizing relaxed feedback. Moreover, if the system is affine in the control, there exists an ordinary stabilizing feedback continuous outside the origin. In general, however, such a feedback fails to exist, as pointed out by Sontag and Sussmann [27] and by Brockett [8] among others [25],[12]. Consequently, a smooth Lyapunov function in general does not exist. This fact leads to the design of time-varying (see [15],[16]) or discontinuous feedbacks. The construction of the latter (see [11]) has used the existence of a continuous control-Lyapunov function (CLF) whose decrease condition is stated in terms of Dini derivates or, equivalently, of proximal subgradients. Moreover, more recently, Clarke, Ledyaev, Stern, and Rifford [10] invoked the existence of such a CLF which is locally Lipschitz in order to construct a discontinuous feedback law which stabilizes the underlying system to any given tolerance and which possesses a robustness property relative to measurement error. The first result of this article is that, under certain mild assumptions on f (a local Lipschitz condition and bounded dynamics near the origin), for globally asymptotically controllable (GAC) systems, such locally Lipschitz CLF always exist. This fact extends the well-known result of Sontag [26] and brings an affirmative answer to a conjecture that has been attributed to Sontag and Sussmann. Furthermore, the main result shows that a semiconcave CLF outside the origin always exists under the same assumptions. The semiconcavity is an intermediate property between Lipschitz continuity and continuous differentiability. Semiconcave functions have been used, for instance, to obtain uniqueness results for weak solutions of Hamilton–Jacobi equations (see [19],[20]). More recently, attention has been focused on the differential properties of such functions (see [1], [2]). In the case of the stabilization problem, the semicon-

*Received by the editors May 10, 1999; accepted for publication (in revised form) July 4, 2000; published electronically November 8, 2000.

http://www.siam.org/journals/sicon/39-4/35603.html

†Institut Girard Desargues, Université Claude Bernard Lyon I 69622 Villeurbanne, France (rifford@desargues.univ-lyon1.fr).

cavity will be exploited to derive an efficient construction for stabilizing discontinuous feedbacks by means of Euler trajectories. Furthermore, the semiconcave regularity of the CLF will be used to obtain some regularity of our discontinuous feedback outside a set of singularities which will be proven small on account of semiconcavity. These results will appear in forthcoming articles [22], [23], [21]. Some works and general references related to this article include [5], [6], [17], [18], [24].

2. Definitions and statements of the results. In this paper, we study systems of the general form

$$\dot{x}(t) = f(x(t), u(t)), \tag{2.1}$$

where the state $x(t)$ takes values in a Euclidean space $\mathbb{X} = \mathbb{R}^n$, the control $u(t)$ takes values in a given set U, and f satisfies the following hypotheses.

Assumption 2.1. *f is locally Lipschitz in x(uniformly in u). That is, for all $x \in \mathbb{X}$, there exists $\mathcal{V}_x$, a neighborhood of x, and $L_x \geq 0$ such that*

$$\|f(y', u) - f(y, u)\| \leq L_x \|y' - y\| \quad \forall y, y' \in \mathcal{V}_x, \quad \forall u \in U.$$

Assumption 2.2. *f is bounded on the ball $R\bar{B} \times U$ for all $R > 0$ (or, equivalently, in view of the preceding assumption for some $R > 0$).*

A special element "0" is distinguished in U, and the state $x = 0$ of $\mathbb{X}$ is an equilibrium point, i.e., $f(0, 0) = 0$. (No linear structure on U is used, however.) The set of admissible controls is the set of measurable and locally essentially bounded functions $u : \mathbb{R}_{\geq 0} \longrightarrow U$. $\mathbb{R}_{\geq 0}$ denotes nonnegative reals, B denotes the open ball $B(0, 1) := \{x : \|x\| < 1\}$ in $\mathbb{X}$, and $\bar{B}$ denotes the closure of B.

We now introduce our definitions and the main result.

Definition 2.3. *The system* (2.1) *is GAC if there exist a nondecreasing function $M : \mathbb{R}_{>0} \longrightarrow \mathbb{R}_{>0}$ such that $\lim_{R \downarrow 0} M(R) = 0$ and a function $T : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \longrightarrow \mathbb{R}_{\geq 0}$ with the following property.*

For any $0 < r < R$, for each initial state ξ, $\|\xi\| \leq R$, there exist a control $u : \mathbb{R}_{\geq 0} \longrightarrow U$ and corresponding trajectory $x(\cdot) : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{X}$ such that

1. $\lim_{t \to \infty} x(t) = 0$,
2. *for all $t \geq 0, \|x(t)\| \leq M(R)$,*
3. *for all $t \geq T(r, R), \|x(t)\| \leq r$.*

Remark 2.4. *A routine argument involving continuity of trajectories with respect to initial states shows that the requirements of the above standard definition are equivalent to the following apparently weaker pair of conditions used in some references (see* [28], [29]*).*

1. *For each $\xi \in \mathbb{X}$ there is a control $u : \mathbb{R}_{\geq 0} \longrightarrow U$ that drives ξ asymptotically to* 0.
2. *For each $\epsilon > 0$, there is a $\delta > 0$ such that for each $\xi \in \mathbb{X}$ with $\|\xi\| \leq \delta$ there is a control $u : \mathbb{R}_{\geq 0} \longrightarrow U$ that drives ξ asymptotically to* 0 *and such that the corresponding trajectory $x(\cdot)$ satisfies $\|x(t)\| \leq \epsilon$ for all $t \geq 0$.*

Moreover, the authors of [28], [29] *add a condition on bounded controls; this one implies Assumption* 2.2 *by restriction to the system near the origin.*

A function $V : \mathbb{X} \longrightarrow \mathbb{R}_{\geq 0}$ is *positive definite* if $V(0) = 0$ and $V(x) > 0$ for $x \neq 0$, and *proper* if $V(x) \to +\infty$ as $\|x\| \to +\infty$.

Definition 2.5. *A Lyapunov pair for the system* (2.1) *is a pair (V, W) consisting of a continuous, positive definite, proper function $V : \mathbb{X} \longrightarrow \mathbb{R}$ and a positive definite* *continuous function $W : \mathbb{X} \longrightarrow \mathbb{R}$, with the property that for each $x \in \mathbb{X} \setminus \{0\}$ we have*

$$\forall \zeta \in \partial_P V(x), \inf_{u \in U} \langle \zeta, f(x,u) \rangle \leq -W(x). \tag{2.2}$$

Here $\partial_P V(x)$ refers to the *proximal subdifferential* of V at x (which may be empty): ζ belongs to $\partial_P V(x)$ if and only if there exists σ and $\eta > 0$ such that

$$V(y) - V(x) + \sigma \|y - x\|^2 \geq \langle \zeta, y - x \rangle \quad \forall y \in x + \eta B.$$

The condition (2.2) is, in fact, equivalent to another one often used in the definition of a nonsmooth Lyapunov function (see [26], [28], [29]); this other notion is based on the notion of directional or Dini subderivate. The equivalence between these two conditions is a consequence of Subbotin's theorem (see, for example, [14], our principal source for the theory of nonsmooth analysis, and [10] for a discussion of the equivalence). We remark that there exists a complete calculus of proximal subdifferentials, one that extends all the theorems of the usual smooth calculus.

DEFINITION 2.6. *A CLF for the system* (2.1) *is a function $V : \mathbb{X} \longrightarrow \mathbb{R}$ such that there exists a continuous positive definite $W : \mathbb{X} \longrightarrow \mathbb{R}$ with the property that (V, W) is a Lyapunov pair for* (2.1)*.*

We will say that V is a locally Lipschitz CLF if V is a CLF which is locally Lipschitz on $\mathbb{X}$. We claim the following theorem.

THEOREM 1. *Let (f, U) be a control system as described above. Then under Assumptions* 2.1 *and* 2.2*, if the system is GAC, there exists a locally Lipschitz CLF.*

REMARK 2.7. *The converse is true and relatively easy if we suppose that f is continuous in u (we need this to obtain the existence of trajectories); we refer to Sontag* [26].

We now recall the definition of a *semiconcave* function [19] in an open set Ω of $\mathbb{X}$.

DEFINITION 2.8. *Let $g : \Omega \longrightarrow \mathbb{R}$ be a continuous function on Ω; it is said to be semiconcave on Ω if for any point $x_0 \in \Omega$ there exist $\rho, C > 0$ such that*

$$g(x) + g(y) - 2g\left(\frac{x+y}{2}\right) \leq C\|x - y\|^2 \tag{2.3}$$

for all $x, y \in x_0 + \rho B$.

We shall deduce as a corollary of the preceding theorem the main result of this article.

THEOREM 2. *Let (f, U) be a control system as described above. Then under Assumptions* 2.1 *and* 2.2*, if the system is GAC, there exists a CLF which is semiconcave on $\mathbb{X} \setminus \{0\}$.*

REMARK 2.9. *The CLF is a viscosity supersolution of*

$$\sup_{u \in U}\{-\langle f(x,u), DV \rangle\} - W \geq 0.$$

We begin by giving some regularity results about certain value functions. Then we give the proof of Theorem 1. In the last section, we conclude with the proof of Theorem 2.

3. A result on value functions in finite time.

Throughout this section, we are given a multifunction F mapping $\mathbb{X}$ to the subsets of $\mathbb{X}$, and we consider the differential inclusion

$$\dot{x}(t) \in F(x(t)) \quad \text{almost everywhere (a.e.).} \tag{3.1}$$

We say that $x(\cdot)$ is a solution of (3.1) on the interval $[a, b]$ if it is an absolutely continuous function $x : [a, b] \longrightarrow \mathbb{X}$ which, together with $\dot{x}$, satisfies (3.1); such an arc will be called an F-trajectory on the interval $[a, b]$. We need for this section two properties of F which turn out to be particularly important.

ASSUMPTION 3.1. *The multifunction F is locally Lipschitz with nonempty compact convex values.*

ASSUMPTION 3.2. *For some positive constants K and M, and for all $x \in \mathbb{X}$,*

$$v \in F(x) \Longrightarrow \|v\| \leq K\|x\| + M$$

(that is called the linear growth condition).

Under these two conditions, for all $x_0 \in \mathbb{X}$, there exists a trajectory of (3.1) defined on $\mathbb{R}_{\geq 0}$ such that $x(0) = x_0$, and for any trajectory with initial data x_0 we have the following estimate:

$$\forall t \geq 0, \|x(t)\| \leq \|x_0\|e^{Kt} + Mte^{Kt}. \tag{3.2}$$

This inequality is an easy consequence of Gronwall's lemma (see [14]).

Let us consider a function $L : \mathbb{X} \longrightarrow \mathbb{R}_{\geq 0}$ and a compact set $\mathcal{T}$ of $\mathbb{X}$ satisfying the following assumption.

ASSUMPTION 3.3. *L is locally Lipschitz and for all $x \in \mathbb{X}$, $L(x) \geq 1$.*

ASSUMPTION 3.4. *There exists $\delta > 0$ such that for all $x \in \mathcal{T}, \delta\bar{B} \subset F(x)$.*

We proceed now to define a value function $V(\cdot)$ on $\mathbb{X}$ in terms of trajectories of F as follows:

$$V(x) := \inf\left\{\int_0^T L(x(t))dt : x(0) = x, \dot{x}(t) \in F(x(t)) \text{ a.e. and } x(T) \in \mathcal{T}\right\}.$$

(Note that T is a choice variable in this "free-time" problem.)

We introduce the notation

$$\mathcal{R} := \{x \in \mathbb{X} : V(x) < +\infty\}.$$

The letter $\mathcal{R}$ stands for reachable; the set of points where V is finite is the set of points which can be driven to the target $\mathcal{T}$ in finite time. We have the following theorem.

THEOREM 3. *Assume* (3.1)–(3.4). *Then*

(i) *$\mathcal{R}$ is open,*

(ii) *V is locally Lipschitz in $\mathcal{R}$,*

(iii) *for all $x \in \mathcal{R} \setminus \mathcal{T}$, for all $\zeta \in \partial_P V(x), \min_{v\in F(x)}\langle\zeta, v\rangle \leq -L(x)$.*

Proof. First, by the Lipschitz condition on F and (3.4), there exists $0 < r \leq 1$ such that for all $x \in \mathcal{T} + r\bar{B}$, $\frac{\delta}{2}\bar{B} \subset F(x)$. Hence, each state x of $\mathcal{T} + r\bar{B}$ can be driven to $\mathcal{T}$ by a trajectory of (3.1) in time $\frac{2}{\delta}d(x, \mathcal{T})$ (where $d(x, \mathcal{T})$ denotes $\min_{\tau\in\mathcal{T}} \|x - \tau\|$). This proves that V is finite on $\mathcal{T} + r\bar{B}$. If we set $m := \max_{x\in\mathcal{T}+r\bar{B}} L(x)$, we have

$$\forall x \in \mathcal{T} + r\bar{B}, V(x) \leq \frac{2m}{\delta}d(x, \mathcal{T}). \tag{3.3}$$

Now fix $x_0 \notin \mathcal{T}$ such that $V(x_0) < +\infty$.

By the definition of V, there exists an F-trajectory $x_0(\cdot)$ and $T > 0$ such that $x_0(0) = x_0, x_0(T) \in \mathcal{T}$ and

$$\int_0^T L(x(s))ds \leq V(x_0) + 1. \tag{3.4}$$

The estimate (3.2) gives

$$\forall t \in [0,T], \|x_0(t)\| \leq \|x_0\| e^{KT} + MTe^{KT}.$$

Let $A := [\|x_0\| + MT]e^{KT}$, and let λ_F be the Lipschitz constant of F on the ball $(A+1)\bar{B}$. Fix $y \in B(x_0, re^{-\lambda_F T})$.

By [4, Cor. 1, p. 121], there exists an F-trajectory $y(\cdot)$ such that $y(0) = y$, verifying

$$\forall t \in [0,T], \|y(t) - x_0(t)\| \leq e^{\lambda_F T}\|y - x_0\|, \text{and } \|y(t)\| \leq A+1. \tag{3.5}$$

Consequently, if we set λ_L as the Lipschitz constant of L on the ball $(A+1)\bar{B}$, we obtain

$$\begin{aligned}\int_0^T L(y(s))ds &\leq \int_0^T L(x_0(s))ds + \int_0^T [L(y(s)) - L(x_0(s))]ds \\ &\leq V(x_0) + 1 + \int_0^T \lambda_L \|y(s) - x_0(s)\| ds \\ &\leq V(x_0) + 1 + T\lambda_L e^{\lambda_F T}\|y - x_0\| \\ &\leq V(x_0) + 1 + T\lambda_L r.\end{aligned}$$

On the other hand, $d(y(T), \mathcal{T}) \leq \|y(T) - x_0(T)\| \leq e^{\lambda_F T}\|y - x_0\| \leq r$; this implies by (3.3) that

$$V(y(T)) \leq \frac{2m}{\delta} d(y(T), \mathcal{T}) \leq \frac{2mr}{\delta}.$$

Consequently, we have that for all $y \in B(x_0, re^{-\lambda_F T})$,

$$V(y) \leq \int_0^T L(y(s))ds + \frac{2mr}{\delta} \tag{3.6}$$

$$\leq V(x_0) + 1 + T\lambda_L r + \frac{2mr}{\delta} =: c < +\infty. \tag{3.7}$$

We have shown that $B(x_0, re^{-\lambda_F T}) \subset \mathcal{R}$, which gives (i).

Now, let $x \in B(x_0, re^{-\lambda_F T})$; then for each positive integer n, there exists an F-trajectory $x_n(\cdot)$ and $T_x^n \geq 0$ such that $x_n(0) = x$, $x_n(T_x^n) \in \mathcal{T}$, and

$$\int_0^{T_x^n} L(x_n(s))ds \leq V(x) + \frac{1}{n}.$$

Thus $L \geq 1$ implies $T_x^n \leq V(x) + \frac{1}{n} \leq c + \frac{1}{n}$ by (3.6). As before, the estimate (3.2) gives for each n

$$\begin{aligned}\forall t \in [0, T_x^n], \|x_n(t)\| &\leq \|x\| e^{KT_x^n} + MT_x e^{KT_x^n} \\ &\leq [\|x\| + M(c+1)]e^{K(c+1)} \\ &\leq [\|x_0\| + 1 + M(c+1)]e^{K(c+1)}.\end{aligned}$$

So we find a uniform bound for $\|\dot{x}_n(\cdot)\|$ on the intervals $[0, T_x^n] \subset [0, V(x)+1]$. Hence, since our trajectories $x_n(\cdot)$ are uniformly bounded and equicontinuous on the compact interval $[0, c+\frac{1}{n}]$, the theorem of Arzela–Ascoli and the compactness of trajectories (see

[14]) imply that there exists a trajectory $x(\cdot)$ with initial data x such that $x(T_x) \in \mathcal{T}$ and

$$V(x) = \int_0^{T_x} L(x(s))ds,$$

with $T \leq V(x)$. That means that the infimum is attained in the definition of V.

We set $A' := [\|x_0\| + 1 + M(c+1)]e^{K(c+1)}$, and λ'_F is the Lipschitz constant of F on the ball $(A'+1)\bar{B}$. We proceed to show that V is Lipschitz on the ball $\bar{B}(x_0, \frac{r}{2}e^{-\lambda'_F(c+1)})$.

Fix x, y in $\bar{B}(x_0, \frac{r}{2}e^{-\lambda'_F(c+1)})$. Then there exists, as above, $x(\cdot)$ as an F-trajectory and $T_x \geq 0$ such that $x(0) = x, x(T_x) \in \mathcal{T}$, and

$$V(x) = \int_0^{T_x} L(x(s))ds.$$

By [4, Cor. 1, p. 121], there exists an F-trajectory $y(\cdot)$ such that $y(0) = y$, verifying

$$\forall t \in [0, T_x], \|y(t) - x(t)\| \leq e^{\lambda'_F T_x}\|y - x\|, \text{and } \|y(t)\| \leq A' + 1.$$

Consequently, if we set, as before, λ'_L as the Lipschitz constant of L on the ball $(A'+1)\bar{B}$, we obtain

$$\begin{aligned}\int_0^{T_x} L(y(s))ds &\leq \int_0^{T_x} L(x(s))ds + \int_0^{T_x} \lambda'_L\|y(s) - x(s)\|ds \\ &\leq V(x) + T_x\lambda'_L e^{\lambda'_F T_x}\|y - x\| \\ &\leq V(x) + c\lambda'_L e^{\lambda'_F c}\|y - x\|.\end{aligned}$$

Now, $V(y(T_x)) \leq \frac{2m}{\delta} d(y(T_x), \mathcal{T}) \leq \frac{2m}{\delta} e^{\lambda'_F c}\|y - x\|$. Hence, we conclude that

$$V(y) \leq V(x) + \left[c\lambda'_L + \frac{2m}{\delta}\right] e^{\lambda'_F c}\|y - x\|.$$

Thus, since all the constants in the preceding inequality are independent of x and y, we find

$$|V(y) - V(x)| \leq \left[(c+1)\lambda'_L + \frac{2m}{\delta}\right] e^{\lambda'_F(c+1)}\|y - x\|,$$

which proves (ii). We now have to prove (iii). For that, consider $x \in \mathcal{R} \setminus \mathcal{T}$ and a trajectory $x(\cdot)$ of (3.1) attaining the infimum in the definition of $V(x)$. Let ζ belong to $\partial_P V(x)$; then there exists σ and $\eta > 0$ such that

$$V(y) - V(x) + \sigma\|y - x\|^2 \geq \langle \zeta, y - x\rangle \quad \forall y \in x + \eta B.$$

By the optimality of the trajectory $x(\cdot)$, for all $t \in [0, T], V(x(t)) = \int_t^T L(x(s))ds$. Then, for t sufficiently small,

$$\int_t^T L(x(s))ds - \int_0^T L(x(s))ds + \sigma\|x(t) - x\|^2 \geq \langle \zeta, x(t) - x\rangle,$$

which gives

$$-\frac{1}{t}\int_0^t L(x(s))ds + t\sigma\left\|\frac{x(t) - x}{t}\right\|^2 \geq \left\langle \zeta, \frac{x(t) - x}{t}\right\rangle.$$

We find (iii) by passing to the limit $t \downarrow 0$. □

REMARK 3.5. *In* [10], *a result of this type is proven differently by an argument based on Hamiltonian necessary conditions.*

REMARK 3.6. *The conclusions of Theorem* 3 *remain true if we weaken Assumption* 3.4 *to the proximal condition*

$$\min_{v \in F(x)} \langle \zeta, v \rangle \leq -\delta \|\zeta\|$$

for all $x \in \mathcal{T}$ and $\zeta \in N^P_{\mathcal{T}}(x)$, where $N^P_{\mathcal{T}}(x)$ denotes the proximal normal cone to $\mathcal{T}$ at x (see the book of Clarke, Ledyaev, Stern, and Wolenski [14]*). (This result is a consequence of proximal criteria for attainability; see* [13], [14]*.) This kind of condition added to the smooth regularity of F is used in* [9] *to obtain the semiconcavity of the minimum-time function. However, these results (on the Lipschitz property or on the semiconcavity property) do not hold if we omit the linear growth condition* (3.2)*; see, for example,* [7, Ex. 1.3, p. 238].

REMARK 3.7. *The conclusion* (iii) *can be strengthened to equality. The value function V is the viscosity solution of a certain Hamilton–Jacobi equation (see* [7], [14]*).*

4. Proof of Theorem 1. We suppose first that we have constructed a CLF V which is continuous on $\mathbb{X}$ and locally Lipschitz on $\mathbb{X} \setminus \{0\}$. Thus, there exists another continuous positive definite function $W : \mathbb{X} \longrightarrow \mathbb{R}_{\geq 0}$ such that (V, W) is a Lyapunov pair for (2.1). We proceed to show that we can deduce the existence of a new CLF which is locally Lipschitz on all the space $\mathbb{X}$. We set for any $0 \leq a \leq b$

$$S_V(b) := \{x; V(x) \leq b\} \text{ and } S_V[a, b] := \{x; a \leq V(x) \leq b\};$$

these are compact sets of $\mathbb{X}$. We proceed to construct a sequence of functions on $\mathbb{X}$ which will converge uniformly to our desired locally Lipschitz CLF.

First, we set $\mathcal{V}_0(x) := \max\{V(x), 1\}$. This function is locally Lipschitz on $\mathbb{X}$, proper, positive, constant on $S_V(1)$, and it verifies

$$\forall x \notin S_V(1), \forall \zeta \in \partial_P \mathcal{V}_0(x), \inf_{u \in \mathcal{U}} \langle \zeta, f(x, u) \rangle \leq -W(x).$$

By assumption, for all $n \geq 0$, V is Lipschitz on $S_V[\frac{1}{2^{n+1}}, \frac{1}{2^n}]$; we denote by $K(\frac{1}{2^{n+1}}, \frac{1}{2^n}) = K_n$ its Lipschitz constant on this set. (Without loss of generality we can choose this constant greater than 1.)

We define now a sequence inductively. Suppose $\mathcal{V}_n$ is given; we set

$$\mathcal{V}_{n+1}(x) := \begin{cases} \mathcal{V}_n(x) & \text{if } x \notin S_V(\frac{1}{2^n}), \\ \mathcal{V}_n(x) + \frac{1}{K_n}[V(x) - \frac{1}{2^n}] & \text{if } x \in S_V[\frac{1}{2^{n+1}}, \frac{1}{2^n}], \\ \mathcal{V}_n(x) - \frac{1}{2^{n+1} K_n} & \text{if } x \in S_V(\frac{1}{2^{n+1}}). \end{cases}$$

We have the following lemma.

LEMMA 4.1. *For all $n \geq 1$, $\mathcal{V}_n$ is 1-Lipschitz on $S_V(1)$, proper, and constant on $S_V(\frac{1}{2^n})$. Moreover, $\mathcal{V}_n$ satisfies the following properties:*

$$\forall x \in \mathbb{X}, \mathcal{V}_n(x) \geq 1 - \sum_{k=1}^{n} \frac{1}{2^k K_{k-1}},$$

and for all $x \in S_V(\frac{1}{2^{n-1}}) \setminus S_V(\frac{1}{2^n})$, *for all* $\zeta \in \partial_P \mathcal{V}_n(x)$,

$$\inf_{u\in\mathcal{U}} \langle \zeta, f(x,u)\rangle \leq -\frac{W(x)}{K_{n-1}}. \tag{4.1}$$

Proof. We are going to prove only the last assertion. The other ones are left to the reader; they are the consequence of an easy inductive proof.

Let $n \geq 1, x \in S_V(\frac{1}{2^{n-1}}) \setminus S_V(\frac{1}{2^n})$, and $\zeta \in \partial_P \mathcal{V}_n(x)$.

We remark that for all y not in $S_V(\frac{1}{2^n})$, we have

$$\mathcal{V}_n(y) = \min\left\{\mathcal{V}_{n-1}(y), \mathcal{V}_{n-1}(y) + \frac{V(y) - \frac{1}{2^{n-1}}}{K_{n-1}}\right\}.$$

For the x chosen above, the minimum is attained in the second term, so

$$\zeta \in \partial_P\left[\mathcal{V}_{n-1}(x) + \frac{V(x) - \frac{1}{2^{n-1}}}{K_{n-1}}\right] = \partial_P\left[\mathcal{V}_{n-1}(x) + \frac{V(x)}{K_{n-1}}\right]. \tag{4.2}$$

First case. $n > 1$. We remark now that for all $y \in S_V(\frac{1}{2^{n-2}})$,

$$\mathcal{V}_{n-1}(y) = \max\left\{C_{n-2} + \frac{V(y) - \frac{1}{2^{n-2}}}{K_{n-2}}, C_{n-2} - \frac{1}{2^{n-1}K_{n-2}}\right\},$$

where C_{n-2} is the value of $\mathcal{V}_{n-2}$ on the set $S_V(\frac{1}{2^{n-2}})$. We deduce by (4.2) that

$$\zeta \in \partial_P\left[\max\left\{\frac{V(x)}{K_{n-2}} + A, A'\right\} + \frac{V(x)}{K_{n-1}}\right],$$

where $A = C_{n-2} - \frac{1}{2^{n-2}K_{n-2}}$ and $A' = C_{n-2} - \frac{1}{2^{n-1}K_{n-2}}$.
Hence, we obtain that ζ is in the set

$$\partial_P\left[\max\left\{\left(\frac{1}{K_{n-2}} + \frac{1}{K_{n-1}}\right)V(x) + A, \frac{V(x)}{K_{n-1}} + A'\right\}\right].$$

Now, by the basic calculus on the proximal subgradients, we have

$$\zeta \in \operatorname{co}\left\{\left(\frac{1}{K_{n-2}} + \frac{1}{K_{n-1}}\right)\partial_P V(x), \frac{1}{K_{n-1}}\partial_P V(x)\right\}.$$

Then, there exists ζ_1 and ζ_2 in $\partial_P V(x)$ and $t \in [0,1]$ such that

$$\begin{aligned}\zeta &= t\left(\frac{1}{K_{n-2}} + \frac{1}{K_{n-1}}\right)\zeta_1 + (1-t)\frac{1}{K_{n-1}}\zeta_2\\ &= \left[t\left(\frac{1}{K_{n-2}} + \frac{1}{K_{n-1}}\right) + (1-t)\frac{1}{K_{n-1}}\right]\hat{\zeta},\end{aligned}$$

where $\hat{\zeta} \in \partial_P V(x)$, because $\partial_P V(x)$ is a convex set. Now, we invoke the decrease property of V, $\inf_{u\in\mathcal{U}}\langle\hat{\zeta}, f(x,u)\rangle \leq -W(x)$. Then

$$\inf_{u\in\mathcal{U}} \langle \zeta, f(x,u)\rangle \leq -\frac{W(x)}{K_{n-1}},$$

which gives the result.

Second case. If $n = 1$, the proof is similar. □

Now, note that for each $x \neq 0$, the sequence $(\mathcal{V}_n(x))_{n\geq 0}$ is stationary; thus it converges. So, we can define

$$\mathcal{V}(x) := \lim_{n\to\infty} \mathcal{V}_n(x) - C,$$

where $C := 1 - \sum_{n=0}^{+\infty} \frac{1}{2^{n+1}K_n} \in [0,1]$ (because the Lipschitz constants have been chosen greater than 1).

By the preceding lemma, $\mathcal{V}_n$ is always positive and for $x = 0$,

$$\mathcal{V}_n(0) = 1 - \sum_{k=1}^{n} \frac{1}{2^k K_{k-1}} \rightarrow_{n\to\infty} C =: \mathcal{V}(0) + C.$$

We deduce that $\mathcal{V}(0) = 0$ and then that $\mathcal{V}$ is positive definite. On the other hand, it is locally Lipschitz everywhere as a simple limit of Lipschitz functions (with the same constant in each compact set on $\mathbb{X}$), and it verifies the decreasing property (2.2) with a continuous positive definite function $\mathcal{W}$ defined as follows:

$$\mathcal{W}(x) := \inf_{y\in\mathbb{X}} \{w(y) + \|x - y\|\} \ \forall\ x \in \mathbb{X},$$

where

$$w(x) := \begin{cases} W(x) & \text{if } x \notin S_V(1), \\ \frac{W(x)}{K_n} & \text{if } x \in S_V(\frac{1}{2^n}) \setminus S_V(\frac{1}{2^{n+1}}), \\ 0 & \text{if } x = 0. \end{cases}$$

The decrease property is an immediate consequence of (4.1).

To complete the proof of Theorem 1, we now have to prove the existence of a CLF which is continuous on $\mathbb{X}$ and locally Lipschitz outside the origin. We begin by defining a multifunction F, which is useful because it is uniformly bounded:

$$\forall x \in \mathbb{X}, F(x) := \text{cl} \quad \text{co} \left\{ \frac{f(x,u)}{1 + \|f(x,u)\|}, u \in \mathcal{U} \right\}.$$

We study the differential inclusion

$$\dot{x}(t) \in F(x(t)) \quad \text{a.e.}$$

This dynamic has the same properties as the system (2.1).

PROPOSITION 1.

(i) *F is locally Lipschitz and compact convex valued.*

(ii) *The system $\dot{x}(t) \in F(x(t))$ is GAC.*

Proof.

(i) First of all, it is clear by construction that for all $x \in \mathbb{X}$, the set $F(x)$ is compact convex. Moreover, since the function

$$x \mapsto \frac{f(x,u)}{1 + \|f(x,u)\|}$$

is locally Lipschitz uniformly in u, we deduce that the multifunction F is also locally Lipschitz.

(ii) We now prove the global asymptotic controllability of the differential inclusion

$$\dot{x}(t) \in F(x(t)) \quad \text{a.e.} \tag{4.3}$$

Let $x \in \mathbb{X}$ with $\|x\| \leq R$ be given. By assumption, there is a trajectory $x(\cdot)$ of (2.1) on $[0, \infty)$ which verifies the assumptions of global asymptotic controllability (Definition 2.3). We set

$$\phi(t) := \int_0^t [1 + \|\dot{x}(s)\|] ds,$$

and we define a function $\tilde{x}$ on $[0, \infty]$ by

$$\tilde{x}(\tau) := x(t),$$

where $t = t(\tau)$ is determined in $[0, \infty]$ by

$$\tau = \int_0^t [1 + \|\dot{x}(s)\|] ds.$$

(This change of variables or time scale is known as the Erdmann transform.) Then

$$\frac{d\tilde{x}}{d\tau} = \frac{\dot{x}(t)}{1 + \|\dot{x}(t)\|} \in F(\tilde{x}(\tau)) \quad \text{a.e.},$$

so that $\tilde{x}$ is an F-trajectory.

But by construction, for all $\tau \geq 0, \|\tilde{x}(\tau)\| \leq M(R)$, and if $\tau \geq \phi(T(r, R))$, then $\|\tilde{x}(\tau)\| \leq r$.

The trajectory $x(\cdot)$ remains in the ball $M(R)\bar{B}$, so if N_R denotes the maximum of $\|f(x, u)\|$ for $x \in M(R)\bar{B}$ and $u \in U$ (finite by Assumption 2.2), we have

$$\forall t \geq 0, \phi(t) \leq t(1 + N_R).$$

We deduce that if $\tau \geq T(r, R)(1+N_R)$, then $\tau \geq \phi(T(r, R))$ and consequently $\|\tilde{x}(\tau)\| \leq r$.

The differential inclusion (4.3) is GAC with suitable values $M(R)$ and $\tilde{T}(r, R) := T(r, R)(1 + N_R)$. □

We shall use the notation $M(\cdot)$ and $\tilde{T}(\cdot, \cdot)$ for the functions of the global asymptotic stability of F.

Remark 4.2. *We have, in fact, by a similar proof the following property.*

Proposition 2. *Let $\beta : \mathbb{X} \longrightarrow \mathbb{R}_{>0}$ be locally Lipschitz. Then the differential inclusion*

$$\dot{x}(t) \in \beta(x(t))F(x(t)) \quad \textit{a.e.}$$

is locally Lipschitz with convex compact values and it is GAC with appropriate constants $M(R) \downarrow 0$ and $\tilde{T}_\beta(r, R) = T(r, R) \max_{x \in M(R)\bar{B}} \{\beta(x)^{-1}\}$.

We proceed now to define iteratively a sequence of value functions for targets shrinking down to the equilibrium. We make an inductive proof which is mainly based on the Theorem 3. In the first step we define the first value function V_0, and then in order to set out the idea of the induction, we construct explicitly the second value function V_1 in the second step. Finally, we finish our proof by giving the induction for all n (third step) and by defining our definitive CLF in the fourth step.

First step. We begin the inductive proof by setting a first value function V_0.

First of all, we define a new multifunction Γ_0 as follows:

$$\Gamma_0(x) := \begin{cases} \left[1 + (\|x\| - M(1))\frac{\tilde{T}(\frac{1}{2},1)}{M(1)^2}\right]^{-1} F(x) & \text{for } \|x\| \geq M(1), \\ F(x) & \text{for } 1 \leq \|x\| \leq M(1), \\ F(x) + 4[1 - \|x\|]\bar{B} & \text{for } \frac{1}{2} \leq \|x\| \leq 1, \\ F(x) + 2\bar{B} & \text{for } \|x\| \leq \frac{1}{2}. \end{cases}$$

By construction (and by Proposition 2), we have immediately the following lemma.

Lemma 4.3. *Γ_0 is compact convex valued, locally Lipschitz, uniformly bounded (by 1), and the differential inclusion $\dot{x} \in \Gamma_0(x(t))$ is GAC.*

On the other hand, $\bar{B} \subset \Gamma_0(x)$ for all x in $\frac{1}{2}\bar{B}$. Hence, Theorem 3 can be applied with $\mathcal{T} = \mathcal{T}_0 := \frac{1}{2}\bar{B}$ and $L = L_0 := 1$. So we define the value function

$$V_0(x) := \inf\left\{T : x(0) = x, \dot{x}(t) \in \Gamma_0(x(t)) \text{ a.e. and } x(T) \in \frac{1}{2}\bar{B}\right\}$$

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

Lemma 4.4. *V_0 is locally Lipschitz on $\mathbb{X}$, positive, proper, and for all $x \notin B$,*

$$\forall \zeta \in \partial_P V_0(x), \min_{v \in F(x)} \langle \zeta, v \rangle \leq -1.$$

Proof. This is an easy corollary of Theorem 3. □

We set $m_0 := \max\{V_0(x); \|x\| \leq 1\}$ and $S_0 := \{x; V_0(x) \leq m_0\}$. We define a new function $\tilde{V_0}$ as follows:

$$\tilde{V_0}(x) := \max\{0, V_0(x) - m_0\}.$$

Lemma 4.5.

(a) *$\tilde{V_0}(x) = 0 \Longleftrightarrow x \in S_0$.*

(b) *$\bar{B} \subset S_0 \subset 3M(1)\bar{B}$.*

(c) *For all $x \notin S_0$, for all $\zeta \in \partial_P \tilde{V_0}(x), \min_{v \in F(x)} \langle \zeta, v \rangle \leq -1$.*

Proof.

(a) The proof of (a) is obvious by the definition of $\tilde{V_0}$.

(b) The first inclusion is given by the definition of S_0. However, the second one is less easy. Since the system $\dot{x} \in F(x)$ is GAC, for all $\alpha \in \bar{B}$ there exists a F-trajectory $x(\cdot)$ such that
   (1) $x(0) = \alpha$ and $\dot{x}(t) \in F(x(t))$ a.e.,
   (2) for all $t \geq 0, \|x(t)\| \leq M(1)$,
   (3) for all $t \geq \tilde{T}(\frac{1}{2}, 1), \|x(t)\| \leq \frac{1}{2}$.

   Now, from the definition of Γ_0, for all $x \in M(1)\bar{B},\ F(x) \subset \Gamma_0(x)$; then

$$V_0(\alpha) \leq \tilde{T}\left(\frac{1}{2}, 1\right).$$

   Consequently, $m_0 \leq \tilde{T}(\frac{1}{2}, 1)$.

   Let us consider now $\alpha \in \mathbb{X}$ such that $\|\alpha\| \geq 3M(1)$.

We remark that for $\|x\| \geq 2M(1)$, we have

$$\|\Gamma_0(x)\| \leq \left[1 + \frac{\tilde{T}(\frac{1}{2},1)}{M(1)}\right]^{-1}.$$

Then the time needed by a Γ_0 trajectory with initial condition α to reach the ball $2M(1)\bar{B}$ is greater than $[1 + \frac{\tilde{T}(\frac{1}{2},1)}{M(1)}]M(1)$.

Hence, $V_0(\alpha) \geq M(1) + \tilde{T}(\frac{1}{2},1) > m_0$.

Consequently, $S_0 \subset 3M(1)\bar{B}$.

(c) This last assertion is a consequence of Lemma 4.4. □

Second step. We present now the construction of the second value function V_1 in order to give the idea of our inductive proof.

We set

$$\Gamma_1(x) := \begin{cases} \left[1 + (\|x\| - M(\frac{1}{2}))\frac{\tilde{T}(\frac{1}{4},\frac{1}{2})}{M(\frac{1}{2})^2}\right]^{-1} F(x) & \text{for } \|x\| \geq M(\frac{1}{2}), \\ F(x) & \text{for } \frac{1}{2} \leq \|x\| \leq M(\frac{1}{2}), \\ F(x) + 8\left[\frac{1}{2} - \|x\|\right]\bar{B} & \text{for } \frac{1}{4} \leq \|x\| \leq \frac{1}{2}, \\ F(x) + 2\bar{B} & \text{for } \|x\| \leq \frac{1}{4}. \end{cases}$$

We have immediately the following result.

LEMMA 4.6. *Γ_1 is compact convex valued, and the differential inclusion $\dot{x}(t) \in \Gamma_1(x(t))$ is GAC (with possible constants $M_1(R) = M(R) \downarrow 0$ and $\tilde{T}_1(r,R)$).*

We need an auxiliary function with the local Lipschitz property. We define for all $x \in \mathbb{X}$

$$B_0(x) := \max\{V_0(y) : \|y\| \leq \|x\| + M(1)\}.$$

As before, the new multifunction leads to a value function R_1 associated to the set $\mathcal{T}_1 := \frac{1}{4}\bar{B}$. We set for all x in $\mathbb{X}$

$$R_1(x) := \inf\left\{\int_0^T L_1(x(t))dt : x(0) = x, \dot{x} \in \Gamma_1(x) \text{ a.e. and } x(T) \in \mathcal{T}_1\right\},$$

where $L_1(x) := 1 + \max\{0, \|x\| - 3M(1)\}\frac{B_0(x)}{\rho_1 M(1)^2}$ and

$$\rho_1 := \frac{m_0/2}{m_0\left[1 + (3M(1) - M(\frac{1}{2}))\frac{\tilde{T}(\frac{1}{4},\frac{1}{2})}{M(\frac{1}{2})^2}\right] + \tilde{T}_1(\frac{1}{4},1)} \leq 1.$$

Theorem 3 gives the following lemma.

LEMMA 4.7.

(a) *R_1 is locally Lipschitz on $\mathbb{X}$.*

(b) *For all $\|x\| \geq \frac{1}{2}$, for all $\zeta \in \partial_P R_1(x), \min_{v \in F(x)}\langle\zeta, v\rangle \leq -L_1(x)$.*

Proof. Since L_1 and Γ_1 are locally Lipschitz and the system associated to Γ_1 is GAC, R_1 is finite everywhere and Theorem 3 proves the assertions. □

As in the first step, we are going to evaluate the size of a certain level set given by R_1. We set $m_{R_1} := \max\{R_1(y) : y \in \frac{1}{2}\bar{B}\}$ and

$$S_{R_1}(m_{R_1}) = \{x : R_1(x) \leq m_{R_1}\}.$$

By Proposition 1, for any $x \in \frac{1}{2}\bar{B}$, there exists an F-trajectory $x(\cdot)$ such that

1. $x(0) = x$,
2. for all $t \geq 0, \|x(t)\| \leq M(\frac{1}{2})$,
3. for all $t \geq \tilde{T}(\frac{1}{4}, \frac{1}{2}), x(t) \in \mathcal{T}_1$.

Moreover, for all $x \in M(\frac{1}{2})\bar{B}$, $F(x) \subset \Gamma_1(x)$ and $L_1(x) = 1$; then

$$R_1(x) \leq \tilde{T}\left(\frac{1}{4}, \frac{1}{2}\right).$$

Consequently, $m_{R_1} \leq \tilde{T}(\frac{1}{4}, \frac{1}{2})$.

Now, we consider an initial state α such that $\|\alpha\| \geq 3M(\frac{1}{2})$.

We remark that for $\|x\| \geq 2M(1)$, we have

$$\|\Gamma_1(x)\| \leq \left[1 + \frac{\tilde{T}(\frac{1}{4}, \frac{1}{2})}{M(\frac{1}{2})}\right]^{-1}.$$

Then the time used by a Γ_1-trajectory with initial condition α to reach the ball $2M(\frac{1}{2})\bar{B}$ is greater than $[1 + \frac{\tilde{T}(\frac{1}{4}, \frac{1}{2})}{M(\frac{1}{2})}]M(\frac{1}{2})$.

Hence, $L_1 \geq 1$ implies $R_1(\alpha) \geq M(\frac{1}{2}) + \tilde{T}(\frac{1}{4}, \frac{1}{2}) > m_{R_1}$.

Consequently,

$$S_{R_1}(m_{R_1}) \subset 3M\left(\frac{1}{2}\right)\bar{B}.$$

Indeed, from the proof follows the following lemma.

Lemma 4.8. *$\frac{1}{2}\bar{B} \subset S_{R_1}(m_{R_1}) \subset 3M(\frac{1}{2})\bar{B}$.*

We want now to compare R_1 with V_0.

Lemma 4.9.

(a) *For all $x \in S_0, \rho_1 R_1(x) \leq \frac{m_0}{2}$.*

(b) *If $\|x\| \geq 5M(1)$, then $V_0(x) \leq \rho_1 R_1(x)$.*

Proof.

(a) Let $x \in S_0$. Indeed, there exists a Γ_0-trajectory $x(\cdot)$ which connects x to the set $\bar{B}$ in time $T_x \leq V_0(x) \leq m_0$. Hence, for all $t \geq 0,\ x(t) \in S_0 \subset 3M(1)\bar{B}$ (by Lemma 4.5(b)). In the zone $\|x\| \in [1, 3M(1)]$ we can write $\Gamma_1(x) \subset \beta(x)\Gamma_0(x)$ with $\beta(x)$ as follows (assuming that $M(\frac{1}{2}) \geq 1$):

$$\beta(x) := \begin{cases} 1 & \text{if } \|x\| \in [\frac{1}{2}, M(\frac{1}{2})], \\ \left[1 + (\|x\| - M(\frac{1}{2}))\frac{\tilde{T}(\frac{1}{4}, \frac{1}{2})}{M(\frac{1}{2})^2}\right]^{-1} & \text{if } \|x\| \in [M(\frac{1}{2}), M(1)], \\ \dfrac{\left[1 + (\|x\| - M(\frac{1}{2}))\frac{\tilde{T}(\frac{1}{4}, \frac{1}{2})}{M(\frac{1}{2})^2}\right]^{-1}}{\left[1 + (\|x\| - M(1))\frac{\tilde{T}(\frac{1}{2}, 1)}{M(1)^2}\right]^{-1}} & \text{if } \|x\| \in [M(1), 3M(1)]. \end{cases}$$

We observe that if $M(\frac{1}{2}) < 1$, we have to omit it in the definition of β. Now, an appropriate change of variables (see Proposition 2) shows that there exists a Γ_1-trajectory $x(\cdot)$ which remains in $3M(1)\bar{B}$ and drives x to $\bar{B}$ in a time $T \leq T_x \max_{\|x\| \in [1, 3M(1)]} \beta(x)^{-1}$.

Thus, we obtain $T \leq m_0[1 + (3M(1) - M(\frac{1}{2}))\frac{\tilde{T}(\frac{1}{4}, \frac{1}{2})}{M(\frac{1}{2})^2}]$.

Now, we can extend this trajectory to $\mathcal{T}_1$ with the following property (by

Lemma 4.6): for all $t \geq T, x(t) \in M(1)\bar{B}$ and $x(T + \tilde{T}_1(\frac{1}{4}, 1)) \in \frac{1}{4}\bar{B}$. Thus, we have constructed a trajectory which remains in $3M(1)\bar{B}$ (where $L_1 = 1$) and reaches the set $\mathcal{T}_1$.

Consequently, $R_1(x) \leq m_0[1 + (3M(1) - M(\frac{1}{2}))\frac{\tilde{T}(\frac{1}{4},\frac{1}{2})}{M(\frac{1}{2})^2}] + \tilde{T}_1(\frac{1}{4}, 1)$.

We conclude by the definition of ρ_1.

(b) Let x be such that $\|x\| \geq 5M(1)$. By the definition of B_0 we have

$$\|y\| \geq \|x\| - M(1) \Longrightarrow B_0(y) \geq V_0(x) \Longrightarrow L_1(y) \geq 1 + \frac{V_0(x)}{\rho_1 M(1)^2}.$$

On the other hand, the time required to go from $\{\|y\| \geq \|x\| - M(1)\}$ to $\{\|y\| \geq \|x\| - 2M(1)\}$ is greater than $M(1)$. (The dynamic is bounded by 1.) Consequently,

$$\begin{aligned} R_1(x) &\geq M(1)\left[1 + \frac{V_0(x)}{\rho_1 M(1)}\right] \\ &\geq \frac{V_0(x)}{\rho_1}. \qquad \square \end{aligned}$$

We finish this step by defining a new function V_1 as follows.

$$\forall x \in \mathbb{X}, V_1(x) := \min\{\tilde{V_0}(x) + m_0, \rho_1 R_1(x)\}.$$

We set $m_1 := \max\{V_1(x) : x \in \frac{1}{2}\bar{B}\}$ and $S_1 := \{x : V_1(x) \leq m_1\}$. We have the following lemma.

Lemma 4.10. *V_1 is locally Lipschitz on $\mathbb{X}$. Moreover, we have*

(a) *$m_1 \leq \frac{m_0}{2}$;*

(b) *for all $x \in S_0 \cup S_1, V_1(x) = \rho_1 R_1(x)$;*

(c) *$\frac{1}{2}\bar{B} \subset S_1 \subset 3M(\frac{1}{2})\bar{B}$;*

(d) *if $\|x\| \geq 5M(1)$, then $V_1(x) = V_0(x)$;*

(e) *for $\frac{1}{2} \leq \|x\| \leq 5M(1)$, for all $\zeta \in \partial_P V_1(x), \min_{v\in F(x)}\langle\zeta, v\rangle \leq -\rho_1$;*

(f) *for $\|x\| > 5M(1)$, for all $\zeta \in \partial_P V_1(x), \min_{v\in F(x)}\langle\zeta, v\rangle \leq -1$.*

Proof.

(a) By Lemma 4.9 (a), for any $x \in S_0, \rho_1 R_1(x) \leq \frac{m_0}{2}$. Hence by definition of V_1, for all $x \in S_0, V_1(x) = \rho_1 R_1(x)$. Thus we conclude by remarking that $\frac{1}{2}\bar{B} \subset S_0$. We have, in fact, $m_1 = \rho_1 m_{R_1}$.

(b) Let $x \in S_0 \cup S_1$. If $x \in S_0$, we have shown the equality in the first assertion. Otherwise, $V_1(x) \leq m_1 \leq \frac{m_0}{2}$ implies the equality.

(c) If $x \in S_1$, then $V_1(x) = \rho_1 R_1(x) \leq m_1$. And by the remark in (a), $R_1(x) \leq m_{R_1}$, which gives the inclusion.

(d) For $\|x\| \geq 5M(1)$, we have that $V_0(x) = \tilde{V_0}(x) + m_0$ (because $S_0 \subset 3M(1)\bar{B}$). We conclude by Lemma 4.9(b).

(e) Let $x \in \mathbb{X}$ such that $\frac{1}{2} \leq \|x\| \leq 5M(1)$ and $\zeta \in \partial_P V_1(x)$. We recall the definition of $V_1(x)$:

$$V_1(x) := \min\{\tilde{V_0}(x) + m_0, \rho_1 R_1(x)\}.$$

First case. The minimum is attained by the second term. Then $\zeta \in \partial_P \rho_1 R_1(x) = \rho_1 \partial_P R_1(x)$. We conclude by Lemma 4.7 (b).

Second case. The minimum is attained by the first term and not by the second one. In this case, $x \notin S_0$ and $\zeta \in \partial_P(\tilde{V_0} + m_0)(x) = \partial_P \tilde{V_0}(x)$. We conclude by Lemma 4.5 (c).

(f) This is an easy consequence of Lemma 4.9(b). □

Third step. We finish the construction of the sequence $(V_n)_{n\geq 0}$ by induction on n.

Assume $(V_k, \mathcal{T}_k, L_k, R_k, \Gamma_k)$ have already been defined for $1 \leq k \leq n$ with the following properties:

1. $\frac{1}{2^k}\bar{B} \subset S_k \subset 3M(\frac{1}{2^k})\bar{B}$,
2. for $\|x\| \geq 5M(\frac{1}{2^{k-1}}), V_k(x) = V_{k-1}(x)$,
3. for $\frac{1}{2^k} \leq \|x\| \leq 5M(\frac{1}{2^{k-1}})$, for all $\zeta \in \partial_P V_k(x), \min_{v\in F(x)}\langle \zeta, v\rangle \leq -\rho_k$,
4. $L_k = 1$ on the ball $3M(\frac{1}{2^{k-1}})\bar{B}$,
5. for all $x \in \mathbb{R}^N, V_k(x) = 0 \iff x \in \frac{1}{2^{k+1}}\bar{B} =: \mathcal{T}_k$,
6. for all $k \in [1, n], m_k \leq \frac{m_{k-1}}{2}$, and $\rho_k \leq \rho_{k-1} \leq 1 =: \rho_0$,

where

$$m_k := \max\left\{V_k(x); \|x\| \leq \frac{1}{2^k}\right\},\ S_k := \{x; V_k(x) \leq m_k\},$$

and the ρ_k's are some positive constants.

As before, we can define a new function V_{n+1}. We proceed as follows: for all $x \in \mathbb{X}$, we set $\Gamma_{n+1}(x) :=$

$$\begin{cases} \left[1 + (\|x\| - M(\frac{1}{2^{n+1}}))\frac{\tilde{T}(\frac{1}{2^{n+2}}, \frac{1}{2^{n+1}})}{M(\frac{1}{2^{n+1}})^2}\right]^{-1} F(x) & \text{if } \|x\| \geq M(\frac{1}{2^{n+1}}), \\ F(x) & \text{if } \|x\| \in [\frac{1}{2^{n+1}}, M(\frac{1}{2^{n+1}})], \\ F(x) + 2^{n+3}[\frac{1}{2^{n+1}} - \|x\|]\bar{B} & \text{if } \|x\| \in [\frac{1}{2^{n+2}}, \frac{1}{2^{n+1}}], \\ F(x) + 2\bar{B} & \text{if } \|x\| \leq \frac{1}{2^{n+2}}. \end{cases}$$

As before, we need an auxiliary function with the local Lipschitz property. We define for all $x \in \mathbb{X}$

$$B_n(x) := \max\left\{V_n(y) : \|y\| \leq \|x\| + M\left(\frac{1}{2^n}\right)\right\}.$$

From this multifunction, we define a value function associated to the set $\mathcal{T}_{n+1} := \frac{1}{2^{n+2}}\bar{B}$. We set for any $x \in \mathbb{X}$

$$R_{n+1}(x) := \inf\left\{\int_0^T L_{n+1}(x(t))dt : x(0) = x, \dot{x} \in \Gamma_{n+1}(x),\ \text{and } x(T) \in \mathcal{T}_{n+1}\right\},$$

where $L_{n+1}(x) := 1 + \max\{0, \|x\| - 3M(\frac{1}{2^n})\}\frac{B_n(x)}{\rho_{n+1}M(\frac{1}{2^n})^2}$ with

$$\rho_{n+1} := \frac{\rho_n m_n/2}{m_n\left[1 + (3M(\frac{1}{2^n}) - M(\frac{1}{2^{n+1}}))\frac{\tilde{T}(\frac{1}{2^{n+2}}, \frac{1}{2^{n+1}})}{M(\frac{1}{2^{n+1}})^2}\right] + \tilde{T}_{n+1}(\frac{1}{2^{n+2}}, \frac{1}{2^n})} \leq \rho_n.$$

The differential inclusion $\dot{x} \in \Gamma_{n+1}(x)$ is GAC; we denote by $\tilde{T}_{n+1}(\cdot, \cdot)$ its new constant. (We saw that we can choose $\tilde{M}_{n+1} = M$.) On the other hand, we set

$$m_{R_{n+1}} := \max\left\{R_{n+1}(y) : y \in \frac{1}{2^{n+1}}\bar{B}\right\},$$

$$\text{and } S_{R_{n+1}}(m_{R_{n+1}}) := \{x : R_{n+1}(x) \leq m_{R_{n+1}}\}.$$

LEMMA 4.11.

(a) *R_{n+1} is locally Lipschitz on $\mathbb{X}$.*

(b) *For all $\|x\| \geq \frac{1}{2^{n+1}}$, for all $\zeta \in \partial_P R_{n+1}(x), \min_{v \in F(x)} \langle \zeta, v \rangle \leq -L_{n+1}(x)$.*

(c) *$\frac{1}{2^{n+1}} \bar{B} \subset S_{R_{n+1}}(m_{R_{n+1}}) \subset 3M(\frac{1}{2^{n+1}})\bar{B}$.*

(d) *For all $x \in S_n, \rho_{n+1} R_{n+1}(x) \leq \frac{m_n}{2}$.*

(e) *If $\|x\| \geq 5M(\frac{1}{2^n})$, then $V_n(x) \leq \rho_{n+1} R_{n+1}(x)$.*

Proof. (a) and (b) are evident by Theorem 1. The assertion (c) is proved as before (Lemma 4.8). We prove now (d) and (e); we begin with (e).

Let $\|x\| \geq 5M(\frac{1}{2^n})$ be given. By the definition of B_n we have

$$\|y\| \geq \|x\| - M\left(\frac{1}{2^n}\right) \Longrightarrow B_n(y) \geq V_n(x) \Longrightarrow L_{n+1}(y) \geq 1 + \frac{V_n(x)}{\rho_{n+1} M(\frac{1}{2^n})^2}.$$

On the other hand, the time required for driving from $\{\|y\| \geq \|x\| - M(\frac{1}{2^n})\}$ to $\{\|y\| \geq \|x\| - 2M(\frac{1}{2^n})\}$ is greater than $M(\frac{1}{2^n})$. (The dynamic is bounded by 1.) Consequently,

$$\begin{aligned} R_{n+1}(x) &\geq M\left(\frac{1}{2^n}\right)\left[1 + \frac{V_n(x)}{\rho_{n+1} M(\frac{1}{2^n})}\right] \\ &\geq \frac{V_n(x)}{\rho_{n+1}}. \end{aligned}$$

We prove now (d). Let $x \in S_n$. Indeed, there exists a Γ_n-trajectory $x(\cdot)$ which takes x to the set $\frac{1}{2^n}\bar{B}$ in time $T_x \leq V_n(x) \leq m_n$ and which remains in S_n (because $S_n \subset 3M(\frac{1}{2^n}\bar{B}$ and $L_n = 1$ on $3M(\frac{1}{2^n}\bar{B})$. In the zone $\|x\| \in [\frac{1}{2^n}, 3M(\frac{1}{2^n})]$ we can write $\Gamma_{n+1}(x) \subset \beta(x)\Gamma_n(x)$ with $\beta(x)$ as follows (assuming that $M(\frac{1}{2^{n+1}}) \geq \frac{1}{2^n}$; we adapt otherwise):

$$\beta(x) := \begin{cases} 1 & \text{for } \|x\| \in [\frac{1}{2^n}, M(\frac{1}{2^{n+1}})], \\ \left[1 + (\|x\| - M(\frac{1}{2^{n+1}}))\frac{\tilde{T}(\frac{1}{2^{n+2}}, \frac{1}{2^{n+1}})}{M(\frac{1}{2^{n+1}})^2}\right]^{-1} & \text{for } \|x\| \in [M(\frac{1}{2^{n+1}}), M(\frac{1}{2^n})], \\ \dfrac{\left[1+(\|x\|-M(\frac{1}{2^{n+1}}))\frac{\tilde{T}(\frac{1}{2^{n+2}}, \frac{1}{2^{n+1}})}{M(\frac{1}{2^{n+1}})^2}\right]^{-1}}{\left[1+(\|x\|-M(\frac{1}{2^n}))\frac{\tilde{T}(\frac{1}{2^{n+1}}, 2^n)}{M(\frac{1}{2^n})^2}\right]^{-1}} & \text{for } \|x\| \in [M(\frac{1}{2^n}), 3M(\frac{1}{2^n})]. \end{cases}$$

An appropriate change of variables (see Proposition 2) shows that there exists a Γ_{n+1}-trajectory $x(\cdot)$ which remains in $3M(\frac{1}{2^n})\bar{B}$ and takes x to $\frac{1}{2^n}\bar{B}$ in a time $T \leq T_x \max_{\|x\| \in [\frac{1}{2^n}, 3M(\frac{1}{2^n})]} \beta(x)^{-1}$.

Thus, we have $T \leq m_n[1 + (3M(\frac{1}{2^n}) - M(\frac{1}{2^{n+1}}))\frac{\tilde{T}(\frac{1}{2^{n+2}}, \frac{1}{2^{n+1}})}{M(\frac{1}{2^{n+1}})^2}]$.

Now, we can extend this trajectory to $\mathcal{T}_{n+1}$ with the following property (by Lemma 4.6): for all $t \geq T, x(t) \in M(\frac{1}{2^n})\bar{B}$ and $x(T + \tilde{T}_{n+1}(\frac{1}{2^{n+2}}, \frac{1}{2^n})) \in \mathcal{T}_{n+1}$. In this way, we have constructed a trajectory which remains in $3M(\frac{1}{2^n})\bar{B}$ (where $L_{n+1} = 1$) and reaches the set $\mathcal{T}_{n+1}$.

Consequently,

$$R_{n+1}(x) \leq m_n \left[1 + \left(3M\left(\frac{1}{2^n}\right) - M\left(\frac{1}{2^{n+1}}\right)\right)\frac{\tilde{T}(\frac{1}{2^{n+2}}, \frac{1}{2^{n+1}})}{M(\frac{1}{2^{n+1}})^2}\right] + \tilde{T}_1\left(\frac{1}{2^{n+2}}, \frac{1}{2^n}\right).$$

We conclude by the definition of ρ_{n+1}. □

We can now define the new function V_{n+1}. We set for all $x \in \mathbb{X}$

$$V_{n+1}(x) := \min\{\tilde{V_n}(x) + m_n, \rho_{n+1} R_{n+1}(x)\},$$

where $\tilde{V_n}(x) := \max\{0, V_n(x) - m_n\}$.

As before, we consider

$$m_{n+1} := \max\left\{V_{n+1}(x) : x \in \frac{1}{2^{n+1}}\bar{B}\right\} \text{ and } S_{n+1} := \{x : V_{n+1}(x) \leq m_{n+1}\}.$$

We have the following lemma.

Lemma 4.12. *V_{n+1} is locally Lipschitz on $\mathbb{X}$. Moreover, we have the following.*

(a) $m_{n+1} \leq \frac{m_n}{2}$.

(b) $\frac{1}{2^{n+1}}\bar{B} \subset S_{n+1} \subset 3M(\frac{1}{2^{n+1}})\bar{B}$.

(c) *If* $\|x\| \geq 5M(\frac{1}{2^n})$, *then* $V_{n+1}(x) = V_n(x)$.

(d) *For* $\frac{1}{2^{n+1}} \leq \|x\| \leq 5M(\frac{1}{2^n})$, *for all* $\zeta \in \partial_P V_{n+1}(x), \min_{v \in F(x)}\langle \zeta, v\rangle \leq -\rho_{n+1}$.

Proof. The proof is similar to the proof of Lemma 4.10. This is left to the reader. □

Fourth step. The function *V*.

We study the convergence of the sequence $(V_n)_{n\geq 0}$; for that, we need a last lemma.

Lemma 4.13. *For all* $k, 0 \leq k \leq n$, *for all* $x \in S_k, V_n(x) \leq m_k$.

Proof. We do an inductive proof. Since the result has already been proved for n=1 (Lemma 4.10), we assume that we have proved the result for $n \geq 1$; we establish the property for $n+1$.

Let us consider $0 \leq k \leq n+1$ and $x \in S_k$.

First case. $k \leq n$.

If $x \notin S_n$, then by definition of V_n, $V_n(x) = \tilde{V_n}(x) + m_n$. Hence, $V_{n+1}(x) \leq V_n(x) \leq m_k$ by induction.

Otherwise, $x \in S_n$. In this case, $\tilde{V_n}(x) = 0$ implies $V_{n+1}(x) \leq m_n \leq m_k$ by the property on the sequence (m_k).

Second case. $k = n+1$.

The property follows from the definition of S_{n+1}. □

We can now conclude. Let us consider a compact set K in $\mathbb{X} \setminus \{0\}$. Then, as $\lim_{n\to\infty} M(\frac{1}{2^n}) = 0$, there exists a positive integer n_K such that

$$\|x\| \geq 5M\left(\frac{1}{2^{n_K}}\right) \quad \forall x \in K.$$

By the second property of the sequence $(V_n)_{n\geq 0}$, for any $n \geq n_K, V_n(x) = V_{n_K}(x)$.

Hence, the sequence $(V_n(x))_{n\geq 0}$ converges for all x in K and the limit is a locally Lipschitz function in K (as a stationary limit of locally Lipschitz functions). On the other hand, for any $n \geq 0, V_n(0) = 0$; so we can define for all $x \in \mathbb{X}$

$$V(x) := \lim_{n\to\infty} V_n(x).$$

By the proof above, V is locally Lipschitz on $\mathbb{X} \setminus \{0\}$, positive definite, and proper (since $V_n = V_0, \quad$ for all n if $\|x\| \geq 5M(1)$); we have to show that V is continuous at the origin. This fact is a consequence of the preceding lemma.

Let us consider $x_p \longrightarrow_{p\to\infty} 0$. We want to show that $f(x_p) \longrightarrow_{p\to\infty} 0$.

Let $\epsilon > 0$. There exists $n_0 \geq 0$ such that $m_{n_0} \leq \epsilon$ (because $m_n \leq \frac{m_0}{2^n}$). Thus, by the last lemma, for all $n \geq n_0$, for all $x \in S_{n_0}, V_n(x) \leq \epsilon$.

There exists $P > 0$ such that if $p \geq P$,

$$x_p \in \frac{1}{2^{n_0}} \bar{B} \subset S_{n_0}.$$

We deduce that for all $n \geq n_0$, for all $p \geq P, V_n(x_p) \leq \epsilon$. By passing to the limit: for all $p \geq P, V(x_p) \leq \epsilon$, which gives the continuity on the origin.

Now, we set for all $x \in \mathbb{X}$

$$w(x) := \begin{cases} \rho_n & \text{if } 5M(\frac{1}{2^{n+1}}) < \|x\| \leq 5M(\frac{1}{2^n}), \\ 1 & \text{if } \|x\| > 5M(1), \\ 0 & \text{if } x = 0. \end{cases}$$

We can now define the function W by

$$\forall x \in \mathbb{X}, W(x) := \inf_{y \in \mathbb{X}} \{w(y) + \|x - y\|\}.$$

W is a positive definite and locally Lipschitz function. The decrease condition (2.2) is the consequence of the stationarity of the sequence $(V_n(x))_{n \geq 0}$ outside the origin and of Lemma 4.12 (d). This completes the proof of Theorem 1.

5. Existence of a semiconcave CLF. We begin with some preliminaries on semiconcavity. It is easy to show that any semiconcave function in Ω is locally Lipschitz. Concave functions are of course semiconcave. Another class of semiconcave functions is that of C^1 functions with locally Lipschitz gradients. Moreover we have the two following lemmas.

LEMMA 5.1. *Let $\Psi : \mathbb{R} \longrightarrow \mathbb{R}$ be an increasing semiconcave function, and let $g : \Omega \longrightarrow \mathbb{R}$ be a semiconcave function on Ω. Then $\Psi \circ g$ is a semiconcave function on Ω.*

LEMMA 5.2. *Let $g, h : \Omega \longrightarrow \mathbb{R}$ be two semiconcave functions on Ω. Then the function* $\min\{g, h\}$ *is semiconcave on Ω.*

A convenient way to build semiconcave approximations of a given function is provided by the method of inf-*convolution*, a standard tool in convex and nonsmooth analysis. Let Ω be a subset of $\mathbb{X}$, and let g be a positive function in Ω. Define, for any $\alpha > 0$,

$$g_\alpha(x) := \inf_{y \in \Omega} \{g(y) + \alpha \|x - y\|^2\}. \tag{5.1}$$

LEMMA 5.3. *Let $g : \mathbb{X} \longrightarrow \mathbb{R}$ be a locally Lipschitz and proper function. Then g_α is semiconcave on $\mathbb{X}$ (in* (5.1) *the infimum is actually a minimum) and, moreover, $g_\alpha \nearrow g$, as $\alpha \rightarrow +\infty$, locally uniformly in $\mathbb{X}$.*

Proof. We leave the proof to the reader. □

We can link the proximal subdifferentials of u and its inf-convolution. We have the following lemma. (We refer to [14, Thm. 5.1, p. 44] for the proof.)

LEMMA 5.4. *Suppose that $x \in \mathbb{X}$ is such that $\partial_P g_\alpha(x)$ is nonempty. Then there exists a point $\bar{y} \in \mathbb{X}$ satisfying the following.*

a) *The infimum in* (5.1) *is attained uniquely at $\bar{y}$.*

b) *The proximal subgradient $\partial_P g_\alpha(x)$ is the singleton $\{2\alpha(x - \bar{y})\}$.*

c) $2\alpha(x - \bar{y}) \in \partial_P g(\bar{y})$.

Proof of Theorem 2. By Theorem 1, there exists a control-Lyapunov pair for the system (2.1); without loss of generality, we can suppose that the function W is

1-Lipschitz on $\mathbb{X}$. (Otherwise, we can set $\tilde{W}(x) := \inf_{y\in\mathbb{X}}\{W(y) + \|x-y\|\}$.) For any $0 < r < R$, we define the following sets:

$$S_V[r,R] := \{x \in \mathbb{X} : V(x) \in [r,R]\} \text{ and } S_V(R) := \{x \in \mathbb{X} : V(x) \leq R\}.$$

Let us consider an integer $n \in \mathbb{N}^*$. By the Lipschitz property of f and V, we can consider $L_f^n \geq 1$ (respectively, $L_V^n \geq 1$) the Lipschitz constant of $f(\cdot,u)$ (respectively, of V) on the level set $S_V(M_n)$, where the constant M_n is defined by

$$M_n := \max\{V(x) : x \in S_V(11n) + \bar{B}\}.$$

On the other hand, we denote by w_n the minimum of W on $S_V[\frac{1}{2n}, 11n]$, and we set

$$\alpha_n := \max\left\{8n(L_V^n)^2 + 1, \frac{2L_V^n(1 + L_V^n L_f^n)}{w_n} + 1, 11n\right\}. \tag{5.2}$$

We define by inf-convolution the function V_{α_n} as follows:

$$V_{\alpha_n}(x) := \inf_{y\in\mathbb{X}}\{V(y) + \alpha_n\|x-y\|^2\}. \tag{5.3}$$

LEMMA 5.5. *Let $x_0 \in S_V(M_n)$. If the infimum in the definition of $V_{\alpha_n}(x_0)$ is attained at $\bar{y}$, then $\|x_0 - \bar{y}\| \leq \min\{\frac{1}{8nL_V^n}, \frac{w_n}{2(1+L_V^n L_f^n)}\}$ and*

$$V(x_0) - \frac{1}{8n} \leq V_{\alpha_n}(x_0) \leq V(x_0).$$

Proof. If the infimum is attained for $\bar{y}$, then $V(\bar{y}) \leq V(x_0) \leq M_n \implies \bar{y} \in S_V(M_n)$. Hence, if $\|x_0 - \bar{y}\| > \min\{\frac{1}{8nL_V^n}, \frac{w_n}{2(1+L_V^n L_f^n)}\}$, then, by definition of L_f^n and L_V^n,

$$\begin{aligned}
V_{\alpha_n}(x_0) &= V(\bar{y}) + \alpha_n\|x_0 - \bar{y}\|^2 \\
&\geq V(x_0) - L_V^n\|x_0 - \bar{y}\| + \alpha_n\|x_0 - \bar{y}\|^2 \\
&\geq V(x_0) + \|x_0 - \bar{y}\|[\alpha_n\|x_0 - \bar{y}\| - L_V^n] \\
&\geq V(x_0) + \|x_0 - \bar{y}\|\left[\alpha_n \min\left\{\frac{1}{8nL_V^n}, \frac{w_n}{2(1+L_V^n L_f^n)}\right\} - L_V^n\right] \\
&> V(x),
\end{aligned}$$

we find a contradiction. Hence, $\|x_0 - \bar{y}\| \leq \min\{\frac{1}{8nL_V^n}, \frac{w_n}{2(1+L_V^n L_f^n)}\}$. On the other hand, we have found the estimate

$$V_{\alpha_n}(x_0) \geq V(x_0) + \|x_0 - \bar{y}\|[\alpha_n\|x_0 - \bar{y}\| - L_V^n].$$

Consequently, $V_{\alpha_n}(x_0) \geq V(x_0) - L_V^n\|x_0 - \bar{y}\|$, which implies the desired inequality by the bound on $\|x_0 - \bar{y}\|$. □

LEMMA 5.6. *Let $x_0 \in S_V[\frac{1}{2n}, 11n]$ and $\zeta \in \partial_P V_{\alpha_n}(x_0)$; then*

$$\inf_{u\in U}\langle\zeta, f(x_0,u)\rangle \leq -\frac{W(x_0)}{2}.$$

Proof. By Lemmas 5.4 and 5.5, the infimum in the definition of $V_{\alpha_n}(x_0)$ is attained uniquely at a point $\bar{y} \in S_V(11n)$ which satisfies $\|x_0 - \bar{y}\| \leq \frac{w_n}{2(1+L_V^n L_f^n)}$ and such that

$\zeta \in \partial_P V(\bar{y})$. Thus, by the Lipschitz properties of f, V, and W, we can write

$$\begin{aligned}\inf_{u\in U}\langle \zeta, f(x_0,u)\rangle &\le \inf_{u\in U}\langle \zeta, f(\bar{y},u)\rangle + \sup_{u\in U}\|\zeta\|\|f(x_0,u)-f(\bar{y},u)\|\\ &\le -W(\bar{y}) + L_V^n L_f^n\|x_0-\bar{y}\| \quad \text{(decrease condition)}\\ &\le -W(x_0) + (1+L_V^n L_f^n)\|x_0-\bar{y}\|\\ &\le -W(x_0) + \frac{w_n}{2} \le -\frac{W(x_0)}{2}. \qquad \square\end{aligned}$$

LEMMA 5.7. *For each n, there exists an increasing, C^∞ function $\Psi_n : \mathbb{R}_{\ge 0} \longrightarrow \mathbb{R}_{\ge 0}$ satisfying the following properties.*

(i) *For all $t \in [0, \frac{1}{2n}], \Psi_n(t) = t + \frac{1}{8n}$.*

(ii) *For all $t \in [\frac{1}{n} - \frac{1}{8n}, 10n], \Psi_n(t) = t$.*

(iii) *For all $t \in [11n - \frac{1}{8n}, \infty), \Psi_n(t) \ge 11n + \max\{V(x) : V_{\alpha_n}(x) \le t\}$.*

(iv) *For all $t \ge 0, \Psi_n'(t) \ge \frac{1}{2}$.*

Proof. The different intervals being disjoint, the different properties (i)–(iv) allow us to define an increasing, piecewise, affine function that can be regularized in order to get a C^∞ function Ψ_n. $\square$

The function $\tilde{V}_n := \Psi_n \circ V_{\alpha_n}$ is semiconcave on $\mathbb{X}$ by Lemma 5.1. The definitive Lyapunov pair $(\mathcal{V}, \mathcal{W})$ is defined for all $x \in \mathbb{X}$ by

$$\mathcal{V}(x) := \min_{n\in\mathbb{N}^*}\{\tilde{V}_n(x)\} \quad \text{and } \mathcal{W}(x) := \frac{W(x)}{4}. \tag{5.4}$$

LEMMA 5.8. *For all $n \in \mathbb{N}^*$, for all $x_0 \in S_V[\frac{1}{n}, 10n], \mathcal{V}(x_0) = \min_{1\le p\le n}\tilde{V}_p(x_0)$. Furthermore, if $\zeta \in \partial_P\mathcal{V}(x_0)$, then*

$$\inf_{u\in U}\langle \zeta, f(x_0,u)\rangle \le -\frac{W(x_0)}{4}. \tag{5.5}$$

Proof. Let be given $n \in \mathbb{N}^*$ and $x_0 \in S_V[\frac{1}{n}, 10n]$. By Lemma 5.5, $V_{\alpha_n}(x_0) \in [\frac{1}{n} - \frac{1}{8n}, 10n]$. Hence, Lemma 5.7 implies that $\tilde{V}_n(x_0) = V_{\alpha_n}(x_0)$. On the other hand, for any $p \ge n$, by construction $\alpha_p \ge \alpha_n$ and then $V_{\alpha_p}(x_0) \ge V_{\alpha_n}(x_0)$. The same argument as above on Ψ_p leads to

$$\tilde{V}_p(x_0) = V_{\alpha_p}(x_0) \ge \tilde{V}_n(x_0) = V_{\alpha_n}(x_0).$$

Consequently, we have shown that $\mathcal{V}(x_0) = \min_{1\le p\le n}\tilde{V}_p(x_0)$. Now, if the minimum in the definition of $\mathcal{V}(x_0)$ is attained for $\tilde{V}_{n_0}(x_0)$ (with $1 \le n_0 \le n$) then

$$\zeta \in \partial_P\mathcal{V}(x_0) \Longrightarrow \zeta \in \partial_P\tilde{V}_{n_0}(x_0) = \Psi'(V_{\alpha_{n_0}}(x_0))\partial_P V_{\alpha_{n_0}}(x_0).$$

We now have to show the inequality (5.5).

First case. If $V(x_0) > 11n_0$ and $V_{\alpha_{n_0}(x_0)} \le 11n_0$, then there exists $\bar{y} \in x_0 + \bar{B}$ (because $\alpha_{n_0} \ge 11n_0$) such that $V_{\alpha_{n_0}}(x_0) = V(\bar{y}) + \alpha_{n_0}\|x_0 - \bar{y}\|^2$. Therefore, $\bar{y} \in S_V(11n_0)$ and $x_0 \in S_V(M_{n_0})$ by definition of M_{n_0}. By Lemmas 5.5 and 5.7 (iii), we obtain $V_{\alpha_{n_0}(x_0)} \ge 11n_0 - \frac{1}{8n_0}$ and $\tilde{V}_{n_0}(x_0) \ge 11n_0 + V(x_0)$. But Lemma 5.7(ii) implies that

$$\tilde{V}_n(x_0) = V_{\alpha_n}(x_0) \le V(x_0).$$

Since n_0 is optimal and $\tilde{V}_{n_0}(x_0) \ge V(x_0)$, we get a contradiction. Therefore, this case cannot appear.

Second case. If $V(x_0) > 11n_0$ and $V_{\alpha_{n_0}(x_0)} > 11n_0$, then Lemma 5.7(iii) implies $\tilde{V}_{n_0}(x_0) \geq 11n_0 + V(x_0)$, and we conclude as in the first case.

Third case. If $V(x_0) < \frac{1}{2n_0}$, then

$$V_{\alpha_{n_0}}(x_0) \leq V(x_0) < \frac{1}{2n_0} \Longrightarrow \tilde{V}_{n_0}(x_0) = V_{\alpha_{n_0}}(x_0) + \frac{1}{8n_0} \geq V(x_0).$$

But we proved that $\tilde{V}_n(x_0) = V_{\alpha_n}(x_0) \leq V(x_0)$, so the minimum is also attained for n; then we have (5.5) by Lemma 5.6.

Fourth case. If $x_0 \in S_V[\frac{1}{2n_0}, 11n_0]$, then we conclude by Lemmas 5.6 and 5.7 (iv). □

This last lemma shows that the minimum in the definition of $\mathcal{V}(x)$ is always attained for $x \neq 0$. Therefore, the function $\mathcal{V}$ is semiconcave outside the origin (by Lemma 5.2). On the other hand, $\mathcal{V}$ is continuous at the origin (because $0 \leq \mathcal{V} \leq V$) and satisfies the decrease condition by (5.5). Consequently, $\mathcal{V}$ provides a CLF, which proves the Theorem 2.

Acknowledgments. The author thanks Professor Francis Clarke for his comments and his suggestions, and two anonymous referees who read the paper with great care pointed out a large number of misprints and made interesting remarks.

REFERENCES

[1] G. Alberti, L. Ambrosio, and P. Cannarsa, *On the singularities of convex functions*, Manuscripta Math., 76 (1992), pp. 421–435.

[2] L. Ambrosio, P. Cannarsa, and H. M. Soner, *On the propagation of singularities of semiconvex functions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 597–616.

[3] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[4] J-P. Aubin and A. Cellina, *Differential Inclusions. Set-Valued Maps and Viability Theory*, Grundlehren Math. Wiss. 264, Springer-Verlag, Berlin, New York, 1984.

[5] A. Bacciotti, *Local Stabilizability of Nonlinear Control Systems*, World Scientific, River Edge, NJ, 1992.

[6] A. Bacciotti and L. Rosier, *Lyapunov and Lagrange stability: Inverse theorems for discontinuous systems*, Math. Control Signals Systems, 11 (1998), pp. 101–128.

[7] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.

[8] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory (Houghton, MI, 1982), Birkhäuser Boston, Boston, MA, 1983, pp. 181–191.

[9] P. Cannarsa and C. Sinestrari, *Convexity properties of the minimum time function*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 273–298.

[10] F.H. Clarke, Yu.S. Ledyaev, L. Rifford, and R.J. Stern, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 25–48.

[11] F.H. Clarke, Yu.S. Ledyaev, E.D. Sontag, and A.I. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[12] F.H. Clarke, Yu.S. Ledyaev, and R.J. Stern, *Asymptotic stability and smooth Lyapunov functions*, J. Differential Equations, 149 (1998), pp. 69–114.

[13] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski, *Qualitative properties of trajectories of control systems: A survey*, J. Dynam. Control Systems, 1 (1995), pp. 1–48.

[14] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math., 178, Springer-Verlag, New York, 1998.

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

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

[17] R.A. Freeman and P.V. Kokotović, *Robust Nonlinear Control Design.* State-Space and Lyapunov Techniques, Birkhäuser Boston, Boston, MA, 1996.

[18] H. Hermes, *Resonance, stabilizing feedback controls, and regularity of viscosity solutions of Hamilton-Jacobi-Bellman equations*, Math. Control Signals Systems, 9 (1996), pp. 59–72.

[19] S. N. Kružkov, *Generalized solutions of Hamilton-Jacobi equations of eikonal type* I, Mat. Sb. (N.S.), 27 (1975), pp. 406–446.

[20] P-L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman (Advanced Publishing Program), Boston, MA, 1982.

[21] L. Rifford, *Semiconcave control-Lyapunov functions and stabilizing feedbacks*, SIAM J. Control Optim., submitted.

[22] L. Rifford, *Problèmes de stabilisation en théorie du controle*, Ph.D. thesis, Université Claude Bernard Lyon I, Lyon, France, 2000.

[23] L. Rifford, *Stabilisation des systèmes globalement asymptotiquement commandables*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 211–216.

[24] L. Rosier, *Étude de quelques problèmes de stabilisation*, Ph.D. thesis, Ecole Normale Supérieure de Cachan, Bretagne, France, 1993.

[25] E.P. Ryan, *On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback*, SIAM J. Control Optim., 32 (1994), pp. 1597 –1604.

[26] E.D. Sontag, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.

[27] E.D. Sontag and H.J. Sussmann, *Remarks on continuous feedback*, in Proceedings of the IEEE Conference Decision and Control, Albuquerque, NM, 1980, pp. 916–921.

[28] E.D. Sontag and H.J. Sussmann, *Nonsmooth control-Lyapunov functions*, in Proceedings of the IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 2799–2805.

[29] E.D. Sontag and H.J. Sussmann, *General classes of control-Lyapunov functions*, in Stability Theory (Ascona, 1995), Birkhäuser, Basel, 1996, pp. 87–96.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1065–1081

STOCHASTIC LINEAR QUADRATIC REGULATORS WITH INDEFINITE CONTROL WEIGHT COSTS. II[*]

SHUPING CHEN[†] AND XUN YU ZHOU[‡]

Abstract. In part I of this paper [S. Chen, X. Li, and X. Zhou, *SIAM J. Control Optim.*, 36 (1998), pp. 1685–1702], an optimization model of stochastic linear quadratic regulators (LQRs) with *indefinite* control cost weighting matrices is proposed and studied. In this sequel, the problem of solving LQR models with system diffusions dependent on *both* state and control variables, which is left open in part I, is tackled. First, the solvability of the associated stochastic Riccati equations (SREs) is studied in the normal case (namely, all the state and control weighting matrices and the terminal matrix in the cost functional are nonnegative definite, with at least one positive definite), which in turn leads to an optimal state feedback control of the LQR problem. In the general indefinite case, the problem is decomposed into two optimal LQR problems, one with a forward dynamics and the other with a *backward* dynamics. The well-posedness and solvability of the original LQR problem are then obtained by solving these two subproblems, and an optimal control is explicitly constructed. Examples are presented to illustrate the results.

Key words. stochastic linear quadratic regulator, well-posedness, stochastic Riccati equation, backward stochastic differential equation

AMS subject classifications. 93E20, 49K45

PII. S0363012998346578

1. Introduction.

The optimal linear quadratic regulator (LQR) problem, initiated by Kalman [15], is one of the most important classes of optimal control problems. In the deterministic case, it is well known that when the control weighting matrix R in the cost function is positive definite, the problem can be solved elegantly via the Riccati equation; see, e.g., [2] for a thorough study of the Riccati approach. For the case when R is possibly singular, the deterministic LQR problem has also been extensively studied as the problem of singular LQ control; refer to, e.g., [11, 14, 22]. The stochastic LQR problem was first studied by Wonham [23] and has since been studied by many other researchers (see, for example, [4, 12, 3] and the references therein); its theory is widely believed to have been well developed and established. However, this belief is challenged by our recent finding [9] that a class of stochastic LQR problems with *indefinite* control weight cost is sensible and well-posed. (A related observation in the context of stochastic stability, namely, that the corresponding Lyapunov function exists even if R is indefinite, was made in [7].) To be precise, let us consider the following stochastic LQR problem:

$$\begin{aligned}&\text{Minimize} \quad J = E\left\{\int_0^T \frac{1}{2}[x(t)'Q(t)x(t) + u(t)'R(t)u(t)]dt + \frac{1}{2}x(T)'Hx(T)\right\}\\ &\text{subject to} \quad \begin{cases} dx(t) = [A(t)x(t) + B(t)u(t)]dt + [C(t)x(t) + D(t)u(t)]dW(t),\\ x(0) = x_0.\end{cases}\end{aligned}$$

[*] Received by the editors October 21, 1998; accepted for publication (in revised form) May 11, 2000; published electronically November 8, 2000.

http://www.siam.org/journals/sicon/39-4/34657.html

[†] Center for Mathematical Sciences, Zhejiang University, Hangzhou 310027, People's Republic of China (amaschen@dial.zju.edu.cn). The research of this author was supported partially by the National Natural Science Foundation of China and the State Education Commission of China.

[‡] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk). The research of this author was supported by the RGC earmarked grant CUHK 4125/97E.

Here $W(t)$ is a Brownian motion and the control variable $u(t)$ takes value in some Euclidean space. In part I of this paper [9], it is revealed that the above problem may be well-posed even when R, the so-called control cost weighting matrix, is *indefinite* (in particular, *negative definite*). This is quite contrary to the finite-dimensional, deterministic case[1] and has very deep reasons behind it, due to the uncertainty present in the system. On the other hand, this phenomenon may occur only when the diffusion coefficient of the system dynamics depends on the control (i.e., $D \neq 0$), meaning that controls *would* or *could* influence the uncertainty scale in the system. For a detailed discussion and many examples, see [9].

In [9], a *stochastic Riccati equation* (SRE) is introduced to solve the LQR problem. In general, when all the coefficients A, B, etc., are random processes, the SRE is a nonlinear *backward stochastic differential equation* (BSDE) of the Pardoux–Peng [20] type. If all the coefficients are deterministic, then the SRE reduces to the following (deterministic) ODE:

$$(1.1)\qquad \begin{cases} \dot{P}(t) = -(P(t)A(t) + A(t)'P(t) + C(t)'P(t)C(t) + Q(t)) \\ \qquad\qquad +(P(t)B(t) + C(t)'P(t)D(t))(R(t) + D(t)'P(t)D(t))^{-1} \\ \qquad\qquad\qquad \times(B(t)'P(t) + D(t)'P(t)C(t)), \\ P(T) = H, \\ K(t) \equiv R(t) + D(t)'P(t)D(t) > 0 \ \ \forall t \in [0, T]. \end{cases}$$

If the SRE admits a solution, then it is shown in [9] that an optimal control of the LQR problem can be constructed explicitly as a linear state feedback via the solution to the SRE. However, note that the SRE is in general different from the conventional Riccati equation (for the deterministic case) due to the presence of the term $(R + D'PD)^{-1}$, and is in fact substantially more difficult to handle. In [9], a special case when $C(t) \equiv 0$ is treated, and necessary and sufficient conditions of solvability of the SRE (1.1) are obtained. The general case when $C(t) \not\equiv 0$ remains an outstanding open problem.

This paper proceeds to tackle this open problem. We study two cases. The first one is the so-called *normal case*, namely, when Q, R, H are all nonnegative definite, with at least one positive definite. The solvability of the SRE (1.1) in this case is shown under the assumption that either R or $D'D$ is nonsingular, via a constructive proof. This case is interesting in its own right, particularly in view of many financial application problems where R is typically zero, whereas $D'D$ is nonsingular under the assumption of a complete market; refer to [26, 16]. Moreover, the normal case is useful in solving the second case, the *indefinite case* with indefinite control weights. In the indefinite case, we propose to decompose the original LQR problem into two LQR problems, one with a forward dynamics (which is solvable via the SRE) and the other with a *backward* dynamics. The stochastic maximum principle for systems with backward dynamics [21, 13] is applied to solve the second problem. This leads to the necessary conditions of the well-posedness and solvability of the original problem as well as an explicit construction of its optimal control.

The rest of the paper is organized as follows. In section 2 the formulation of optimal stochastic LQR models is given and some preliminary results are presented. In section 3 we study a stochastic LQR problem with nonhomogeneous dynamics. Sections 4 and 5 are devoted to the normal and indefinite cases, respectively. Finally, section 6 gives some concluding remarks.

[1] It was recently found that some infinite-dimensional deterministic LQR problems are well-posed, even though the control weight operator is negative; see [19, 8].

2. Problem formulation and preliminaries. We consider in this paper a stochastic optimal control problem. The system is governed by the following linear Ito's stochastic differential equation (SDE):

$$\left\{\begin{array}{l} dx(t) = [A(t)x(t) + B(t)u(t)]dt + [C(t)x(t) + D(t)u(t)]dW(t), \\ x(0) = x_0, \end{array}\right. \tag{2.1}$$

where $W(t)$ is a given one-dimensional standard Brownian motion on $[0, T]$ (with $W(0) = 0$); and $u(\cdot)$, an admissible control, is an R^m-valued, square integrable $\mathcal{F}_t$-adapted measurable process with

$$\mathcal{F}_t = \sigma\{W(s) : 0 \le s \le t\}. \tag{2.2}$$

The set of all such admissible controls is denoted by U_{ad}. Note that we assumed the Brownian motion to be one-dimensional just for simplicity; there is no essential difficulty in the analysis below for the multidimensional case.

For each $u(\cdot) \in U_{ad}$, the associated cost is

$$J(u(\cdot)) = E\left\{\int_0^T \frac{1}{2}[x(t)'Q(t)x(t) + u(t)'R(t)u(t)]dt + \frac{1}{2}x(T)'Hx(T)\right\}. \tag{2.3}$$

The solution $x(\cdot)$ of the SDE (2.1) is called the *response* of the control $u(\cdot) \in U_{ad}$, and $(x(\cdot), u(\cdot))$ is called an *admissible pair.* The objective of the optimal control problem is to minimize the cost function $J(u(\cdot))$ over all $u(\cdot) \in U_{ad}$. An admissible pair $(x^*(\cdot), u^*(\cdot))$ is called *optimal* if $u^*(\cdot)$ achieves the infimum of $J(u(\cdot))$. The optimization problem (2.1)–(2.3) is called *well-posed* if $\inf_{u(\cdot)\in U_{ad}} J(u(\cdot)) > -\infty$. Since the data (A, B, C, D, Q, R, H) completely determine the optimal LQR problem (2.1)–(2.3), sometimes for simplicity we may interchangeably use (A, B, C, D, Q, R, H) to denote the problem.

Notation. We use the following notation in this paper:

M' : the transpose of any vector or matrix M;
M^j : the jth entry of any vector M;
$|M|$: $= \sqrt{\sum_{i,j} m_{ij}^2}$ for any matrix or vector $M = (m_{ij})$;
S^n : the space of all $n \times n$ symmetric matrices;
S_+^n : the subspace of all nonnegative definite matrices of S^n;
$\hat{S}_+^n$: the subspace of all positive definite matrices of S^n;
$C([0, T]; X)$: the Banach space of X-valued continuous functions on $[0, T]$ endowed with the maximum norm $\|\cdot\|$ for a given Hilbert space X;
ρ_x : the gradient or Jacobian of a function ρ with respect to the variable x;
ρ_{xx} : the Hessian of a scalar function ρ with respect to the variable x.

Given a probability space $(\Omega, \mathcal{F}, P)$ with a filtration $\{\mathcal{F}_t : a \le t \le b\}$ $(-\infty \le a < b \le +\infty)$, a Hilbert space X with the norm $\|\cdot\|_X$, and p $(1 \le p \le +\infty)$, define the Banach space $L^p_{\mathcal{F}}(a, b; X) = \{\phi(\cdot) = \{\phi(t, \omega) : a \le t \le b\} |\ \phi(\cdot)$ is an $\mathcal{F}_t$-adapted,

X-valued measurable process on $[a,b]$, and $E\int_a^b \parallel \phi(t,\omega) \parallel_X^p \, dt < +\infty\}$, with the norm

$$\parallel \phi(\cdot) \parallel_{\mathcal{F},p} = \left(E\int_a^b \parallel \phi(t,\omega) \parallel_X^p \, dt \right)^{\frac{1}{p}}.$$

In the rest of this paper, we shall employ the usual convention of suppressing the ω-dependence of all random functions. Sometimes we even write A for a (deterministic or stochastic) process $A(t)$, omitting the variable t, whenever no confusion arises. Under this convention, when $A \in C([0,T];S^n)$, $A \geq (>) 0$ means $A(t) \geq (>) 0$ $\forall t \in [0,T]$.

The following basic assumption will be in force throughout this paper.

(A) The data appearing in the LQR problem satisfy

$$\begin{array}{l} A, C \in L^\infty(0,T;R^{n\times n}), \\ B, D \in L^\infty(0,T;R^{n\times m}), \\ Q \in L^\infty(0,T;S^n_+), \\ R \in L^\infty(0,T;S^m), \\ H \in S^n_+. \end{array}$$

Note that we restrict Q and H, but *not* R, to be nonnegative definite.

Let us consider the SRE (1.1). A function $P \in C([0,T];S^n)$ is called a solution of (1.1) if it satisfies *all* the constraints in (1.1) (in particular, the third inequality constraint). The following result is taken from [9, Theorem 3.2].

THEOREM 2.1. *If the SRE* (1.1) *admits a solution P, then the stochastic LQR problem (A,B,C,D,Q,R,H) is well-posed. In addition, the feedback control*

$$u^*(t,x) = -K^{-1}(t)[B(t)'P(t) + D(t)'P(t)C(t)]x, \tag{2.4}$$

which results in a unique solution of the state equation (2.1), *is optimal with the optimal value*

$$\inf_{u(\cdot)\in U_{ad}} J(u(\cdot)) = \frac{1}{2}x_0'P(0)x_0. \tag{2.5}$$

Remark 2.1. Theorem 2.1 implies that if the SRE (1.1) is solvable, then the original LQR problem is completely solved. In general we call the problem (A,B,C,D,Q,R,H) solvable via SRE if (1.1) admits a solution P and (2.4) gives an optimal feedback control. Note that, completely different from the deterministic case, the solvability of the problem (A,B,C,D,Q,R,H) does not necessarily imply its solvability via SRE, or the solvability of (1.1).

While the existence of the SRE (1.1) remains a hard problem, the uniqueness can be proved in a rather routine way, which is presented in the following theorem for the readers' convenience.

THEOREM 2.2. *There is at most one solution to* (1.1).

Proof. Suppose $P_1, P_2 \in C([0,T];S^n)$ are two solutions to (1.1). Consider the following LQR problem:

$$\begin{array}{ll} \text{Minimize} & J(s,y;u(\cdot)) = E\left\{ \int_s^T \frac{1}{2}[x(t)'Q(t)x(t) + u(t)'R(t)u(t)]dt + \frac{1}{2}x(T)'Hx(T) \right\} \\ \text{subject to} & \left\{ \begin{array}{l} dx(t) = [A(t)x(t) + B(t)u(t)]dt + [C(t)x(t) + D(t)u(t)]dW(t), \\ x(s) = y, \end{array} \right. \end{array}$$

where $(s,y)\in[0,T]\times R^n$. Theorem 2.1 implies that both $\frac{1}{2}y'P_1(s)y$ and $\frac{1}{2}y'P_2(s)y$ are the minimum value of the cost functional $J(s,y;u(\cdot))$, and hence they must be identical. The desired result follows then from the arbitrariness of (s,y). □

3. Nonhomogeneous LQR problem.

In this section we consider the optimal LQR problem where the cost is the same as that of (2.3) while there are *random* nonhomogeneous terms in the dynamics:

$$\begin{cases} dx(t)=[A(t)x(t)+B(t)u(t)+f(t)]dt+[C(t)x(t)+D(t)u(t)+g(t)]dW(t),\\ x(0)=x_0, \end{cases} \tag{3.1}$$

with $f,g\in L^2_{\mathcal{F}}(0,T;R^n)$. While the study of this kind of system is interesting on its own, it will be useful later in section 5.

Introduce an equation

$$\begin{cases} d\phi(t)=-[\bar{A}(t)'\phi(t)+\bar{C}(t)'\lambda(t)+P(t)f(t)+\bar{C}(t)'P(t)g(t)]dt+\lambda(t)dW(t),\\ \phi(T)=0, \end{cases} \tag{3.2}$$

where

$$\bar{A}=A-BK^{-1}L,\quad \bar{C}=C-DK^{-1}L,\quad L=B'P+D'PC,\quad K=R+D'PD, \tag{3.3}$$

and $P(\cdot)$ is the solution to the SRE (1.1) (assuming that (1.1) is solvable). Equation (3.2) is a BSDE whose solution is a *pair* of processes (ϕ,λ). If (1.1) admits a solution $P\in C([0,T];S^n)$, then it is well known that (3.2) must have an $\mathcal{F}_t$-adapted solution $(\phi,\lambda)\in L^2_{\mathcal{F}}(0,T;R^n)\times L^2_{\mathcal{F}}(0,T;R^n)$. See [5, 6] for the origin of the linear BSDE and [25, Chapter 7] for a systematic account of the latest BSDE theory.

THEOREM 3.1. *If equations* (1.1) *and* (3.2) *admit solutions* $P\in C([0,T];S^n)$ *and* $(\phi,\lambda)\in L^2_{\mathcal{F}}(0,T;R^n)\times L^2_{\mathcal{F}}(0,T;R^n)$, *respectively, then the stochastic LQR problem consisting of* (3.1) *and* (2.3) *has an optimal feedback control*

$$u^*(t,x)=-K^{-1}(t)[L(t)x+h(t)], \tag{3.4}$$

where $h(t)=B(t)'\phi(t)+D(t)'\lambda(t)+D(t)'P(t)g(t)$. *Moreover, the optimal value is*

$$\begin{aligned}\inf_{u(\cdot)\in U_{ad}} J(u(\cdot))=\frac{1}{2}E\int_0^T&[-h'K^{-1}h+g'Pg+2\phi'f+2\lambda'g](t)dt\\ &+\frac{1}{2}x_0'P(0)x_0+\phi(0)'x_0.\end{aligned} \tag{3.5}$$

Proof. Applying Ito's formula, we get

$$\begin{aligned}\frac{1}{2}d(x'Px)=&\frac{1}{2}[u'D'PDu+x'(-Q+L'K^{-1}L)x\\ &+2u'Lx+2x'(Pf+C'Pg)+2u'D'Pg+g'Pg](t)dt\\ &+\frac{1}{2}\{\ldots\}dW(t)\end{aligned} \tag{3.6}$$

and

$$\begin{aligned}d(x'\phi)=\ &[x'(L'K^{-1}B'\phi-Pf-C'Pg+L'K^{-1}D'Pg+L'K^{-1}D'\lambda)\\ &+u'B'\phi+u'D'\lambda+\phi'f+\lambda'g](t)dt+\{\ldots\}dW(t).\end{aligned} \tag{3.7}$$

Integrating both (3.6) and (3.7) from 0 to T, taking expectations, adding them together, and noting (2.3), one obtains

(3.8)
$$
\begin{aligned}
J(u(\cdot)) &\equiv J(u(\cdot)) + E\left\{\int_0^T d\left(\frac{1}{2}x'Px + x'\phi\right) - \left(\frac{1}{2}x'Px + x'\phi\right)\Big|_0^T\right\} \\
&= \frac{1}{2}E\int_0^T [u'Ku + 2u'(B'Px + D'PCx + D'Pg + B'\phi + D'\lambda) + x'L'K^{-1}Lx \\
&\qquad +2x'L'K^{-1}B'\phi + 2x'L'K^{-1}D'(Pg+\lambda) + g'Pg + 2\phi'f + 2\lambda'g](t)dt \\
&\qquad +\frac{1}{2}x_0'P(0)x_0 + \phi(0)'x_0 \\
&= \frac{1}{2}E\int_0^T \{[u + K^{-1}(Lx+h)]'K[u + K^{-1}(Lx+h)] - h'K^{-1}h \\
&\qquad\qquad +g'Pg + 2\phi'f + 2\lambda'g\}(t)dt + \frac{1}{2}x_0'P(0)x_0 + \phi(0)'x_0.
\end{aligned}
$$

It follows immediately that the optimal feedback control is given by (3.4) and the optimal value by (3.5), provided that the corresponding state exists under such a feedback control. To see if the latter assertion is true, notice that under (3.4), the system (3.1) reduces to

$$
(3.9)\qquad \begin{cases} dx(t) = [\bar{A}(t)x(t) + \bar{f}(t)]dt + [\bar{C}(t)x(t) + \bar{g}(t)]dW(t), \\ x(0) = x_0, \end{cases}
$$

where $\bar{A}$ and $\bar{C}$ are given by (3.3), $\bar{f} = f - BK^{-1}h$, and $\bar{g} = g - DK^{-1}h$. Equation (3.9) is a nonhomogeneous linear SDE. Since $P \in C([0,T];S^n)$, $K^{-1} \in C([0,T];S^n_+)$, and $(\phi,\lambda) \in L^2_{\mathcal{F}}(0,T;R^n) \times L^2_{\mathcal{F}}(0,T;R^n)$, the coefficients $\bar{A}(t)$ and $\bar{C}(t)$ are uniformly bounded functions and the nonhomogeneous terms $\bar{f},\bar{g} \in L^2_{\mathcal{F}}(0,T;R^n)$. Hence (3.9) admits one and only one solution by standard SDE theory. This completes the proof. □

4. Solvability of SRE in normal case. We say that the (homogeneous) LQR problem (2.1)–(2.3) (or (A,B,C,D,Q,R,H) or even (Q,R,H)) is in the *normal case* if

$$
(4.1)\qquad Q \geq 0, \quad R \geq 0, \quad H \geq 0,
$$

with *at least* one *strictly* positive definite at *any* time. It is seen directly from the cost functional (2.3) that the LQR problem must be well-posed in the normal case (with a nonnegative infimum). As we mentioned, the well-posedness of the LQR problem in general does not necessarily lead to the solvability of the SRE (1.1). However, we will show that in the normal case there does exist a solution to the SRE (1.1), under the additional assumption that either R or $D'D$ is positive definite, and therefore the LQR problem is solvable via SRE. Note that, once again, the existence result to be presented in this section is not only interesting for its own sake but also useful in dealing with the case with indefinite control weights in section 5.

First we give a technical lemma.

LEMMA 4.1. *Given $F,G \in L^\infty(0,T;R^{n\times n})$, $Q \in L^\infty(0,T;S^n)$, and $H \in S^n$, the linear matrix equation*

$$
(4.2)\qquad \begin{cases} \dot{X}(t) + X(t)F(t) + F(t)'X(t) + G(t)'X(t)G(t) + Q(t) = 0, \\ X(T) = H, \end{cases}
$$

admits a unique solution $X \in C([0,T]; S^n)$. *Moreover,* $X(t)$ *is nonnegative definite for all* $t \in [0,T]$ *if* $Q(t)$ *and* H *are nonnegative definite for all* $t \in [0,T]$, *and* $X(t)$ *is positive definite for all* $t \in [0,T]$ *if either* $Q(t)$ *or* H *is positive definite for all* $t \in [0,T]$.

Proof. Since (4.2) is linear with bounded coefficients, the existence and uniqueness are clear. Now let Φ be the solution of the following matrix stochastic differential equation:

$$\begin{cases} d\Phi(t) = F(t)\Phi(t)dt + G(t)\Phi(t)dW(t), \\ \Phi(0) = I. \end{cases} \tag{4.3}$$

Clearly this equation admits a unique solution. Moreover, $\Phi(t)$ is invertible for all $t \in [0,T]$, P-a.s. Indeed, by Ito's formula, it is easy to show that the solution $\Psi(\cdot)$ to the following equation:

$$\begin{cases} d\Psi(t) = -\Psi(t)[F(t) - G^2(t)]dt - \Psi(t)G(t)dW(t), \\ \Phi(0) = I, \end{cases} \tag{4.4}$$

satisfies $\Psi(t)\Phi(t) = I$ for all $t \in [0,T]$, P-a.s. Now applying Ito's formula again, we obtain

$$d(\Phi(t)'X(t)\Phi(t)) = -\Phi(t)'Q(t)\Phi(t)dt + \Phi(t)'[G(t)'X(t) + X(t)G(t)]\Phi(t)dW(t).$$

Hence, noting that $X(t)$ is deterministic, we have

$$X(t) = E\left\{\Phi^{-1}(t)'\Phi(T)'H\Phi(T)\Phi^{-1}(t) + \int_t^T [\Phi^{-1}(t)'\Phi'(r)Q(r)\Phi(r)\Phi^{-1}(t)]dr\right\}. \tag{4.5}$$

This proves the second part of the lemma. □

THEOREM 4.1. *In the normal case, the SRE* (1.1) *admits a solution* $P \in C([0,T]; S^n_+)$ *if either of the following two conditions holds:*

(a) $R(t) > 0$ $\forall t \in [0,T]$*;*

(b) $R(t)$ *is singular, and* $D(t)'D(t) > 0$ $\forall t \in [0,T]$.

Proof. First, (1.1) can be rewritten as

$$\begin{cases} \dot{P} + PF + F'P + G'PG + M'RM + Q = 0, \\ P(T) = H, \end{cases} \tag{4.6}$$

where

$$F = A - BM, \;\; G = C - DM, \;\; M = K^{-1}L \equiv (R + D'PD)^{-1}(B'P + D'PC). \tag{4.7}$$

We now introduce an iterative scheme to construct the solution to (4.6). Initially, let

$$P_0 \equiv I, \; M_0 = (R + D'P_0D)^{-1}(B'P_0 + D'P_0C), \qquad F_0 = A - BM_0, \; G_0 = C - DM_0. \tag{4.8}$$

Note that $R + D'P_0D = R + D'D$ is invertible at any time under either of the cases (a) or (b). For $i = 0, 1, 2, \ldots$, let P_{i+1} be the solution to the following linear matrix equation:

$$\begin{cases} \dot{P}_{i+1} + P_{i+1}F_i + F_i'P_{i+1} + G_i'P_{i+1}G_i + M_i'RM_i + Q = 0, \\ P_{i+1}(T) = H, \end{cases} \tag{4.9}$$

and set

$$M_{i+1}=(R+D'P_{i+1}D)^{-1}(B'P_{i+1}+D'P_{i+1}C),\ F_{i+1}=A-BM_{i+1},\ G_{i+1}=C-DM_{i+1}. \tag{4.10}$$

Note that in the normal case, (4.9) must admit a solution $P_{i+1}\in C([0,T];S^n_+)$ in view of Lemma 4.1. In case (a), $R+D'P_{i+1}D$ always has an inverse. In case (b), either $Q(t)$ or H is positive definite for all $t\in[0,T]$, resulting in $P_{i+1}>0$ by Lemma 4.1. Therefore, in either case (a) or (b), the inverse in defining M_{i+1} in (4.10) always exists. Setting $\Delta_{i+1}=P_{i+1}-P_i$, one can show that Δ_{i+1} satisfies the following:

$$\begin{cases}\dot\Delta_{i+1}+\Delta_{i+1}F_i+F_i'\Delta_{i+1}+G_i'\Delta_{i+1}G_i-(M_i-M_{i-1})'(R+D'P_iD)(M_i-M_{i-1})=0,\\ \Delta_{i+1}(T)=0.\end{cases} \tag{4.11}$$

Once again, by Lemma 4.1 we conclude that $\Delta_{i+1}\le 0$. Hence $\{P_i(\cdot)\}$ is a decreasing sequence in $C([0,T];S^n_+)$ and therefore has a limit (with respect to the max norm of $C([0,T];S^n_+)$), denoted by $P(\cdot)$. Clearly $P(\cdot)$ is the solution to (4.6), and hence to (1.1). □

Equations (4.8)–(4.10) actually constitute a numerical algorithm to compute the solution of the SRE (1.1). The following proposition presents an estimate for the convergence speed of this algorithm.

PROPOSITION 4.1. *Let the assumptions of Theorem* 4.1 *hold. Let*

$$\{P_i\}\subset C([0,T];S^n_+)$$

be constructed by the algorithm (4.8)–(4.10) *and* $P\in C([0,T];S^n_+)$ *be the solution to the Riccati equation* (1.1). *Then*

$$|P_i(t)-P(t)|\le c\sum_{j=i}^{\infty}\frac{(c')^{j-2}}{(j-2)!}(T-t)^{j-2},\qquad i=2,3,\dots, \tag{4.12}$$

where c *and* c' *are constants that depend only on the coefficients of* (1.1).

Proof. Note that $\Delta_{i+1}\equiv P_{i+1}-P_i$ satisfies (4.11). Set $K_i=R+D'P_iD$; then

$$M_i-M_{i-1}=K_i^{-1}(B'\Delta_i+D'\Delta_iC)-K_{i-1}^{-1}D'\Delta_iDK_i^{-1}(B'P_{i-1}+D'P_{i-1}C). \tag{4.13}$$

Putting this into (4.11) and taking integration, we obtain

$$\begin{aligned}\Delta_{i+1}(t)=\int_t^T&[\Delta_{i+1}F_i+F_i'\Delta_{i+1}+G_i'\Delta_{i+1}G_i\\ &-(\Delta_iB+C'\Delta_iD)K_i^{-1}(B'\Delta_i+D'\Delta_iC)\\ &+2(\Delta_iB+C'\Delta_iD)K_{i-1}^{-1}D'\Delta_iDK_i^{-1}(B'P_{i-1}+D'P_{i-1}C)\\ &-(P_{i-1}B+C'P_{i-1}D)K_i^{-1}D'\Delta_iDK_{i-1}^{-1}K_iK_{i-1}^{-1}D'\Delta_iDK_i^{-1}\\ &\times(B'P_{i-1}+D'P_{i-1}C)](s)ds.\end{aligned} \tag{4.14}$$

By the proof of Theorem 4.1, the sequences $\{|P_i|\},\{|K_i|\}$, and $\{|K_i^{-1}|\}$ are uniformly bounded, and as a consequence so are the other sequences involved in (4.14). Hence,

$$|\Delta_{i+1}(t)|\le c\int_t^T[|\Delta_{i+1}(s)|+|\Delta_i(s)|]ds. \tag{4.15}$$

Denote $v_i(t) = \int_t^T |\Delta_{i+1}(s)|ds$. Then (4.15) reads

$$\dot{v}_i(t) + cv_i(t) + cv_{i-1}(t) \geq 0,$$

which implies

$$v_i(t) \leq ce^{cT} \int_t^T v_{i-1}(s)ds \equiv c' \int_t^T v_{i-1}(s)ds.$$

By induction, we deduce that

$$v_i(t) \leq \frac{(c')^{i-1}}{(i-1)!}(T-t)^{i-1}v_1(0).$$

It then follows from (4.15) that

$$|\Delta_{i+1}(t)| \leq c\left\{\frac{(c')^{i-1}}{(i-1)!}(T-t)^{i-1} + \frac{(c')^{i-2}}{(i-2)!}(T-t)^{i-2}\right\}v_1(0). \tag{4.16}$$

This easily yields (4.12). □

Remark 4.1. Theorem 4.1 gives sufficient conditions for an LQR problem to be solved completely via the solution to the SRE in the normal case, with an optimal linear feedback control explicitly given by (2.4). It should be noted that the assumptions of Theorem 4.1 do not exclude the possibility that the control weights $R(t)$ might be singular or even identically zero. In this case $D(t)'D(t)$ needs to be positive definite in order for the problem to be well-posed (see condition (b) of Theorem 4.1). The term $D(t)'D(t)$ measures how the control would influence the uncertainty, leading to what we call the *uncertainty cost* or *risk cost*. It is this indirect cost that makes the problem meaningful despite the fact that there is no direct control cost. Such a situation typically occurs in financial applications. For example, in a mean-variance portfolio selection problem, the risk (measured by the covariance matrix) is the only factor that concerns the investor, which can be translated into an uncertainty cost for the investor to make a decision. See Zhou and Li [26] for a detailed investigation of such a model and Kohlmann and Zhou [16] for a related Black–Scholes model.

5. LQR control with indefinite control weights.

Now we consider the problem where the control weighting matrix R is possibly indefinite. In view of Theorem 2.1, the problem boils down to the one of finding solutions to the SRE (1.1). However, the solvability of (1.1) in general remains an open problem at the moment. In this section, we are going to derive some necessary conditions for the LQR problem to be solvable without directly dealing with the SRE (1.1). The problem that we are able to handle is one that can be decomposed into two subproblems: one can be solved via SRE, and the other cannot. Let us make it more precise. Since $R(t)$ is symmetric, it may be assumed to be a diagonal matrix for simplicity of exposition. Write $R(t)$ as

$$R(t) = \begin{pmatrix} R_1(t) & 0 \\ 0 & R_2(t) \end{pmatrix}, \tag{5.1}$$

where the sizes of $R_1(t)$ and $R_2(t)$ are m_1 and m_2 (independent of t), respectively, with $m_1 > 0$ and $m_1 + m_2 = m$. We then divide the matrices B and D accordingly, namely,

$$B(t) = (B_1(t), B_2(t)), \qquad D(t) = (D_1(t), D_2(t)). \tag{5.2}$$

Suppose that the LQR problem $(A, B_1, C, D_1, Q, R_1, H)$ is solvable via SRE. A typical case is when $(A, B_1, C, D_1, Q, R_1, H)$ satisfies the assumptions of Theorem 4.1 or, in particular, when $R_1 > 0$; see section 4.

Using the above decomposition, we can rewrite system (2.1) and cost functional (2.3) as

$$\begin{cases} dx(t) = [A(t)x(t) + B_1(t)u_1(t) + B_2(t)u_2(t)]dt \\ \qquad\qquad +[C(t)x(t) + D_1(t)u_1(t) + D_2(t)u_2(t)]dW(t), \\ x(0) = x_0, \end{cases} \tag{5.3}$$

and

$$\begin{aligned} J(u(\cdot)) = E\Bigg\{ \int_0^T \frac{1}{2}[x(t)'Q(t)x(t) + u_1(t)'R_1(t)u_1(t) \\ +u_2(t)'R_2(t)u_2(t)]dt + \frac{1}{2}x(T)'Hx(T) \Bigg\}. \end{aligned} \tag{5.4}$$

The main idea of our approach is to first fix $u_2(\cdot)$, viewing $f(t) \equiv B_2(t)u_2(t)$ and $g(t) \equiv D_2(t)u_2(t)$ as nonhomogeneous terms (which are random since $u_2(\cdot)$ is random in general) for an LQR problem where the dynamics is (5.3) with the control $u_1(\cdot)$ and the cost is (5.4). After this problem (parameterized by $u_2(\cdot)$) is solved, we solve the remaining control problem in terms of $u_2(\cdot)$. Namely, we carry out a two-step procedure, which is made precise by the following lemma.

LEMMA 5.1. *We have the following:*

(a) $\inf_{u(\cdot)} J(u(\cdot)) \equiv \inf_{(u_1(\cdot),u_2(\cdot))} J(u_1(\cdot), u_2(\cdot)) = \inf_{u_2(\cdot)} \inf_{u_1(\cdot)} J(u_1(\cdot), u_2(\cdot))$.

(b) $u^*(\cdot) = (u_1^*(\cdot), u_2^*(\cdot))$ *minimizes* $J(u(\cdot))$ *if and only if* $u_2^*(\cdot)$ *minimizes* $\inf_{u_1(\cdot)} J(u_1(\cdot), u_2(\cdot))$ *and* $u_1^*(\cdot)$ *minimizes* $J(u_1(\cdot), u_2^*(\cdot))$.

Proof. (a) The conclusion is straightforward by the definition of infimum.

(b) *Necessity.* If $u_2^*(\cdot)$ does not minimize $\inf_{u_1(\cdot)} J(u_1(\cdot), u_2(\cdot))$, then there exists $\bar{u}_2(\cdot)$ so that

$$\inf_{u_1(\cdot)} J(u_1(\cdot), \bar{u}_2(\cdot)) < \inf_{u_1(\cdot)} J(u_1(\cdot), u_2^*(\cdot)).$$

By (a), we have

$$\begin{aligned} J(u_1^*(\cdot), u_2^*(\cdot)) &= \inf_{u_2(\cdot)} \inf_{u_1(\cdot)} J(u_1(\cdot), u_2(\cdot)) \\ &\le \inf_{u_1(\cdot)} J(u_1(\cdot), \bar{u}_2(\cdot)) \\ &< \inf_{u_1(\cdot)} J(u_1(\cdot), u_2^*(\cdot)) \\ &\le J(u_1^*(\cdot), u_2^*(\cdot)), \end{aligned}$$

which leads to a contradiction. This proves the first assertion. As a consequence,

$$\inf_{u_1(\cdot)} J(u_1(\cdot), u_2^*(\cdot)) = \inf_{u_2(\cdot)} \inf_{u_1(\cdot)} J(u_1(\cdot), u_2(\cdot)) = J(u_1^*(\cdot), u_2^*(\cdot)).$$

This shows the second assertion.

Sufficiency. If $u^*(\cdot) = (u_1^*(\cdot), u_2^*(\cdot))$ does not minimize $J(u(\cdot))$, then there is $\bar{u}(\cdot) = (\bar{u}_1(\cdot), \bar{u}_2(\cdot)) \in U_{ad}$ such that

$$J(\bar{u}_1(\cdot), \bar{u}_2(\cdot)) < J(u_1^*(\cdot), u_2^*(\cdot)) = \inf_{u_1(\cdot)} J(u_1(\cdot), u_2^*(\cdot)).$$

On the other hand,

$$J(\bar{u}_1(\cdot),\bar{u}_2(\cdot)) \geq \inf_{u_2(\cdot)}\inf_{u_1(\cdot)} J(u_1(\cdot),u_2(\cdot)) = \inf_{u_1(\cdot)} J(u_1(\cdot),u_2^*(\cdot)).$$

This leads to a contradiction. □

In the first step, we consider the parameterized LQR problem (5.3)–(5.4), where the state and control variables are $x(\cdot)$ and $u_1(\cdot)$, respectively, with $u_2(\cdot)$ regarded as a parameter. This problem is one with nonhomogeneous terms which has been studied in section 3. Equations (1.1) and (3.2) specialize to the following equations in the present case:

$$(5.5)\qquad \begin{cases} \dot{P} + PA + A'P + C'PC + Q \\ \qquad -(PB_1 + C'PD_1)(R_1 + D_1'PD_1)^{-1}(B_1'P + D_1'PC) = 0, \\ P(T) = H, \\ K_1(t) \equiv R_1(t) + D_1'(t)P(t)D_1(t) > 0 \ \forall t \in [0,T], \end{cases}$$

and

$$(5.6)\qquad \begin{cases} d\phi = -[A_1'\phi + C_1'\lambda + (B_2'P + D_2'PC)'u_2]dt + \lambda dW(t), \\ \phi(T) = 0, \end{cases}$$

where P is the solution to (5.5), and

$$(5.7)\quad A_1 = A - B_1K_1^{-1}(B_1'P + D_1'PC), \quad C_1 = C - D_1K_1^{-1}(B_1'P + D_1'PC).$$

Owing to the way we decompose the matrix R into R_1 and R_2, the SRE (5.5) admits a solution $P \in C([0,T];S^n)$. Then by Theorem 3.1, an optimal feedback control of the problem (5.3)–(5.4) (with $u_2(\cdot)$ regarded as a parameter) is

$$(5.8)\qquad \begin{aligned} u_1(t,x) = -K_1^{-1}(t)\{&[B_1(t)'P(t) + D_1(t)'P(t)C(t)]x + B_1(t)'\phi(t) \\ &+D_1(t)'\lambda(t) + D_1(t)'P(t)D_2(t)u_2(t)\}, \end{aligned}$$

with the optimal value (which is a functional of $u_2(\cdot)$)

(5.9)

$$\begin{aligned} J_2(u_2(\cdot)) &\equiv \inf_{u_1(\cdot)} J(u_1(\cdot),u_2(\cdot)) \\ &= \frac{1}{2}E\int_0^T \{u_2'[D_2'(P - PD_1K_1^{-1}D_1'P)D_2 + R_2]u_2 \\ &\qquad +2\phi'(B_2 - B_1K_1^{-1}D_1'PD_2)u_2 + 2\lambda'(D_2 - D_1K_1^{-1}D_1'PD_2)u_2 \\ &\qquad -2\phi'B_1K_1^{-1}D_1'\lambda - \lambda'D_1K_1^{-1}D_1'\lambda - \phi'B_1K_1^{-1}B_1'\phi\}(t)dt \\ &\quad +\frac{1}{2}x_0'P(0)x_0 + \phi(0)'x_0 \\ &= \frac{1}{2}E\int_0^T \{u_2'\bar{R}u_2 + 2\phi'\bar{B}u_2 + 2\lambda'\bar{D}u_2 - (B_1'\phi + D_1'\lambda)'K_1^{-1}(B_1'\phi + D_1'\lambda)\}(t)dt \\ &\quad +\frac{1}{2}x_0'P(0)x_0 + \phi(0)'x_0, \end{aligned}$$

where

$$(5.10)\qquad \begin{cases} \bar{R} = D_2'(P - PD_1K_1^{-1}D_1'P)D_2 + R_2, \\ \bar{B} = B_2 - B_1K_1^{-1}D_1'PD_2, \\ \bar{D} = D_2 - D_1K_1^{-1}D_1'PD_2. \end{cases}$$

The next step is then to choose $u_2(\cdot)$ so as to minimize (5.9) subject to (5.6). We see that this is an LQR problem where the system dynamics (5.6) is a linear *backward* stochastic differential equation (with $(\phi(\cdot), \lambda(\cdot))$ as a *pair* of state variables) and the cost functional (5.9) involves an *initial* cost. Let us call such a problem a *backward LQR problem.* Our objective now is to solve this problem.

Peng [21] derived a set of necessary conditions (maximum principle) for optimal controls of nonlinear forward-backward systems, which include the backward LQR problem as a special space. However, the maximum principle presented in [21] is in a local form. Recently Dokuchaev and Zhou [13] proved the global maximum principle. For the readers' convenience, let us first state the maximum principle for nonlinear backward systems and then specialize to the backward LQR model.

Given $\xi \in R^n$,

$$\text{minimize } J(u(\cdot)) = E\left[\int_0^T l(t,x(t),z(t),u(t))dt + h(x(0))\right] \tag{5.11}$$

$$\text{subject to } \begin{cases} dx(t) = f(t,x(t),z(t),u(t))dt + z(t)dW(t), \\ x(T) = \xi. \end{cases} \tag{5.12}$$

The set of admissible controls is defined as $BU_{ad} = L^2_{\mathcal{F}}(0,T;R^m)$. We assume that

(B) f, l, h are continuous in their argument(s) and continuously differentiable in (x,z,u). Moreover, the derivatives of f are uniformly bounded, the derivatives of l are bounded by $C(1+|x|+|z|+|u|)$, and the derivative of h is bounded by $C(1+|x|)$.

Under the above assumption, given $u(\cdot) \in BU_{ad}$, the nonlinear backward SDE (5.12) admits a unique $\mathcal{F}_t$-adapted solution pair $(x(\cdot), z(\cdot))$ (see Pardoux and Peng [20]) and the cost (5.11) is well defined.

Define a Hamiltonian

$$H(t,x,z,u,q) = q'f(t,x,z,u) - l(t,x,z,u) \tag{5.13}$$

for $(t,x,z,u,q) \in [0,T] \times R^n \times R^n \times R^m \times R^n$.

THEOREM 5.1 (see [13, Theorem 3.1]). *If $(x^*(\cdot), z^*(\cdot), u^*(\cdot))$ is optimal for problem* (5.11)–(5.12), *then it must satisfy*

(5.14)
$$H(t,x^*(t),z^*(t),u^*(t),q(t)) = \max_{u\in R^m} H(t,x^*(t),z^*(t),u,q(t)), \quad P\text{-a.s.}, \quad \text{a.e. } t \in [0,T],$$

where $q(\cdot)$ is the solution of the adjoint equation

(5.15)
$$\begin{cases} dq(t) = -H_x(t,x^*(t),z^*(t),u^*(t),q(t))dt - H_z(t,x^*(t),z^*(t),u^*(t),q(t))dW(t), \\ q(0) = -h_x(x^*(0)). \end{cases}$$

We now apply the above result to the backward LQR problem consisting of (5.6) and (5.9). If $(\phi^*(\cdot), \lambda^*(\cdot), u_2^*(\cdot))$ is an optimal triple, then the corresponding adjoint equation (5.15) reduces to

$$\begin{cases} dq(t) = [A_1q + B_1u_2^* - B_1K_1^{-1}B_1'\phi^* - B_1K_1^{-1}D_1'\lambda^*](t)dt \\ \qquad\qquad +[C_1q + D_1u_2^* - D_1K_1^{-1}B_1'\phi^* - D_1K_1^{-1}D_1'\lambda^*](t)dW(t), \\ q(0) = -x_0. \end{cases} \tag{5.16}$$

Theorem 5.2. *If* $(\phi^*(\cdot), \lambda^*(\cdot), u_2^*(\cdot))$ *is optimal for* (5.6) *and* (5.9), *then*

$$\bar{R}(t) \equiv D_2(t)'[P(t) - P(t)D_1(t)K_1^{-1}(t)D_1(t)'P(t)]D_2(t) + R_2(t) \geq 0 \;\; \forall t \in [0,T]. \tag{5.17}$$

Moreover, if $\bar{R}(t) > 0 \; \forall t \in [0,T]$, *then the optimal control* $u_2^*(\cdot)$ *must satisfy*

$$u_2^*(t) = -\bar{R}^{-1}(t)[(B_2'P + D_2'PC_1)q + B_1'\phi^* + D_1'\lambda^*](t). \tag{5.18}$$

Proof. The derivative of the Hamiltonian in u in the present LQR case is

$$H_u(t,\phi,\lambda,u_2,q) = -\bar{R}(t)u_2 - [(B_2'P + D_2'PC_1)(t)q + B_1(t)'\phi + D_1(t)'\lambda]. \tag{5.19}$$

Therefore the desired (5.18) and (5.17) follow from the first-order and second-order necessary conditions of the maximum condition (5.14), respectively. □

To summarize the above results, we have the following theorem.

Theorem 5.3. *Assume that* $R(t)$ *is decomposed according to* (5.1) *such that the SRE* (5.5) *has a solution* $P(\cdot)$. *If the stochastic LQR problem* (5.3)–(5.4) *has an optimal control* $u^*(\cdot) = (u_1^*(\cdot), u_2^*(\cdot))$, *then the following must hold:*

$$D_2'[P - PD_1(R_1 + D_1'PD_1)^{-1}D_1'P]D_2 + R_2 \geq 0. \tag{5.20}$$

Moreover, if we assume that the inequality in (5.20) *is strict* (*for all* $t \in [0,T]$), *then the optimal control* $u^*(\cdot) = (u_1^*(\cdot), u_2^*(\cdot))$ *can be obtained by the following:*

$$u_2^*(t) = -\bar{R}^{-1}(t)\{[B_2(t)'P(t) + D_2(t)'P(t)C_1(t)]q(t) + B_1(t)'\phi(t) + D_1(t)'\lambda(t)\}, \tag{5.21}$$

where $(q(\cdot), \phi(\cdot), \lambda(\cdot))$ *is the solution to the following forward-backward SDE:*

$$\begin{cases} dq(t) = \{[A_1 - B_1\bar{R}^{-1}(B_2'P + D_2'PC_1)]q - B_1(\bar{R}^{-1} + K_1^{-1})B_1'\phi \\ \qquad -B_1(\bar{R}^{-1} + K_1^{-1})D_1'\lambda\}(t)dt \\ \qquad +\{[C_1 - D_1\bar{R}^{-1}(B_2'P + D_2'PC_1)]q - D_1(\bar{R}^{-1} + K_1^{-1})B_1'\phi \\ \qquad -D_1(\bar{R}^{-1} + K_1^{-1})D_1'\lambda\}(t)dW(t), \\ d\phi(t) = -\{[A_1 - B_1\bar{R}^{-1}(B_2'P + D_2'PC_1)]'\phi + [C_1 - D_1\bar{R}^{-1}(B_2'P + D_2'PC_1)]'\lambda \\ \qquad +(B_2'P + D_2'PC_1)'\bar{R}^{-1}(B_2'P + D_2'PC_1)\}(t)dt + \lambda(t)dW(t), \\ q(0) = -x_0, \\ \phi(T) = 0, \end{cases} \tag{5.22}$$

and

$$\begin{aligned} u_1^*(t) = -K_1^{-1}(t)\{&[B_1(t)'P(t) + D_1(t)'P(t)C(t)]x(t) + B_1(t)'\phi(t) \\ &+D_1(t)'\lambda(t) + D_1(t)'P(t)D_2(t)u_2^*(t)\}, \end{aligned} \tag{5.23}$$

with $x(\cdot)$ *being the solution to the following SDE:*

$$\begin{cases} dx(t) = [A_1x - (B_1K_1^{-1}D_1'PD_2 - B_2)u_2^* - B_1K_1^{-1}B_1'\phi - B_1K_1^{-1}D_1'\lambda](t)dt \\ \qquad +[C_1x - (D_1K_1^{-1}D_1'PD_2 - D_2)u_2^* - D_1K_1^{-1}B_1'\phi - D_1K_1^{-1}D_1'\lambda](t)dW(t), \\ x(0) = x_0. \end{cases} \tag{5.24}$$

Proof. First, (5.22) is obtained by replacing $u_2^*(t)$ with the right-hand side of (5.18) in (5.16) and (5.6). This is an essentially *decoupled* forward-backward SDE. Therefore, one can first solve the backward equation for (ϕ, λ) and then solve the forward equation for q. The existence and uniqueness of the solutions to the two equations are clear, as they are linear with bounded coefficients and square-integrable nonhomogeneous terms. Next, if $u^*(\cdot) = (u_1^*(\cdot), u_2^*(\cdot))$ is optimal for the problem (5.3)–(5.4), then by Lemma 5.1, $u_2^*(\cdot)$ must minimize $J_2(u_2(\cdot)) \equiv \inf_{u_1(\cdot)} J(u_1(\cdot), u_2(\cdot))$, which is a backward LQR problem where the dynamics is (5.6) and the cost is (5.9). Then (5.20) is implied by Theorem 5.2. Furthermore, as mentioned, if we substitute the optimal control obtained in (5.18) to (5.6) and (5.16), then we get exactly (5.22). Thus (5.21) is given by (5.18). Second, (5.24) is obtained by replacing $u_1^*(t)$ with (5.8) in (5.3), which evidently has a unique solution. Then (5.23) is identical to (5.8). This completes the proof. □

Remark 5.1. Condition (5.20) gives the precise trade-off between the "good" part (R_1, D_1) and the "bad" part (R_2, D_2) of the overall system in order for the problem to be well-posed and solvable. For example, suppose $R_1 > 0$ and $R_2 \leq 0$. Then (5.20) shows that R_2 cannot be more negative than $-D_2'[P - PD_1(R_1 + D_1'PD_1)^{-1}D_1'P]D_2$. On the other hand, we see that the larger D_2 is (which is the part of the noise corresponding to the "bad" component), or the smaller D_1 is (which is the part of the noise corresponding to the "good" component), the more likely it is that the LQR problem will be well-posed.

Example 5.1. Consider the following LQR problem:

$$\begin{array}{ll} \text{Minimize} & J = E\left\{\int_0^1 \frac{1}{2}[x^2(t) + r_1(t)u_1^2(t) + r_2(t)u_2(t)^2]dt + \frac{1}{2}x(1)^2\right\} \\ \text{subject to} & \begin{cases} dx(t) = (u_1(t) + u_2(t))dW(t), \\ x(0) = 0, \end{cases} \end{array} \tag{5.25}$$

where both $r_1(t)$ and $r_2(t)$ are possibly negative. In this case, $A(t) = B_1(t) = B_2(t) = C(t) = 0$, $D_1(t) = D_2(t) = 1$, $Q(t) = 1$, $(R_1(t), R_2(t)) = (r_1(t), r_2(t))$, and $H = 1$. The SRE (5.5) corresponding to u_1 is

$$\dot{P}(t) = -1, \qquad P(1) = 1.$$

Hence $P(t) = 2 - t$. The equation has a solution when $r_1(t) + 2 - t > 0$ $\forall t \in [0, 1]$, which is equivalent to $r_1(t) > -1$. In this case $K_1(t) = r_1(t) + 2 - t$. It is easy to see that the necessary condition (5.20) for the overall problem to be well-posed reduces to

$$\frac{r_1(t)(2-t)}{r_1(t) + 2 - t} + r_2(t) \geq 0 \qquad \forall t \in [0, 1]. \tag{5.26}$$

Moreover, when the above inequality is strict, the forward-backward equation (5.22) specializes to

$$\begin{cases} dq(t) = 0, \;\; d\phi(t) = \lambda(t)dW(t), \\ q(0) = 0, \;\; \phi(1) = 0, \end{cases}$$

which, by the uniqueness of its solutions, has the solution $(q(t), \phi(t), \lambda(t)) = (0, 0, 0)$. Therefore the optimal (feedback) controls (5.23) and (5.21) (if they exist) reduce to $u_1^*(t) = u_2^*(t) = 0$.

To see that the above control is indeed optimal under (5.26), by applying Ito's formula to $x(t)^2$ one easily gets the following:

$$\begin{aligned} J(u(\cdot)) &= \frac{1}{2}E\int_0^1 \{(2-t)[u_1(t)+u_2(t)]^2 + r_1(t)u_1(t)^2 + r_2(t)u_2(t)^2\}dt \\ &= \frac{1}{2}E\int_0^1 [(r_1(t)+2-t)u_1(t)^2 + 2(2-t)u_1(t)u_2(t) + (2-t+r_2(t))u_2(t)^2]dt \\ &\geq \frac{1}{2}E\int_0^1 \left[(r_1(t)+2-t)u_1(t)^2 + 2(2-t)u_1(t)u_2(t) + \frac{(2-t)^2}{r_1(t)+2-t}u_2(t)^2\right]dt \\ &= \frac{1}{2}E\int_0^1 (r_1(t)+2-t)\left[u_1(t) + \frac{2-t}{r_1(t)+2-t}u_2(t)\right]^2 dt, \end{aligned}$$

where the last inequality is due to (5.26). Therefore the optimal value must be nonnegative. However, the control $(u_1^*(t), u_2^*(t)) = (0,0)$ gives rise to the zero cost, and hence must be optimal.

Let us now look at a more substantial example.

Example 5.2. Consider the following LQR problem:

$$\text{(5.27)} \qquad \begin{array}{ll} \text{Minimize} & J = E\left\{\displaystyle\int_0^1 \frac{1}{2}[r_1u_1(t)^2 + r_2u_2(t)^2]dt + \frac{1}{2}x(1)^2\right\} \\ \text{subject to} & \left\{\begin{array}{l} dx(t) = u_1(t)dt + u_2(t)dW(t), \\ x(0) = x_0, \end{array}\right. \end{array}$$

where $r_1 > 0$ and r_2 are given constants. In this case, $A(t) = B_2(t) = C(t) = D_1(t) = 0$, $B_1(t) = D_2(t) = 1$, $Q(t) = 0$, $(R_1(t), R_2(t)) = (r_1, r_2)$, and $H = 1$. We see that $u_1(\cdot)$ and $u_2(\cdot)$ control the drift and diffusion parts, respectively. The SRE (5.5) corresponding to u_1 is

$$\text{(5.28)} \qquad \left\{\begin{array}{l} \dot{P}(t) - P(t)^2r_1^{-1} = 0, \\ P(1) = 1, \\ r_1 > 0. \end{array}\right.$$

The solution to this equation is

$$\text{(5.29)} \qquad P(t) = \frac{r_1}{1+r_1-t}.$$

We can then calculate via (5.7) to get

$$\text{(5.30)} \qquad A_1(t) = \frac{1}{t-1-r_1}, \qquad C_1(t) = 0.$$

Moreover, condition (5.20) reduces to

$$\text{(5.31)} \qquad \bar{R}(t) \equiv \frac{r_1}{1+r_1-t} + r_2 \geq 0 \;\; \forall t \in [0,1],$$

or, equivalently,

$$\text{(5.32)} \qquad \frac{r_1}{1+r_1} + r_2 \geq 0.$$

This gives the trade-off between r_1 and r_2. Furthermore, when the above inequality is strict, the equation for $(\phi(\cdot), \lambda(\cdot))$ in (5.22) becomes a homogeneous linear equation with terminal value being 0. Hence by the uniqueness of its solution we have $\phi(t) \equiv \lambda(t) \equiv 0$. Therefore (5.21) gives $u_2^*(t) = 0$. Putting $u_2^*(\cdot), \phi(\cdot)$ and $\lambda(\cdot)$ into (5.24), we get

$$\dot{x}(t) = \frac{1}{t-1-r_1}x(t), \quad x(0) = x_0. \tag{5.33}$$

The solution to this equation is $x(t) = \frac{1+r_1-t}{1+r_1}x_0$. From (5.23), it follows that $u_1^*(t) = -\frac{1}{1+r_1}x_0$. To conclude, the optimal control of this problem, if it ever exists, must be given by $u^*(t) = (-\frac{1}{1+r_1}x_0, 0)$. A direct computation shows that the corresponding optimal value is $\frac{1}{2}\frac{r_1}{1+r_1}x_0^2$ (which is nonnegative).

6. Concluding remarks.

The stochastic LQR problem with indefinite control weights is not only mathematically interesting and challenging, but also practically important in touching the deep nature of the uncertainty as well as suggesting how to control the uncertainty. In [9], it is shown that the stochastic LQR problem can be completely solved if the corresponding SRE has a solution. In this paper, we presented the sufficient conditions for the existence and uniqueness of solutions to SRE in the normal case. As for the indefinite case (with $C \neq 0$), we are still not able to prove the solvability of SRE. However, we construct an optimal feedback control for the original LQR problem based on a decomposition approach.

It is interesting to see that the above decomposition approach leads to a *backward* LQR problem. This problem is actually interesting in its own right. More generally, we may consider a nonlinear stochastic control problem with backward dynamics. This kind of problem may also have potential applications in finance.

The work in [9] has led to a series of in-depth research projects on indefinite stochastic LQR control. For example, the problem with integral constraints was studied in [17], and the discrete-time case was treated in [18]. In [10] the problem with random coefficients was tackled by using functional analysis and forward-backward SDE theory. The computational aspect of the problem was first investigated in [1] by means of linear matrix inequalities and semidefinite programming, followed by a thorough study in [24]. Applications to finance problems were discussed in [26, 16].

To conclude, this is a very exciting research domain. Many fundamentally important problems remain open, and the resolutions to these problems will in turn give rise to new problems.

Acknowledgment. The authors would like to thank the two referees for their constructive comments that led to an improved version of the paper.

REFERENCES

[1] M. Ait Rami and X.Y. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls*, IEEE Trans. Automat. Control, to appear.

[2] B.D.O. Anderson and J.B. Moore, *Optimal Control—Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[3] M. Athens, *Special issues on linear-quadratic-Gaussian problem*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 527–869.

[4] A. Bensoussan, *Lectures on stochastic control*, in Nonlinear Filtering and Stochastic Control, Lecture Notes in Math. 972, Springer-Verlag, New York, 1982, pp. 1–62.

[5] J.-M. Bismut, *Analyse Convexe et Probabilites*, These, Faculte des Sciences de Paris, 1973.

[6] J.-M. Bismut, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62–78.

[7] V. Brusin and V. Ugrinovskii, *Absolute stability approach to stochastic stability of infinite dimensional nonlinear systems*, Automatica J. IFAC, 31 (1995), pp. 1453–1458.

[8] F. Bucci and L. Pandolfi, *The regulator problem with indefinite quadratic cost for boundary control systems: The finite horizon case*, Systems Control Lett., 39 (2000), pp. 79–86.

[9] S. Chen, X. Li, and X.Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[10] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems,* Appl. Math. Optim., to appear.

[11] D.J. Clements and B.D.O. Anderson, *Singular Optimal Control: The Linear–Quadratic Problem*, Springer-Verlag, Berlin, Heidelberg, 1978.

[12] M.H.A. Davis, *Linear Estimation and Stochastic Control*, Chapman and Hall, London, 1977.

[13] N. Dokuchaev and X. Zhou, *Stochastic controls with terminal contingent conditions*, J. Math. Anal. Appl., 238 (1999), pp. 143–165.

[14] M.L.J. Hautus and L.M. Silverman, *System structure and singular control*, Linear Algebra Appl., 50 (1983), pp. 369–402.

[15] R.E. Kalman, *Contributions to the theory of optimal control,* Bol. Soc. Math. Mexicana, 5 (1960), pp. 102–119.

[16] M. Kohlmann and X.Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.

[17] A.E.B. Lim and X.Y. Zhou, *Stochastic optimal LQR control with integral quadratic constraints and indefinite control weights*, IEEE Trans. Automat. Control, 44 (1999), pp. 1359–1369.

[18] J.B. Moore, X.Y. Zhou, and A.E.B. Lim, *Discrete time LQG controls with control dependent noise*, Systems Control Lett., 36 (1999), pp. 199–206.

[19] L. Pandolfi, *The Kalman–Yakubovich–Popov theorem for stabilizable hyperbolic boundary control systems*, Integral Equations Operator Theory, 34 (1999), pp. 478–493.

[20] E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.

[21] S. Peng, *Backward stochastic differential equations and applications to optimal control*, Appl. Math. Optim., 27 (1993), pp. 125–144.

[22] J.C. Willems, A. Kitapci, and L.M. Silverman, *Singular optimal control: A geometric approach*, SIAM J. Control Optim., 24 (1986), pp. 323–337.

[23] W.M. Wonham, *On a matrix Riccati equation of stochastic control,* SIAM J. Control, 6 (1968), pp. 681–697.

[24] D. Yao, S. Zhang, and X.Y. Zhou, *Stochastic LQ control via semidefinite programming*, SIAM J. Control Optim., to appear.

[25] J. Yong and X.Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.

[26] X.Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. Control Optim.
Vol. 39, No. 4, pp. 1082–1111

ON STATE CONSTRAINT REPRESENTATIONS AND MESH-DEPENDENT GRADIENT PROJECTION CONVERGENCE RATES FOR OPTIMAL CONTROL PROBLEMS*

J. C. DUNN†

Abstract. Two distinct nonlinear programming formulations are investigated for ODE optimal control problems with pointwise state and control constraints. The first formulation treats the differential equations of state as an equality constraint in the conventional manner. The second formulation employs a different equality constraint entailing the integrated state transition map. Related convergence rate estimates are developed for augmented gradient projection methods and discrete-time approximations to a large representative class of ODE control problems. In the first formulation, the rate estimates are mesh-dependent, and the predicted number of inner loop gradient projection iterations needed to achieve a fixed small deviation from the optimal value of the augmented Lagrangian is inversely proportional to the square of the mesh width. In the second formulation, the convergence rates and predicted iteration counts are mesh-invariant. The computational costs-per-iteration in the two formulations are comparable. These estimates elucidate previously published numerical experiments with augmented gradient projection methods and constrained regulator problems.

Key words. optimal control, pointwise state/control constraints, regulator problems, nonlinear programs, constraint representations, discrete-time approximations, gradient projection methods, mesh-dependent convergence rates

AMS subject classifications. 49M07, 49M10, 49K15, 65K10, 90C06

PII. S0363012999351656

1. Introduction.

An ODE or PDE optimal control problem can be formulated in various ways as an infinite-dimensional nonlinear program (NLP) with specially structured cost function and constraints. Each such formulation, combined with a standard iterative NLP method and an approximate finite difference or finite element representation for the differential equations of state, yields a different numerical algorithm for the control problem. In general, the effectiveness of the resulting algorithm depends not only on the iterative method and discrete approximation scheme, but also on the underlying NLP formulation. The numerical experiments reported in [17] demonstrate this for an unscaled augmented gradient projection (AGP) method and the Euler finite difference approximation expressed in two different NLP formulations of a prototype ODE regulator problem with pointwise state and control constraints.

The first NLP formulation addressed in [17] is the conventional one for state constrained control problems, i.e., state and control vectors x and u are the primal variables and the discrete approximation to the differential equations of state is treated as an equality constraint that implicitly defines a functional relationship between x and u. The second formulation in [17] is new for control problems with pointwise state constraints. In this scheme, the state vectors in the primal variable set are replaced by artificial variables v linked to u through an equality constraint $\phi(u)-v=0$ that entails the explicit "integrated" state solution map $u \to x = \phi(u)$ for the discrete equations of state. Computational costs-per-iteration for efficient implementations of the unscaled

*Received by the editors February 1, 1999; accepted for publication (in revised form) August 2, 2000; published electronically November 15, 2000. This research was supported by NSF research grant DMS-9803755.

http://www.siam.org/journals/sicon/39-4/35165.html

†Mathematics Department, Box 8205, North Carolina State University, Raleigh, NC 27695-8205 (joe_dunn@ncsu.edu).

AGP method are similar in the two NLP formulations, and inversely proportional to the Euler mesh width in each case. On the other hand, the mesh-dependent convergence properties of the AGP method are very different in the two schemes. In the conventional NLP formulation, the number of inner and outer AGP iterates required to achieve fixed residual tolerances increases rapidly with decreasing mesh width, whereas the comparable iteration counts in the alternative formulation are essentially mesh-invariant. This formulation-dependent disparity in iteration counts is clearly reflected in the ordinates of the FLOP count graphs of [17], which can already differ by several orders of magnitude even on meshes that are still too coarse to support good approximations to the limiting ODE control problem.

The present article reinforces the numerical experiments of [17] with an analysis of the AGP inner loop based on the convergence theory of [11] and new root-mean-square (RMS) norm quadratic growth and Lipschitz norm estimates for the augmented Lagrangian and its gradient. The latter estimates are derived separately for the two NLP formulations of the uniform mesh Euler discrete-time approximation for an important class of ODE optimal control problems with strongly convex costs, linear state equations, and pointwise quasi-convex inequality constraints on control and state vectors. The Lagrangian quadratic growth constants behave similarly in both formulations as the Euler mesh-width approaches zero; however, the corresponding Lagrangian gradient Lipschitz norms act quite differently. In the conventional formulation, the Lipschitz norm is inversely proportional to the square of the mesh width. In the alternative formulation, the Lipschitz norm remains bounded as the mesh width approaches zero. These estimates produce disparate cost value convergence rate bounds that mirror the behavior observed in the numerical calculations of [17]. Although a comparable analysis for higher order discrete approximation schemes with uniform or nonuniform time steps is not attempted in the present investigation, a consideration of the $\mathbb{L}^2$ continuity properties of the continuous-time augmented Lagrangians suggests that similar formulation-dependent convergence rate disparities are likely to occur for any consistent numerical integration scheme. It is interesting to note here that mesh-invariant Lipschitz continuity properties for derivatives are also crucial in the mesh independence principle of Allgower et al. [2] for consistent finite-dimensional implementations of Newton's method for nonlinear equations $F(x) = 0$ in Banach spaces.

When pointwise constraints are imposed separately on the state and control variables, the associated alternative NLP formulation proposed in [17] enjoys two additional advantages. In this commonly encountered situation, the sets of admissible primal variable values $(u(t), v(t))$ are finite-dimensional Cartesian products, the corresponding augmented Lagrangian can be minimized *explicitly* with respect to the artificial variable v, the resulting reduced augmented Lagrangian is often locally Lipschitz continuously differentiable, and hence gradient projection methods can be used to minimize this reduced Lagrangian approximately with respect to the remaining primal variable u. (Augmented Lagrangian methods employing similar reduced Lagrangians have been proposed by Rockafellar [28] and Bertsekas [5] in conjunction with squared slack variable treatments of general inequality-constrained finite-dimensional NLPs.) It is readily seen that the elimination of v from the primal variable set will generally yield a major reduction in the cost of each inner iteration in the AGP scheme. Moreover, estimates derived here for a gradient projection (GP) method prototype applied with and without explicit minimization of the augmented Lagrangian in v

indicate that the former AGP scheme not only has a smaller computational cost per inner iteration, but also may converge more rapidly.

An extreme instance of separated pointwise constraints occurs when the state is entirely unrestricted. In this case, the augmented Lagrangian in the alternative NLP formulation of [17] is trivially minimized over v for each admissible u by setting the (unconstrained) artificial variable v equal to the $\phi(u) + \lambda/c$, where λ is the equality constraint's multiplier and c is the penalty constant in the augmented Lagrangian. The resulting reduced augmented Lagrangian is then merely the control problem's cost function with x replaced by $\phi(u)$, the artificial variable v, and the equality constraint and the multiplier λ disappear from the formulation at the first AGP multiplier update, and the AGP scheme effectively collapses to a basic single loop GP iteration for the given control-constrained optimal control problem. Certain mesh-independent ϵ-active constraint identification results have already been proved by Kelley and Sachs [23] for a standard GP method and Euler discrete-time approximations to ODE optimal control problems with bounded scalar control variables and unrestricted states. In addition, infinite-dimensional local convergence rate theorems have also been proved in [29] and [16] for this class of bounded input ODE optimal control problems, and for vector-valued extensions with polyhedral admissible control input sets.

For general NLPs, deteriorating convergence rates are also seen in AGP inner loop iterations when the penalty constant c is increased without bound and the corresponding inner relaxed AGP optimization problem thereby becomes increasingly ill-conditioned. First order scaling operators designed to address this issue have been investigated by Luenberger [25] and Hager [20] for unconstrained relaxed optimization problems with penalty terms, and by Hager [21] for analogous constrained problems with penalized objective functions. These scaling operators are prima facie well suited to the conventional NLP formulation of the pointwise constrained ODE optimal control problem, since they derive from the Jacobian of the penalized dynamic equation constraint, and the inherent sparse "staircase structure" in this Jacobian should be exploitable in the calculation of the iteration maps for associated scaled gradient-related descent methods. (This is clearly true for the unconstrained relaxed penalized problems that arise in the absence of pointwise constraints, but somewhat less obvious for pointwise inequality-constrained relaxed penalized problems and related scaled GP methods.) Moreover, since the mesh width parameter appears explicitly in the discrete-time dynamic equation constraints, it has been suggested that the scaling principles in [25], [20], and [21] may enhance the convergence behavior of associated gradient-related descent methods, not only as c increases without bound *but also as the mesh width approaches* 0. Unfortunately, this interesting conjecture is at present completely unsupported by any numerical or theoretical investigations for pointwise inequality-constrained ODE control problems and related AGP methods. Even in the absence of pointwise constraints, there are no pertinent published numerical results, and it is not known how the c-invariant upper limits on geometric convergence ratios in the extant general scaled steepest descent convergence theorems of [25] and [20] behave as the mesh width approaches 0. In particular, if these upper limits are not bounded away from 1 as the mesh width approaches 0, the associated scaled steepest descent methods may exhibit mesh-dependent deteriorating convergence rates.

2. The ODE optimal control problem. In the continuous-time optimal control problems treated here, the control schedules $u(\cdot)$ are considered to be bounded measurable functions from $[0, 1]$ to $\mathbb{R}^m$, the state trajectories $x(\cdot)$ are absolutely continuous functions from $[0,1]$ to $\mathbb{R}^n$, the cost functionals are integrated running loss functions,

$$\int_0^1 f^0(t,u(t),x(t))\,dt, \tag{2.1}$$

and the constraints are differential side conditions,

$$\frac{dx}{dt}(t) \overset{a.e.}{=} f(t,u(t),x(t)), \qquad t \in [0,1], \tag{2.2a}$$

$$x(0) = \eta_0, \tag{2.2b}$$

and pointwise vector-valued inequalities,

$$\gamma(t,u(t),x(t)) \leq 0, \qquad t \in [0,1]. \tag{2.3}$$

The ensuing analysis is restricted to Lipschitz continuously differentiable strongly convex cost functionals, continuous quasi-convex constraint component functions $\gamma_1(t,\cdot),\ldots,\gamma_r(t,\cdot)$, and linear rate functions,

$$f(t,\xi,\eta) = A(t)\eta + B(t)\xi, \tag{2.4}$$

with bounded measurable matrix-valued coefficient functions $A(\cdot)$ and $B(\cdot)$. The class of optimal control problems that meet these restrictions is broadly representative, and includes the prototype ODE linear-quadratic regulator (LQR) problem with pointwise bounds on the state and control vector components.

3. Nonlinear programming formulations.

The ODE optimal control problem of section 2 can be cast as an infinite-dimensional NLP,

$$\min_{w\in\mathbb{S}} J(w), \tag{3.1a}$$

subject to the constraints

$$g(w) \leq 0 \tag{3.1b}$$

and

$$h(w) = 0, \tag{3.1c}$$

where $\mathbb{S}$ is a linear variety (i.e., translated subspace) in a normed vector space $\mathbb{W}$, J is a real-valued functional on $\mathbb{W}$, g and h map $\mathbb{W}$ to normed vector spaces $\mathbb{Z}_g$ and $\mathbb{Z}_h$, respectively, and "$\leq$" is a partial order relation on $\mathbb{Z}_g$. Many such formulations are possible. Two are considered here.

3.1. Formulation I.

The ODE optimal control problem of section 2 is commonly treated as an infinite-dimensional nonlinear program (3.1) in a direct sum $\mathbb{W} = \mathbb{U} \oplus \mathbb{X}$, where $\mathbb{U}$ is a vector space of bounded measurable functions $u(\cdot) : [0,1] \to \mathbb{R}^m$, $\mathbb{X}$ is a vector space of absolutely continuous functions $x(\cdot) : [0,1] \to \mathbb{R}^n$ with bounded measurable derivatives, $\mathbb{S}$ is the linear variety of pairs $w = (u,x) \in \mathbb{W}$ for which $x(0) = \eta_0$, $\mathbb{Z}_g$ is a space of bounded measurable functions from $[0,1]$ to $\mathbb{R}^r$ with the standard pointwise partial order relation, $\mathbb{Z}_h$ is a space of bounded measurable functions from $[0,1]$ to $\mathbb{R}^n$, and J, g, and h are defined by

$$J(w) = \int_0^1 f^0(t, u(t), x(t))\, dt, \tag{3.2a}$$

$$g(w)(t) = \gamma(t, u(t), x(t)), \tag{3.2b}$$

and

$$h(w)(t) \overset{a.e}{=} \frac{dx}{dt}(t) - A(t)x(t) - B(t)u(t) \tag{3.2c}$$

for $t \in [0,1]$. Most often, $\mathbb{Z}_g$ and $\mathbb{Z}_h$ are equipped with standard $\mathbb{L}^\infty$ norms, and the norm on $\mathbb{W}$ is comprised of matching $\mathbb{L}^\infty$ and Sobolev norms on the u and x components of w, respectively. Under reasonable conditions on f^0 and γ, the functions J, g, and h are then well defined and continuously Fréchet differentiable at least once. However, since finite-dimensional discrete-time augmented gradient projection calculations are typically implemented in simple RMS norms on the vectors that approximate w and $h(w)$, it can be seen that smoothness properties of J, g, and h relative to function space $\mathbb{L}^2$ counterparts of the RMS norms on x are potentially more interesting for present considerations than the contrasting $\mathbb{L}^\infty$-Sobolev properties. Hence, it is significant that the equality constraint map h in (3.2c) and an associated augmented Lagrangian,

$$\begin{aligned} L(\lambda, w) &= J(w) + \langle \lambda, h(w) \rangle + \frac{1}{2}c\, \|h(w)\|^2 \\ &= J(w) + \frac{1}{2}c \left(\left\| \frac{\lambda}{c} + h(w) \right\|^2 - \left\| \frac{\lambda}{c} \right\|^2 \right), \end{aligned} \tag{3.3}$$

are not even continuous, much less continuously differentiable, when $\mathbb{U}$, $\mathbb{X}$, and $\mathbb{Z}_h$ are provided with $\mathbb{L}^2$ inner products and norms. This $\mathbb{L}^2$ smoothness singularity suggests that the convergence properties of discrete-time AGP implementations related to the present continuous-time NLP formulation may deteriorate on increasingly refined meshes.

Note 1. When the subspaces $\mathbb{X}$ and $\mathbb{Z}_h$ are equipped with $\mathbb{L}^2$ norms, it is easily shown that the linear differential operator in (3.2c) is unbounded and hence discontinuous. As might be expected, a related construction in Example 1 of section 7 demonstrates that the RMS norms of the associated bounded Euler finite-difference operators on uniform meshes for the interval $[0,1]$ increase without limit as the mesh width approaches zero, and the resulting unbounded growth in Lipschitz norms of the Lagrangian gradient has potentially adverse consequences for AGP convergence rates.

On the other hand, suppose that $\mathbb{X}$ is supplied with the inner product induced $\mathbb{X}_1^2$ Sobolev norm, and that $\mathbb{U}$, $\mathbb{Z}_g$, and $\mathbb{Z}_h$ are provided with the inner product induced $\mathbb{L}^2$ norms. The differential operator d/dt is now trivially bounded, and for a restricted but still important class of functions f^0 and γ, the related maps h, g and J and L are Lipschitz continuously Fréchet differentiable with respect to the inherited inner product induced norm on the direct sum $\mathbb{W} = \mathbb{U} \oplus \mathbb{X}$, and in addition, the augmented Lagrangian L simultaneously satisfies strong convexity and quadratic coercivity conditions in this norm. (The latter properties can be established with constructions similar to those used in the proof of Lemma 8.3.) This suggests that discrete-time AGP implementations in finite-dimensional inner product space counterparts of $\mathbb{W}$ may exhibit asymptotically mesh-independent convergence rates as mesh widths are reduced to zero. Unfortunately, in practice, the required gradient/projection computations in the Sobolev-like finite-dimensional weighted inner products are prohibitively expensive.

3.2. Formulation II. Reference [17] proposes an alternative NLP formulation for the optimal control problem of section 2 in a direct sum $\mathbb{W} = \mathbb{U} \oplus \mathbb{V}$, where $\mathbb{U}$ and $\mathbb{V}$ are vector spaces of bounded measurable functions from $[0,1]$ to $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, the linear variety $\mathbb{S}$ is $\mathbb{W}$ itself, and $\mathbb{Z}_g$ and $\mathbb{Z}_h$ are as before in section 3.1. In this setting, J, g, and h are prescribed by

$$J(u) = \int_0^1 f^0(t, u(t), \phi(u)(t))\, dt, \tag{3.4a}$$

$$g(w)(t) = \gamma(t, u(t), v(t)) \tag{3.4b}$$

and

$$h(w)(t) = \phi(u)(t) - v(t) \tag{3.4c}$$

for $t \in [0,1]$, where $\phi(u)$ is the unique absolutely continuous solution of the linear ODE initial value problem

$$\frac{dx}{dt}(t) \overset{a.e.}{=} A(t)x(t) + B(t)u(t), \qquad t \in [0,1], \tag{3.5a}$$

$$x(0) = \eta_0. \tag{3.5b}$$

Equivalently, ϕ is the affine map defined by

$$\phi(u) = \Theta u + \theta, \tag{3.6a}$$

where

$$(\Theta u)(t) = \int_0^t \Phi(t,s)B(s)u(s)\, ds \tag{3.6b}$$

and

$$\theta(t) = \Phi(t,0)\eta_0 \tag{3.6c}$$

for $t \in [0,1]$, and $\Phi(\cdot, s)$ is the fundamental solution matrix prescribed by the initial value problem

$$\frac{\partial}{\partial t}\Phi(t,s) \overset{a.e.}{=} A(t)\Phi(t,s), \qquad t \in [0,1], \tag{3.7a}$$

$$\Phi(s,s) = I \tag{3.7b}$$

for $s \in [0,1]$ [8]. Note that the cost J in (3.4a) depends only on the u component of w. Note also that if v and the equality constraint function (3.4c) were removed and if $x(t)$ were replaced by $\phi(u)(t)$ in γ, then the resulting inequality constraints $g(u)(t) = \gamma(t, u(t), \phi(u)(t)) \leq 0$ would also refer only to u but would *not* impose simple pointwise conditions on u in the interesting cases where $\gamma(t, u, x) \leq 0$ implies nontrivial restrictions on the state variable x.

The introduction of v and the equality constraint function (3.4c) not only simplifies the inequality constraint $g(w) \leq 0$ in this context, but also ensures differentiability of the corresponding augmented Lagrangian (3.3) in norms suitable for AGP methods. If the function spaces $\mathbb{U}$, $\mathbb{V}$, and $\mathbb{Z}_h$ in this NLP formulation are considered to be pre-Hilbert spaces with $\mathbb{L}^2$ inner products and norms, then the affine constraint function h in (3.4c) is Lipschitz continuously Fréchet differentiable, and for a large and important class of running loss functions f^0, the augmented Lagrangian (3.3) is also $\mathbb{L}^2$-Lipschitz continuously differentiable [16]. In such cases, there is no $\mathbb{L}^2$ smoothness singularity for the augmented Lagrangian in the present alternative continuous-time NLP formulation, and it seems plausible that the convergence properties of the related discrete-time AGP implementations may *not* deteriorate on increasingly refined meshes.

The heuristic arguments set forth here and in section 3.1 are vindicated by the subsequent AGP convergence analyses for finite-dimensional discrete-time approximations to the continuous-time optimal control problem of section 2.

4. Riemann–Euler discrete-time approximations.

Let k be a positive integer greater than 1, put $\Delta t = 1/k$, and construct the uniform mesh

$$0 = t_0 < \cdots < t_k = 1 \tag{4.1a}$$

with

$$t_{i+1} = t_i + \Delta t \tag{4.1b}$$

for $i = 0, \ldots, k-1$. Then the integral cost functional, linear ODE state equations, and pointwise inequality constraints in the optimal control formulation of section 2 have the corresponding first order Riemann–Euler approximations:

$$\sum_{i=0}^{k-1} f^0(t_i, u_i, x_i)\, \Delta t, \tag{4.2}$$

$$\frac{x_{i+1} - x_i}{\Delta t} = A(t_i)x_i + B(t_i)u_i, \qquad 0 = 1, \ldots, k-2, \tag{4.3a}$$

$$x_0 = \eta_0, \tag{4.3b}$$

and

$$\gamma(t_i, u_i, x_i) \leq 0, \qquad i = 0, \ldots, k-1. \tag{4.4}$$

For each control sequence $u = (u_0, \ldots, u_{k-1}) \in \mathbb{R}^{km}$, the discrete-time state equations (4.3) have a unique solution $\phi(u) = (\phi(u)_0, \ldots, \phi(u)_{k-1}) \in \mathbb{R}^{kn}$. More specifically, ϕ is the affine map defined by

$$\phi(u) = \Theta u + \theta, \tag{4.5a}$$

where

$$(\Theta u)_0 = 0, \tag{4.5b}$$

$$(\Theta u)_i = \sum_{j=0}^{i-1} \Phi_{i,j} B(t_j) u_j \Delta t, \qquad i = 1, \dots, k-1, \tag{4.5c}$$

$$\theta_i = \Phi_{i,0}\eta_0, \qquad i = 0, \dots, k-1, \tag{4.5d}$$

and $\Phi_{\cdot,j}$ is the fundamental solution matrix prescribed by the initial value problem

$$\frac{\Phi_{i+1,j} - \Phi_{i,j}}{\Delta t} = A(t_i)\Phi_{i,j}, \qquad i = j, \dots, k-2, \tag{4.6a}$$

$$\Phi_{j,j} = I \tag{4.6b}$$

for $j = 0, \dots, k-2$. Note that (4.5) immediately yields

$$\Phi_{i,j} = \prod_{l=j}^{i-1} (I + A(t_l)\Delta t) \tag{4.7}$$

for $0 \leq j < i \leq k-1$.

The foregoing constructions now lead directly to discrete-time counterparts of the NLP formulations in sections 3.1 and 3.2.

4.1. Formulation I. A discrete-time approximation to the continuous-time NLP formulation (3.2) is obtained by letting $\mathbb{U} = \mathbb{R}^{km}$, $\mathbb{X} = \mathbb{R}^{kn}$, $\mathbb{W} = \mathbb{U} \oplus \mathbb{X}$, $\mathbb{S} = \{w \in \mathbb{W} : x_0 = \eta_0\}$, $\mathbb{Z}_g = \mathbb{R}^{kr}$, $\mathbb{Z}_h = \mathbb{R}^{(k-1)n}$ and defining the cost and constraint functions J, g, and h by

$$J(w) = \sum_{i=0}^{k-1} f^0(t_i, u_i, x_i)\, \Delta t, \tag{4.8a}$$

$$g(w)_i = \gamma(t_i, u_i, x_i), \qquad i = 0, \dots, k-1, \tag{4.8b}$$

and

$$h(w)_i = \frac{x_{i+1} - x_i}{\Delta t} - A(t_i)x_i - B(t_i)u_i, \qquad i = 0, \dots, k-2. \tag{4.8c}$$

RMS inner products and norms are provided here for the spaces $\mathbb{U}$, $\mathbb{X}$, $\mathbb{W}$, $\mathbb{Z}_g$, and $\mathbb{Z}_h$ as natural discrete-time analogues of continuous-time $\mathbb{L}^2$ inner products and norms. In particular, on $\mathbb{W}$,

$$\langle w^a, w^b \rangle_{rms} = \langle u^a, u^b \rangle_{rms} + \langle x^a, x^b \rangle_{rms} \tag{4.9a}$$

and

$$\|w\|_{rms}^2 = \|u\|_{rms}^2 + \|x\|_{rms}^2, \tag{4.9b}$$

where

$$\langle u^a, u^b\rangle_{rms} = \sum_{i=0}^{k-1} \langle u_i^a, u_i^b\rangle \,\Delta t, \qquad \langle x^a, x^b\rangle_{rms} = \sum_{i=0}^{k-1} \langle x_i^a, x_i^b\rangle \,\Delta t \tag{4.9c}$$

and

$$\|u\|_{rms}^2 = \sum_{i=0}^{k-1} \|u_i\|^2 \,\Delta t, \qquad \|x\|_{rms}^2 = \sum_{i=0}^{k-1} \|x_i\|^2 \,\Delta t, \tag{4.9d}$$

and where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the ordinary Euclidean inner product and norm on $\mathbb{R}^m$ or $\mathbb{R}^n$.

4.2. Formulation II. A discrete-time approximation to the alternative continuous-time NLP formulation (3.4) is obtained by letting $\mathbb{U} = \mathbb{R}^{km}$, $\mathbb{V} = \mathbb{R}^{kn}$, $\mathbb{W} = \mathbb{U}\oplus\mathbb{V}$, $\mathbb{S} = \mathbb{W}$, $\mathbb{Z}_g = \mathbb{R}^{kr}$, $\mathbb{Z}_h = \mathbb{R}^{kn}$, and defining the cost and constraint functions J, g, and h by

$$J(u) = \sum_{i=0}^{k-1} f^0(t_i, u_i, \phi(u)_i) \,\Delta t, \tag{4.10a}$$

$$g(w)_i = \gamma(t_i, u_i, v_i), \tag{4.10b}$$

and

$$h(w)_i = \phi(u)_i - v_i \tag{4.10c}$$

for $i = 0, \ldots, k-1$, where $\phi(\cdot)$ is the affine state equation solution map in (4.5)–(4.6).

As in section 4.1, RMS inner products and norms are provided for the spaces $\mathbb{U}$, $\mathbb{V}$, $\mathbb{W}$, $\mathbb{Z}_g$, and $\mathbb{Z}_h$. In particular, on $\mathbb{W}$,

$$\langle w^a, w^b\rangle_{rms} = \langle u^a, u^b\rangle_{rms} + \langle v^a, v^b\rangle_{rms} \tag{4.11a}$$

and

$$\|w\|_{rms}^2 = \|u\|_{rms}^2 + \|v\|_{rms}^2, \tag{4.11b}$$

where

$$\langle u^a, u^b\rangle_{rms} = \sum_{i=0}^{k-1} \langle u_i^a, u_i^b\rangle \,\Delta t, \qquad \langle v^a, v^b\rangle_{rms} = \sum_{i=0}^{k-1} \langle v_i^a, v_i^b\rangle \,\Delta t \tag{4.11c}$$

and

$$\|u\|_{rms}^2 = \sum_{i=0}^{k-1} \|u_i\|^2 \,\Delta t, \qquad \|v\|_{rms}^2 = \sum_{i=0}^{k-1} \|v_i\|^2 \,\Delta t, \tag{4.11d}$$

and where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ once again denote the Euclidean inner product and norm on $\mathbb{R}^m$ or $\mathbb{R}^n$.

5. Multiplier methods. The multiplier method proposed by Hestenes [22] and Powell [27] treats finite-dimensional equality-constrained minimization problems,

$$\min_{w\in\mathbb{R}^N} J(w) \tag{5.1a}$$

subject to

$$h(w) = 0, \tag{5.1b}$$

by solving a related sequence of relaxed unconstrained minimization problems,

$$\min_{w\in\mathbb{R}^N} L(\lambda^{(i)}, w), \tag{5.2}$$

where $L(\lambda, w)$ is an augmented Lagrangian (3.3) with nonnegative penalty constant c, and $\{\lambda^{(i)}\}$ is a sequence of multiplier vectors generated by the recursion

$$\lambda^{(i+1)} = \lambda^{(i)} + c\, h(w^{(i+1)}) \tag{5.3}$$

in which $w^{(i+1)}$ denotes an approximation to some exact solution of the unconstrained minimization problem (5.2). The exact solution is generally inaccessible and the approximate minimizer $w^{(i+1)}$ is typically produced by a truncated inner iterative calculation that begins at $w^{(i)}$ for the unconstrained problem (5.2). In the classic formulations of [22] and [27], the domain $\mathbb{W} = \mathbb{R}^N$ is equipped with the standard Euclidean norm, the codomain $\mathbb{Z}_h$ is $\mathbb{R}^q$, λ is a vector in $\mathbb{R}^q$, and $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ in (3.3) are the standard Euclidean inner product and norm on $\mathbb{R}^q$.

5.1. AGP methods. Many elaborations of the Hestenes–Powell multiplier method have been proposed and analyzed for equality-constrained minimization problems and NLPs with equality and inequality constraints (see [5], [6], and the bibliographies therein). The AGP methods addressed in [3], [15], [17], and the present investigation are intended for NLPs (3.1) that have difficult equality constraints but tractable quasi-convex inequality constraints with simple closed convex solutions sets,

$$\Omega_g = \{w \in \mathbb{S} : g(w) \le 0\}, \tag{5.4}$$

whose associated proximal point projector maps P_{Ω_g} are easily computed. The simplest of these methods employs the first order Hestenes–Powell multiplier rule (5.3) and a Goldstein–Levitin–Polyak GP method [18], [24] to compute approximate solutions $w^{(i)}$ for the corresponding relaxed NLPs,

$$\min_{w\in\Omega_g} L(\lambda^{(i)}, w). \tag{5.5}$$

More specifically, $w^{(i+1)}$ is obtained from $w^{(i)}$ by a truncated GP iteration,

$$\zeta^{(0)} = w^{(i)}, \tag{5.6a}$$

$$\zeta^{(j+1)} = P_{\Omega_g}(\zeta^{(j)} - \alpha s_j \nabla_w L(\lambda^{(i)}, \zeta^{(j)})), \qquad j = 0, \dots, j_i, \tag{5.6b}$$

$$w^{(i+1)} = \zeta^{(j_i+1)}, \tag{5.6c}$$

where α is a positive scaling parameter and s_j is a real number in $(0,1]$ prescribed by the Bertsekas–Armijo step length rule [4], i.e., with β and δ fixed in $(0,1)$, s_j is the largest number s such that

$$s \in \{1, \beta, \beta^2, \ldots\} \tag{5.7a}$$

and

(5.7b)
$$L(\lambda,\zeta) - L(\lambda, P_{\Omega_g}(\zeta - \alpha s \nabla_w L(\lambda,\zeta))) \geq \delta \langle \nabla_w L(\lambda,\zeta), \zeta - P_{\Omega_g}(\zeta - \alpha s \nabla_w L(\lambda,\zeta))\rangle$$

with $\lambda = \lambda^{(i)}$ and $\zeta = \zeta^{(j)}$. In these formulas, α, β, and δ are constants; however, the convergence rate estimates in subsequent sections are readily modified to accomodate variable scaling parameters $\alpha^{(j)}$ that are bounded away from 0 and ∞, and step length parameters $\beta^{(j)}$ and $\delta^{(j)}$ that are bounded away from 0 and 1.

5.2. Reduced AGP methods. If $\mathbb{W}$ is a direct sum $\mathbb{W}^a \oplus \mathbb{W}^b$ and the feasible set in the relaxed NLP (5.5) is also a Cartesian product,

$$\Omega_g = \Omega_g^a \times \Omega_g^b, \tag{5.8}$$

of closed convex sets $\Omega_g^a \subset \mathbb{W}^a$ and $\Omega_g^b \subset \mathbb{W}^b$, then (5.5) reduces to

$$\min_{w^a \in \Omega_g^a} \hat{L}(\lambda, w^a), \tag{5.9a}$$

where

$$\hat{L}(\lambda, w^a) = \min_{w^b \in \Omega_g^b} L(\lambda, (w^a, w^b)). \tag{5.9b}$$

In some cases, explicit formulas can be found for an *exact* minimizer $\hat{w}^b(\lambda, w^a)$ of $L(\lambda, (w^a, \cdot))$ in Ω_g^b, and for the corresponding reduced augmented Lagrangian $\hat{L}(\lambda, w^a)$. If $\hat{L}(\lambda, w^a)$ retains sufficient smoothness in the variable w^a, then GP methods can be applied to the simpler reduced NLP (5.9) to obtain an approximate minimizer $w = (w^a, \hat{w}^b(\lambda, w^a))$ for $L(\lambda, w)$ in Ω_g. A reduced augmented gradient projection (RAGP) method of this type for control problems with separated pointwise constraints on the control and state variables was proposed in [17], and is investigated further here. Similar strategies for slack variable formulations of general finite-dimensional inequality-constrained NLPs are proposed and analyzed in [28] and [5].

5.3. The AGP method in formulation I. For the discrete-time NLP of section 4.1, the primal variable is a vector $w = (u, x)$ in the direct sum $\mathbb{W} = \mathbb{R}^{km} \oplus \mathbb{R}^{kn}$, and the relaxed NLP (5.5) becomes

$$\min_{w \in \Omega_g} L(\lambda, w), \tag{5.10a}$$

with

$$\Omega_g = \{w : x_0 = \eta_0,\ \gamma(t_i, u_i, x_i) \leq 0, \quad i = 0, \ldots, k-1\}, \tag{5.10b}$$

and

$$L(\lambda, w) = J(w) + \frac{1}{2} c \left(\left\| \frac{\lambda}{c} + \mathcal{D}x - \mathcal{A}x - \mathcal{B}u \right\|_{rms}^2 - \left\| \frac{\lambda}{c} \right\|_{rms}^2 \right), \tag{5.10c}$$

where

$$J(w) = \sum_{i=0}^{k-1} f^0(t_i, u_i, x_i)\,\Delta t, \tag{5.10d}$$

$$\lambda = (\lambda_0, \ldots, \lambda_{k-2}) \in \mathbb{R}^{(k-1)n}, \tag{5.10e}$$

$$(\mathcal{A}x)_i = A(t_i)x_i, \tag{5.10f}$$

$$(\mathcal{B}u)_i = B(t_i)u_i, \tag{5.10g}$$

and

$$(\mathcal{D}x)_i = \frac{x_{i+1} - x_i}{\Delta t} \tag{5.10h}$$

for $i = 0, \ldots, k-2$.

For AGP methods, it is significant that the RMS projection operator for the set Ω_g has the pointwise decomposition formula

$$P_{\Omega_g}(u, x) = (\xi, \eta) \tag{5.11a}$$

with

$$(\xi_i, \eta_i) = P_{Z_i}(u_i, x_i), \qquad i = 0, \ldots, k-1, \tag{5.11b}$$

$$Z_0 = \{(u_0, x_0) : x_0 = \eta_0, \;\; \gamma(t_0, u_0, x_0) \le 0\}, \tag{5.11c}$$

and

$$Z_i = \{(u_i, x_i) : \gamma(t_i, u_i, x_i) \le 0\}, \qquad i = 1, \ldots, k-1. \tag{5.11d}$$

In these equations, P_{Z_i} denotes the standard Euclidean projection operator for sets $Z_i \subset \mathbb{R}^m \oplus \mathbb{R}^n$.

5.4. The AGP method in formulation II. The primal variable in the discrete-time NLP of section 4.2 is a vector $w = (u, v)$ in the direct sum $\mathbb{W} = \mathbb{R}^{km} \oplus \mathbb{R}^{kn}$, and the relaxed NLP (5.5) becomes

$$\min_{w \in \Omega_g} \; L(\lambda, w) \tag{5.12a}$$

with

$$\Omega_g = \{w : \gamma(t_i, u_i, v_i) \le 0, \quad i = 0, \ldots, k-1\}, \tag{5.12b}$$

and

$$L(\lambda, w) = J(u) + \frac{1}{2}c\left(\left\|\frac{\lambda}{c} + \phi(u) - v\right\|_{rms}^2 - \left\|\frac{\lambda}{c}\right\|_{rms}^2\right), \tag{5.12c}$$

where

$$J(u) = \sum_{i=0}^{k-1} f^0(t_i, u_i, \phi(u)_i)\ \Delta t, \tag{5.12d}$$

$$\lambda = (\lambda_0, \ldots, \lambda_{k-1}) \in \mathbb{R}^{kn}, \tag{5.12e}$$

and $\phi(\cdot)$ is the affine state equation solution map in (4.5)–(4.6).

Once again, the RMS projection operator for the set Ω_g in this scheme has the useful pointwise decomposition formula

$$P_{\Omega_g}(u, v) = (\xi, \eta) \tag{5.13a}$$

with

$$(\xi_i, \eta_i) = P_{Z_i}(u_i, v_i), \qquad i = 0, \ldots, k-1, \tag{5.13b}$$

and

$$Z_i = \{(u_i, v_i) : \gamma(t_i, u_i, v_i) \leq 0\}, \qquad 0 = 1, \ldots, k-1. \tag{5.13c}$$

As before, the symbol P_{Z_i} in these equations denotes the standard Euclidean projection operator for sets $Z_i \subset \mathbb{R}^m \oplus \mathbb{R}^n$.

Note that formula (5.12c) can be rewritten as

$$L(\lambda, w) = \sum_{i=0}^{k-1} l^0(t_i.\lambda_i, u_i, v_i, \phi(u)_i)\ \Delta t, \tag{5.14a}$$

where

$$\begin{aligned} l^0(t_i.\lambda_i, u_i, v_i, x_i) &= f^0(t_i, u_i, x_i) \\ &\quad + \frac{1}{2} c \left(\left\| \frac{\lambda_i}{c} + x_i - v_i \right\|^2 - \left\| \frac{\lambda_i}{c} \right\|^2 \right). \end{aligned} \tag{5.14b}$$

Thus, the relaxed NLP (5.12) amounts to a new discrete-time optimal control problem with unconstrained states, pointwise constrained control inputs (u_i, v_i), running losses $l^0(t_i.\lambda_i, u_i, v_i, x_i)$, and the original state equations (4.3). As explained in [17], it follows that the RMS gradient of $L(\lambda, \cdot)$ can be calculated efficiently in $O(k)$ floating point operations (FLOPS) with adjoint recursions for the Euler discrete-time equations (4.3). In fact, when f^0 is twice continuously differentiable, it is also possible to compute Newtonian scaling operators for $L(\lambda, \cdot)$ in $O(k)$FLOPS with auxiliary dynamic programming recursions [13], [14].

5.5. The RAGP method in formulation II. The reduced augmented gradient projection scheme is applicable in formulation II when the pointwise inequality constraints on state and control variables are "separated" in the sense that

$$\gamma(t, \xi, \eta) = (\gamma^u(t, \xi), \gamma^x(t, \eta)), \tag{5.15}$$

where γ^u maps $\mathbb{R}^1 \oplus \mathbb{R}^m$ to $\mathbb{R}^{r_u}$, γ^x maps $\mathbb{R}^1 \oplus \mathbb{R}^n$ to $\mathbb{R}^{r_x}$, and $r_u + r_x = r$. In this frequently encountered special case, the feasible set Ω_g in (5.12b) is a Cartesian product,

$$\Omega_g = \Omega_g^u \times \Omega_g^v, \tag{5.16a}$$

where

$$\Omega_g^u = \{u : \gamma^u(t_i, u_i) \le 0, \quad i = 0, \ldots, k-1\} \tag{5.16b}$$

and

$$\Omega_g^v = \{v : \gamma^x(t_i, v_i) \le 0, \quad i = 0, \ldots, k-1\}. \tag{5.16c}$$

Hence, the corresponding reduced NLP (5.9) becomes

$$\min_{u \in \Omega_g^u} \hat{L}(\lambda, u) \tag{5.17a}$$

with

$$\begin{aligned} \hat{L}(\lambda, u) &= \left(J(u) - \frac{1}{2} c \left\| \frac{\lambda}{c} \right\|_{rms}^2 \right) + \frac{1}{2} c \min_{v \in \Omega_g^v} \left\| \left(\frac{\lambda}{c} + \phi(u) \right) - v \right\|_{rms}^2 \\ &= \left(J(u) - \frac{1}{2} c \left\| \frac{\lambda}{c} \right\|_{rms}^2 \right) + \frac{1}{2} c \left\| \left(\frac{\lambda}{c} + \phi(u) \right) - P_{\Omega_g^v} \left(\frac{\lambda}{c} + \phi(u) \right) \right\|_{rms}^2 , \end{aligned} \tag{5.17b}$$

where

$$J(u) = \sum_{i=0}^{k-1} f^0(t_i, u_i, \phi(u)_i) \, \Delta t, \tag{5.17c}$$

and

$$\lambda = (\lambda_0, \ldots, \lambda_{k-1}) \in \mathbb{R}^{kn}. \tag{5.17d}$$

Note that the components of the product set Ω_g are now also Cartesian products with associated RMS projector decomposition formulas. More specifically,

$$\Omega_g^u = \prod_{i=0}^{k-1} U_i \qquad \Omega_g^v = \prod_{i=0}^{k-1} V_i \tag{5.18a}$$

with

$$U_i = \{\xi \in \mathbb{R}^m : \gamma^u(t_i, \xi) \le 0\}, \qquad V_i = \{\eta \in \mathbb{R}^n : \gamma^x(t_i, \eta) \le 0\} \tag{5.18b}$$

for $i = 0, \ldots, k-1$ and

$$P_{\Omega_g^u}(u)_i = P_{U_i}(u_i), \qquad P_{\Omega_g^v}(v)_i = P_{V_i}(v_i) \tag{5.19}$$

for $i = 0, \ldots, k-1$. The decomposition formula for $P_{\Omega_g^u}$ facilitates the implementation of GP methods for the reduced NLP (5.17). On the other hand, the formula for $P_{\Omega_g^v}$ simplifies the construction of $\hat{L}(\lambda, \cdot)$ and its RMS gradient. By rewriting (5.9b) as

$$\hat{L}(\lambda, u) = \sum_{i=0}^{k-1} \hat{l}^0(t_i, \lambda_i, u_i, \phi(u)_i) \, \Delta t \tag{5.20a}$$

with

$$\hat{l}^0(t_i,\lambda_i,u_i,x_i) = \left(f^0(t_i,u_i,x_i) - \frac{1}{2}c\left\|\frac{\lambda_i}{c}\right\|^2\right)$$

$$+\frac{1}{2}c\left\|\left(\frac{\lambda_i}{c}+x_i\right) - P_{V_i}\left(\frac{\lambda_i}{c}+x_i\right)\right\|^2, \tag{5.20b}$$

it becomes evident that the reduced NLP (5.17) is a discrete-time control problem with unconstrained states, pointwise constrained control inputs u_i, running loss functions $\hat{l}^0(t_i,\lambda_i,u_i,x_i)$, and the original state equations (4.3). Moreover, it turns out that $\hat{l}^0$ is smooth enough in (u_i,x_i) to admit the standard efficient adjoint recursive calculation of the RMS gradient of $\hat{L}(\lambda,\cdot)$ (cf. section 8.2 and [17]).

6. General GP convergence rate estimates. Reference [11] provides a comprehensive convergence rate analysis of basic GP methods for constrained minimization problems,

$$\min_{w\in\Omega} F(w), \tag{6.1}$$

with closed convex feasible sets Ω in a real Hilbert space $\mathbb{W}$, and convex Lipschitz continuously Fréchet differentiable cost functions $F:\mathbb{W}\to\mathbb{R}^1$. The convergence rate estimates in [11] are directly applicable to the AGP inner loop calculation outlined in section 5 for discrete-time optimal control problems cast as NLPs in finite-dimensional Euclidean spaces with RMS inner products. In this setting, Ω is a solution set for the control problem's pointwise quasi-convex state/control inequality constraints, and $F(\cdot)$ is either an augmented Lagrangian $L(\lambda,\cdot)$ or a reduced augmented Lagrangian $\hat{L}(\lambda,\cdot)$. The structure of the feasible set is similar in NLP formulations I and II of section 4; however, the augmented Lagrangians in these formulations are quite different, and these differences are strongly manifested in the corresponding GP convergence estimates as the mesh width Δt approaches zero.

In the general context of (6.1), the GP iteration of section 5.1 becomes

$$w^{(j+1)} = P_\Omega(w^{(j)} - \alpha s_j\nabla F(w^{(j)})), \qquad j=0,1,2,\ldots, \tag{6.2a}$$

where α is a fixed positive scaling parameter and s_j is the largest number s such that

$$s\in\{1,\beta,\beta^2,\ldots\} \tag{6.2b}$$

and

$$F(w) - F(P_\Omega(w-\alpha s\nabla F(w))) \geq \delta\langle\nabla F(w), w - P_\Omega(w-\alpha s\nabla F(w))\rangle \tag{6.2c}$$

with $w=w^{(j)}$, and β and δ fixed in $(0,1)$. When the convex function F has a unique minimizer $\overline{w}\in\Omega$, the convergence rate of the value sequences $\{F(w^{(n)}\}$ generated by (6.2) correlates strongly with the gradient Lipschitz norm,

$$\Lambda = \sup_{w^a\neq w^b}\frac{\|\nabla F(w^a)-\nabla F(w^b)\|}{\|w^a-w^b\|}, \tag{6.3}$$

and the growth properties of the nondecreasing nonnegative real function,

$$\gamma(d) = \inf\{F(w)-F(\overline{w}) : w\in\Omega \text{ and } \|w-\overline{w}\|\geq d\} \qquad (d\geq 0). \tag{6.4}$$

Roughly speaking, the convergence rate estimates derived in [11] for $\{F(w^{(n)})\}$ improve as Λ decreases or the growth rate of $\gamma(\cdot)$ increases. One such estimate will suffice for present purposes.

THEOREM 6.1. *Suppose that* Ω *is a nonempty closed convex set in a real Hilbert space* $\mathbb{W}$, *that* F *is a convex Lipschitz continuously Fréchet differentiable real function on* $\mathbb{W}$, *and that* Λ *is the Lipschitz norm of* ∇F *in* (6.3). *In addition, suppose that* F *has a unique minimizer* $\overline{w}$ *in* Ω *and that the corresponding function* $\gamma(\cdot)$ *in* (6.4) *grows quadratically with*

$$\Gamma = \inf_{d>0} \frac{\gamma(d)}{d^2} > 0. \tag{6.5}$$

Let $\{w^{(j)}\}$ *be any sequence in* Ω *generated by the gradient projection method* (6.2). *For* $j = 0, 1, 2, \ldots$, *let* s_j *be the associated Bertsekas–Armijo step length and let* $r_j = F(w^{(j)}) - F(\overline{w}) \geq 0$. *Then* $\{s_j\}$ *is bounded away from zero by*

$$\sigma = \min\left\{\alpha, \frac{2\beta(1-\delta)}{\Lambda}\right\}. \tag{6.6}$$

Moreover, either $r_j = 0$ *eventually or* $\{r_j\}$ *converges geometrically to zero with*

$$0 \leq \frac{r_{j+1}}{r_j} \leq \nu, \qquad j = 0, 1, 2, \ldots, \tag{6.7a}$$

and

$$\nu = 1 - \frac{4\Gamma\sigma\delta}{(\sqrt{1+4\Gamma\sigma}+1)^2}. \tag{6.7b}$$

Proof. Directly from the proof of Theorem 4.3 in [11], and the step length lower bound estimate in [4]. □

Note that in all cases, the convergence factor ν in (6.7) lies in the interval $1-\delta < \nu < 1$. Moreover, if $\alpha \geq \frac{2\beta(1-\delta)}{\Lambda}$, then $\sigma = \frac{2\beta(1-\delta)}{\Lambda}$ and the formula for ν in Theorem 6.1 can be written as

$$\nu = (1-\delta) + \delta\left(\frac{2}{\sqrt{1+\rho}+1}\right) \tag{6.8a}$$

with

$$\rho = 8\beta(1-\delta)\left(\frac{\Gamma}{\Lambda}\right). \tag{6.8b}$$

It is now apparent that if $\alpha \geq \frac{2\beta(1-\delta)}{\Lambda}$, then the convergence factor ν is independent of the scale factor α, and $\nu \to 1^-$ as $\Gamma/\Lambda \to 0^+$. On the other hand, if $\alpha < \frac{2\beta(1-\delta)}{\Lambda}$, then $\sigma = \alpha$ and

$$\nu = 1 - \frac{4\Gamma\alpha\delta}{(\sqrt{1+4\Gamma\alpha}+1)^2},$$

in which case ν is independent of Λ and $\nu \to 1^-$ as $\Gamma\alpha \to 0^+$.

Note also that for convex cost functions on finite-dimensional spaces $\mathbb{W}$, the growth function γ for a unique minimizer $\overline{w}$ is always positive-definite [12]; however, in general, the rate at which γ grows with d depends on the structure of F and Ω. For present purposes, consideration is limited to Lipschitz continuously differentiable cost functions F whose gradients satisfy the *strong monotonicity* condition

$$\exists \mu > 0 \text{ for all } w^a, w^b \in \mathbb{W}, \quad \langle \nabla F(w^a) - \nabla F(w^b), w^a - w^b \rangle \geq \mu \|w^a - w^b\|^2 \tag{6.9}$$

(cf. [30]). Such functions are *strongly convex*, and the following result is readily proved with elementary techniques originally developed by Vainberg, Zarantonello, and Browder et al. (e.g., [30], [7], and the still earlier unpublished and currently inaccessible technical report [31] cited in [7]).

THEOREM 6.2. *Suppose that Ω is a nonempty closed convex set in a real Hilbert space $\mathbb{W}$, that F is a Lipschitz continuously Fréchet differentiable real function on $\mathbb{W}$, and that ∇F satisfies the strong monotonicity condition* (6.9). *Then F is strongly convex in the sense of* [6], *the corresponding convex program* (6.1) *has a unique minimizer $\overline{w}$, and the quadratic growth condition* (6.5) *in Theorem* 6.1 *holds with $\Gamma \geq \frac{1}{2}\mu$.*

Proof. The existence and uniqueness of a minimizer $\overline{w}$ for F in the closed convex set Ω can be established with the contraction mapping principle, as outlined in [7]. For $s > 0$, define the associated map $T_s : \Omega \to \Omega$ by the rule

$$\text{for all } w \in \Omega \quad T_s(w) = P_\Omega(w - s\nabla F(w)).$$

With elementary calculus and the Hilbert space projection theorem, it is now easily seen that the following statements are equivalent for all $\overline{w}$:

(a) $\overline{w}$ is a minimizer of the convex real function F in the nonempty closed convex set Ω.

(b) For all $w \in \Omega$, $\langle \nabla F(\overline{w}), w - \overline{w} \rangle \geq 0$.

(c) For some $s > 0$, $\overline{w} = T_s(\overline{w})$.

(d) For all $s > 0$, $\overline{w} = T_s(\overline{w})$.

Moreover, since the projector map P_Ω is nonexpansive [26], it follows from (6.9) and the Lipschitz continuity of $\nabla F(\cdot)$ that

$$\begin{aligned} \|T_s(w^a) - T_s(w^b)\|^2 &\leq \|w^a - w^b - s(\nabla F(w^a) - \nabla F(w^b))\|^2 \\ &\leq (1 - 2\mu s + \Lambda^2 s^2)\ \|w^a - w^b\|^2 \end{aligned}$$

for all $s > 0$ and all w^a and w^b in Ω, where μ is the positive constant in the strong monotonicity condition (6.9) and Λ is the finite Lipschitz norm for $\nabla F(\cdot)$ in (6.3). Evidentially, T_s is a contraction map for sufficiently small fixed positive values of s. In addition, since $\mathbb{W}$ is complete, the closed set Ω is also complete in the metric induced by $\langle \cdot, \cdot \rangle$. Hence the assertions developed above imply that the maps T_s have a common unique fixed point $\overline{w}$, that $\overline{w}$ is the unique global minimizer of F in Ω, and that $\langle \nabla F(\overline{w}), w - \overline{w} \rangle \geq 0$ for all $w \in \Omega$. Now note that for all w in $\mathbb{W}$,

$$\begin{aligned} F(w) - F(\overline{w}) &= \int_0^1 \frac{d}{dt} F(\overline{w} + t(w - \overline{w}))\ dt \\ &= \int_0^1 \langle \nabla F(\overline{w} + t(w - \overline{w})), w - \overline{w} \rangle\ dt, \end{aligned}$$

and hence for all w in Ω,

$$\begin{aligned} F(w) - F(\overline{w}) &\geq \int_0^1 \langle \nabla F(\overline{w} + t(w - \overline{w})) - \nabla F(\overline{w}), w - \overline{w} \rangle\ dt \\ &\geq \frac{1}{2}\mu\ \|w - \overline{w}\|^2. \qquad \square \end{aligned}$$

Note 2. The Lipschitz continuity hypothesis and strong monotonicity condition (6.9) in Theorem 6.2 automatically hold if F is twice continuously Fréchet differentiable and the spectral sets for the corresponding Hessian operators $\nabla^2 F(w)$ are bounded away from zero and infinity as w ranges over $\mathbb{W}$.

Note 3. The proof strategy employed here for Theorem 6.2 is valid in infinite-dimensional Hilbert spaces; hence this theorem applies to certain continuous-time ODE optimal control problems cast in $\mathbb{L}^2$ function spaces, as well as their finite-dimensional discrete-time RMS approximations. A finite-dimensional counterpart of Theorem 6.2 can be proved with a compactness argument when F is merely continuously differentiable (however, even in finite-dimensional spaces, the GP convergence rate estimate in Theorem 6.1 still requires Lipschitz continuity of $\nabla F(\cdot)$). More specifically, if $\nabla F(\cdot)$ is continuous, then, as in the proof of Theorem 6.2, the strong monotonicity condition (6.9) yields

$$F(w) - F(w_0) \geq \frac{1}{2}\mu \left\| w - w_0 \right\|^2 + \langle \nabla F(w_0), w - w_0 \rangle$$

for all w and w_0. It is now easily seen that every level set of F is bounded and closed, and therefore compact in finite-dimensional normed spaces. But in this case, F must attain its infimum at some $\overline{w}$ in the closed set Ω. Since Ω is convex, it follows that $\langle \nabla F(\overline{w}), w - \overline{w} \rangle \geq 0$ for all w in Ω, and the foregoing estimate applied to the increment $F(w) - F(\overline{w})$ immediately proves uniqueness of $\overline{w}$ and the quadratic growth property (6.5).

7. Mesh-dependent rate estimates in formulation I. Theorem 6.1 yields AGP inner loop convergence factors ν that can approach 1 as Δt goes to zero in the discrete-time formulation I of section 5.3. This does not prove that the *actual* cost value errors $r_j = L(\lambda, w_j) - L(\lambda, \overline{w})$ converge more slowly on increasingly refined meshes, since ν^j provides only an *upper* bound on the ratio r_j/r_0. On the other hand, the numerical results in [17] do demonstrate rapidly deteriorating mesh-dependent convergence behavior in formulation I that is at least qualitatively consistent with the behavior of ν as Δt goes to zero. Accordingly, it seems worthwhile to explore briefly how ν may depend on Δt in formulation I for a representative linear-quadratic curve follower problem with pointwise bounds on the state variable.

Example 1. Let $x^{ref}(\cdot)$ be a fixed continuous real valued function on $[0, 1]$ and let $m = n = r = 1$, $f^0(t, \xi, \eta) = \frac{1}{2}(\xi^2 + (\eta - x^{ref}(t))^2)$, $f(t, \xi, \eta) = \xi$, $\gamma(t, \xi, \eta) = |\eta| - 1$ and $|\eta_0| \leq 1$ in the continuous-time ODE model (2.1)–(2.3). With reference to (5.10b)–(5.10c), the associated discrete-time feasible set and augmented Lagrangian in formulation I are then

$$\Omega_g = \{w = (u, x) : x_0 = \eta_0,\ |x_i| \leq 1, \quad i = 0, \ldots, k-1\} \tag{7.1a}$$

and

$$L(\lambda, w) = \frac{1}{2}\|w - w^{ref}\|^2_{rms} + \frac{1}{2}c\left(\left\|\frac{\lambda}{c} + \mathcal{D}x - \mathcal{B}u\right\|^2_{rms} - \left\|\frac{\lambda}{c}\right\|^2_{rms}\right), \tag{7.1b}$$

where

$$w^{ref} = (0, x^{ref}), \tag{7.1c}$$

$$(\mathcal{B}u)_i = u_i, \tag{7.1d}$$

and

$$(\mathcal{D}x)_i = \frac{x_{i+1} - x_i}{\Delta t} \tag{7.1e}$$

for $i = 0, \ldots, k-2$. It will now be shown that the corresponding growth constant Γ for $L(\lambda, \cdot)$ and Lipschitz norm Λ for $\nabla_w L(\lambda, \cdot)$ satisfy

$$\Gamma \leq \frac{1}{2}(1 + c) \tag{7.2}$$

and

$$\Lambda \geq 1 + 2\left(\frac{c}{\Delta t^2}\right) \tag{7.3}$$

for $0 < \Delta t \leq \frac{1}{2}$ and $\lambda \in \mathbb{R}^{k-1}$. With these estimates and (6.7), it is readily seen that $\nu = 1 - O(\Delta t^2)$. More specifically, if $\alpha \geq 1$ and $c \geq 1$, then

$$\nu \geq 1 - \beta \Delta t^2 \tag{7.4}$$

for all $\Delta t \in (0, 1/2]$, uniformly in λ.

To prove (7.2), fix λ and note that both terms in the augmented Lagrangian (7.1b) are convex and Lipschitz continuously differentiable in the RMS norm, and that the first term satisfies the strong monotonicity condition (6.9) with $\mu = 1$. Hence, $L(\lambda, \cdot)$ satisfies (6.9) with some $\mu \geq 1$, and by Theorem 6.2, must therefore have a unique minimizer $\overline{w}$ in the nonempty closed convex set (7.1a) with an associated quadratic growth constant $\Gamma > 0$ that depends on Δt (i.e., on k). Since the pointwise constraints in (7.1a) restrict x alone, assertions (a)–(d) in the proof of Theorem 6.2 yield the first order optimality condition

$$\nabla_u L(\lambda, \overline{w}) = 0. \tag{7.5}$$

A simple "completion of squares" now gives

$$\begin{aligned} L(\lambda, w) - L(\lambda, \overline{w}) = {} & \langle \nabla_x L(\lambda, \overline{w}), x - \overline{x} \rangle_{rms} \\ & + \frac{1}{2}\|w - \overline{w}\|_{rms}^2 + \frac{1}{2}c\, \|\mathcal{D}(x - \overline{x}) - \mathcal{B}(u - \overline{u})\|_{rms}^2. \end{aligned} \tag{7.6}$$

Moreover, the pairs $(u, \overline{x})$ lie in Ω_g for all $u \in \mathbb{R}^k$, and thus the growth function γ for $L(\lambda, \cdot)$ satisfies

$$\begin{aligned} \gamma(d) &= \inf\{L(\lambda, w) - L(\lambda, \overline{w}) : w \in \Omega_g \text{ and } \|w - \overline{w}\|_{rms} \geq d\} \\ &\leq \inf\{L(\lambda, (u, \overline{x})) - L(\lambda, (\overline{u}, \overline{x})) : u \in \mathbb{R}^k \text{ and } \|u - \overline{u}\|_{rms} \geq d\} \\ &= \inf\left\{\frac{1}{2}\|u - \overline{u}\|_{rms}^2 + \frac{1}{2}c\, \|\mathcal{B}(u - \overline{u})\|_{rms}^2 : u \in \mathbb{R}^k \text{ and } \|u - \overline{u}\|_{rms} \geq d\right\} \\ &\leq \inf\left\{\frac{1}{2}(1 + c)\|u - \overline{u}\|_{rms}^2 : u \in \mathbb{R}^k \text{ and } \|u - \overline{u}\|_{rms} \geq d\right\} \\ &= \frac{1}{2}(1 + c)d^2, \end{aligned} \tag{7.7}$$

in which case

$$\Gamma = \inf_{d > 0} \frac{\gamma(d)}{d^2} \leq \frac{1}{2}(1 + c)$$

for all k and $\lambda \in \mathbb{R}^{k-1}$, as claimed.

To prove (7.3), note that the augmented Lagrangian $L(\lambda, \cdot)$ in (7.1b) has the partial RMS gradients

$$\nabla_u L(\lambda, w) = u - \mathcal{B}^*(\lambda + c(\mathcal{D}x - \mathcal{B}u)) \tag{7.8a}$$

and

$$\nabla_x L(\lambda, w) = x - x^{ref} + \mathcal{D}^*(\lambda + c(\mathcal{D}x - \mathcal{B}u)), \tag{7.8b}$$

where $\mathcal{D}^*$ and $\mathcal{B}^*$ denote RMS adjoints of the bounded linear operators $\mathcal{D}$ and $\mathcal{B}$. Therefore, for all w^a and w^b in $\mathbb{R}^k \oplus \mathbb{R}^k$, Cauchy's inequality produces

$$\begin{aligned} &\|w^a - w^b\|_{rms} \cdot \|\nabla_w L(\lambda, w^a) - \nabla_w L(\lambda, w^b)\|_{rms} \\ &\quad \geq \langle w^a - w^b, \nabla_w L(\lambda, w^a) - \nabla_w L(\lambda, w^b)\rangle_{rms} = \|w^a - w^b\|^2_{rms} \\ &\qquad\qquad + c\, \|\mathcal{D}(x^a - x^b) - \mathcal{B}(u^a - u^b)\|^2_{rms}, \end{aligned} \tag{7.9}$$

and therefore

$$\begin{aligned} \Lambda &= \sup_{w^a \neq w^b} \frac{\|\nabla_w L(\lambda, w^a) - \nabla_w L(\lambda, w^b)\|_{rms}}{\|w^a - w^b\|_{rms}} \\ &\geq 1 + c \sup_{(\xi,\eta)\neq 0} \frac{\|\mathcal{D}\eta - \mathcal{B}\xi\|^2_{rms}}{\|\xi\|^2_{rms} + \|\eta\|^2_{rms}} \\ &\geq 1 + c \left(\sup_{\|\eta\|_{rms}=1} \|\mathcal{D}\eta\|_{rms} \right)^2 . \end{aligned} \tag{7.10}$$

Now construct $\hat{\eta} \in \mathbb{R}^k$ with $\hat{\eta}_i = (-1)^i$ for $i = 0, 1, \ldots k-1$, and consider that

$$\begin{aligned} \|\hat{\eta}\|^2_{rms} &= \sum_{i=0}^{k-1} [(-1)^i]^2\, \Delta t \\ &= \sum_{i=0}^{k-1} \frac{1}{k} \\ &= 1 \end{aligned} \tag{7.11}$$

and

$$\begin{aligned} \|\mathcal{D}\hat{\eta}\|^2_{rms} &= \sum_{i=0}^{k-2} \left(\frac{\hat{\eta}_{i+1} - \hat{\eta}_i}{\Delta t} \right)^2 \Delta t \\ &= \frac{k-1}{k} \left(\frac{2}{\Delta t} \right)^2 \\ &\geq \frac{2}{\Delta t^2} \end{aligned} \tag{7.12}$$

for all $k \geq 2$ and $\lambda \in \mathbb{R}^{k-1}$ (cf. Note 1). Thus,

$$\Lambda \geq 1 + 2 \left(\frac{c}{\Delta t^2} \right) \tag{7.13}$$

for $0 < \Delta t \leq \frac{1}{2}$ and $\lambda \in \mathbb{R}^{k-1}$, as claimed.

Note 4. If c is decreased to keep $c/\Delta t^2$ constant as $\Delta t \to 0$, then the Lipschitz norms Λ in (7.3) need not increase without bound and the AGP inner loop geometric convergence ratios ν in (7.4) may remain bounded away from 1. On the other hand, such reductions in c would quickly destroy convergence of the Hestenes–Powell multiplier iteration

$$\lambda_i^{(j+1)} = \lambda_i^{(j)} + c\left(\frac{x_{i+1}^{(j+1)} - x_i^{(j+1)}}{\Delta t} - f(t_i, u_i^{(j+1)}, x_i^{(j+1)})\right),$$

and other AGP outer loop multiplier update formulas. Note that the c-reduction scheme just described amounts to employing augmented Lagrangian and multiplier update formulas in which λ is replaced by $\hat{\lambda} = \lambda/\Delta t$, c is replaced by $\hat{c} = c/\Delta t^2$, the "rate" form of the Euler discrete-time equality constraint function is replaced by the difference form,

$$h(u,x)_i = x_{i+1} - x_i - f(t_i, u_i, x_i)\,\Delta t,$$

and $\hat{c}$ is kept constant as $\Delta t \to 0$. To maintain adequate AGP outer loop convergence in this scheme, it would be necessary to *increase* $\hat{c}$ as $\Delta t \to 0$, and this would drive the inner loop GP convergence ratios ν toward 1 as before.

The inequality (7.4) for ν immediately yields a corresponding lower bound N_ϵ on the number of GP iterations needed to make the error estimate ν^j in (6.7) smaller than some fixed value $\epsilon \in (0,1]$, namely,

$$\begin{aligned} N_\epsilon &> \frac{|\ln \epsilon|}{|\ln(1-\beta\Delta t^2)|} \\ &\geq \frac{3|\ln\epsilon|}{4\beta}\left(\frac{1}{\Delta t^2}\right) \\ &= \frac{3|\ln\epsilon|}{4\beta}k^2 \end{aligned} \tag{7.14}$$

for $k \geq 2$. Since the cost of computing one GP iterate for discrete-time optimal control problems is proportional to k, the estimate (7.14) suggests that the total number of FLOPS required to make the AGP inner loop errors $r = L(\lambda, w) - L(\lambda, \overline{w})$ smaller than a fixed threshold ϵ may increase like k^3. Direct comparisons with the numerical results in [17] are not possible here since the exact solution for the example in [17] is unknown and it was therefore necessary to terminate inner loop AGP iterations on small computed residuals related to the first order necessary conditions. Nevertheless, the monotonically increasing concave graphs of FLOPS versus k in [17] are again qualitatively like the potential $O(k^3)$ growth admitted by the present analysis.

8. Mesh-invariant rate estimates in formulation II. In the convergence theorems of this section, the discrete-time cost functions $J(\cdot)$ in (4.10a) are required to be RMS strongly convex and Lipschitz continuously differentiable uniformly in the mesh width Δt (i.e., uniformly in the number k of mesh points). Related mesh-invariant RMS strong convexity and gradient Lipschitz continuity conditions can then be deduced for the associated discrete-time augmented Lagrangians (5.12c) and reduced augmented Lagrangians (5.17b). By Theorems 6.1 and 6.2, the Lagrangian properties then support mesh-invariant geometric convergence rate estimates for the related AGP and RAGP inner loop iterations.

The required mesh-invariant properties for J in (4.10a) are implied by a k-uniform boundedness property of the discrete-time state transition operators Θ defined in (4.5)–(4.6) and by certain t-uniform convexity and gradient Lipschitz continuity conditions on the continuous-time running loss functions $f^0(t,\cdot)$, $t \in [0,1]$.

LEMMA 8.1. *For each positive integer $k \geq 2$, let $\Theta : \mathbb{R}^{km} \to \mathbb{R}^{kn}$ be the bounded linear operator in the discrete-time state equation solution* (4.5)–(4.6) *with RMS norm*

$$\|\Theta\|_{rms} = \sup_{\|u\|_{rms}=1} \|\Theta u\|_{rms}. \tag{8.1}$$

In addition, let

$$a = \sup_{t\in[0,1]} \|A(t)\| < \infty, \qquad b = \sup_{t\in[0,1]} \|B(t)\| < \infty, \tag{8.2}$$

where $A(t) : \mathbb{R}^n \to \mathbb{R}^n$ and $B(t) : \mathbb{R}^m \to \mathbb{R}^n$ are the uniformly bounded coefficient operators in the continuous-time rate functions (2.4) *and $\|\cdot\|$ denotes the associated Euclidean operator norm in either case. Then*

$$\text{for all } k \geq 2, \qquad \|\Theta\|_{rms} \leq b \ \exp a. \tag{8.3}$$

Proof. By a straightforward application of (4.5)–(4.7), Cauchy's inequality, the formula $\sum_{i=1}^{n}, \ i = n(n+1)/2$, and the estimate, $(1 + a/k)^k < \exp a$ for $a > 0$ and all positive integers k. □

LEMMA 8.2. *For $t \in [0,1]$, let $f^0(t,\cdot) : \mathbb{R}^m \oplus \mathbb{R}^n \to \mathbb{R}^1$ be the running loss functions in* (4.10a) *for the discrete-time cost functions $J(\cdot) : \mathbb{R}^{km} \to \mathbb{R}^1$ in formulation* II. *Put $\zeta = (\xi, \eta) \in \mathbb{R}^m \oplus \mathbb{R}^n$ and let $\nabla_\zeta f^0(t,\zeta)$ denote the standard Euclidean gradient of $f^0(t,\cdot)$ at ζ. Assume that there are numbers $\mu_0 > 0$ and $\Lambda_0 \geq 0$ such that for all $t \in [0,1]$ and all ζ^a and ζ^b in $\mathbb{R}^m \oplus \mathbb{R}^n$,*

$$\langle \nabla_\zeta f^0(t,\zeta^a) - \nabla_\zeta f^0(t,\zeta^b), \zeta^a - \zeta^b \rangle \geq \mu_0 \ \|\xi^a - \xi^b\|^2, \tag{8.4}$$

and

$$\|\nabla_\zeta f^0(t,\zeta^a) - \nabla_\zeta f^0(t,\zeta^b)\| \leq \Lambda_0 \ \|\zeta^a - \zeta^b\|, \tag{8.5}$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product on $\mathbb{R}^m \oplus \mathbb{R}^n$, and $\|\cdot\|$ signifies the norm on $\mathbb{R}^m$ and $\mathbb{R}^m \oplus \mathbb{R}^n$ in (8.4) *and* (8.5), *respectively. Then $J(\cdot)$ is strongly convex and Lipschitz continuously differentiable in the RMS sense,* uniformly for all positive integers $k \geq 2$. *More precisely, there are numbers $\mu_J > 0$ and $\Lambda_J \geq 0$ such that for all $k \geq 2$ and all u^a and u^b in $\mathbb{R}^{km}$,*

$$\langle \nabla J(u^a) - \nabla J(u^b), u^a - u^b \rangle_{rms} \geq \mu_J \ \|u^a - u^b\|_{rms}^2 \tag{8.6}$$

and

$$\|\nabla J(u^a) - \nabla J(u^b)\|_{rms} \leq \Lambda_J \ \|u^a - u^b\|_{rms}. \tag{8.7}$$

Proof. Let $\theta \in \mathbb{R}^{kn}$ and $\Theta : \mathbb{R}^{km} \to \mathbb{R}^{kn}$ be the vector and bounded linear operator in the discrete-time state equation solution formula (4.5)–(4.6). For $u \in \mathbb{R}^{km}$, define $\alpha_u(u) \in \mathbb{R}^{km}$ and $\alpha_x(u) \in \mathbb{R}^{kn}$ by the rules

$$\alpha_u(u)_i = \nabla_\xi f^0(t_i, u_i, \phi(u)_i), \qquad \alpha_x(u)_i = \nabla_\eta f^0(t_i, u_i, \phi(u)_i) \tag{8.8}$$

for $i = 0, \dots, k-1$, where $\nabla_\xi f^0$ and $\nabla_\eta f^0$ are the (ξ, η) components of $\nabla_\zeta f^0$. In addition, let $\Theta^* : \mathbb{R}^{kn} \to \mathbb{R}^{km}$ denote the RMS bounded linear adjoint operator for Θ, i.e., for all $x \in \mathbb{R}^{kn}$ and $u \in \mathbb{R}^{km}$,

$$\langle x, \Theta u \rangle_{rms} = \langle \Theta^* x, u \rangle_{rms}, \tag{8.9}$$

where $\langle \cdot, \cdot \rangle_{rms}$ signifies the RMS inner product in $\mathbb{R}^{kn}$ and $\mathbb{R}^{km}$ on the left and right sides of (8.9), respectively. Then for all $u \in \mathbb{R}^{km}$, the corresponding RMS gradient of $J(\cdot)$ is readily shown to be

$$\nabla J(u) = \alpha_u(u) + \Theta^* \alpha_x(u). \tag{8.10}$$

Hence for all u^a and u^b in $\mathbb{R}^{km}$, conditions (8.4) and (8.9) imply that

$$\begin{aligned}\langle \nabla J(u^a) - \nabla J(u^b), u^a - u^b \rangle_{rms} &\geq \langle \alpha_u(u^a) - \alpha_u(u^b), u^a - u^b \rangle_{rms} \\ &\qquad + \langle \alpha_x(u^a) - \alpha_x(u^b), \Theta(u^a - u^b) \rangle_{rms} \\ &= \langle \alpha_u(u^a) - \alpha_u(u^b), u^a - u^b \rangle_{rms} \\ &\qquad + \langle \alpha_x(u^a) - \alpha_x(u^b), \phi(u^a) - \phi(u^b) \rangle_{rms} \\ &\geq \mu_0 \, \|u^a - u^b\|^2_{rms}.\end{aligned}$$

Furthermore, since it is known that $\|\Theta^*\|_{rms} = \|\Theta\|_{rms}$, condition (8.5) yields

$$\begin{aligned}\|\nabla J(u^a) - \nabla J(u^b)\|^2_{rms} &\leq 2(\|\alpha_u(u^a) - \alpha_u(u^b)\|^2 + \|\Theta^*(\alpha_x(u^a) - \alpha_x(u^b))\|^2_{rms}) \\ &\leq 2 \ \max\{1, \|\Theta^*\|^2_{rms}\}(\|\alpha_u(u^a) - \alpha_u(u^b)\|^2_{rms} \\ &\qquad + \|\alpha_x(u^a) - \alpha_x(u^b)\|^2_{rms}) \\ &\leq 2 \ \max\{1, \|\Theta\|^2_{rms}\} \Lambda_0^2 \ (\|u^a - u^b\|^2_{rms} \\ &\qquad + \|\phi(u^a) - \phi(u^b)\|^2_{rms}) \\ &= 2 \ \max\{1, \|\Theta\|^2_{rms}\} \Lambda_0^2 \ (\|u^a - u^b\|^2_{rms} \\ &\qquad + \|\Theta(u^a - u^b)\|^2_{rms}) \\ &\leq 2 \ \max\{1, \|\Theta\|^2_{rms}\}(1 + \|\Theta\|^2_{rms}) \Lambda_0^2 \ \|u^a - u^b\|^2_{rms}.\end{aligned}$$

With reference to Lemma 8.1, these estimates now give (8.6) and (8.7) with $\mu_J = \mu_0$ and $\Lambda_J = \Lambda_0 \sqrt{2 \max\{1, b^2 \exp 2a\}(1 + b^2 \exp 2a)}$. □

Note 5. The t-uniform convexity and Lipschitz continuity conditions (8.4) and (8.5) are satisfied by the prototype autonomous LQR loss function

$$f^0(t, \xi, \eta) = \frac{1}{2} \ \langle \xi, S\xi \rangle + \frac{1}{2} \ \langle \eta, Q\eta \rangle$$

when $S : \mathbb{R}^m \to \mathbb{R}^m$ is a symmetric positive-definite bounded linear operator and $Q : \mathbb{R}^n \to \mathbb{R}^n$ is a symmetric positive-semidefinite bounded linear operator.

Note 6. The linear operator Θ in the continuous-time state equation solution (3.6) is bounded in the $\mathbb{L}^2$ sense, and an $\mathbb{L}^2$ counterpart of Lemma 8.2 can therefore be proved for the continuous-time cost function $J(\cdot)$ in (3.4a) of formulation II, provided that measurabilty restrictions of the Carathéodory type [8], [1] for $f^0(\cdot, \zeta)$ are added to the hypotheses on $f^0(t, \cdot)$ in (8.4) and (8.5).

8.1. Unseparated pointwise constraints. The following key lemma establishes (k, λ)-uniform RMS strong convexity and Lipschitz continuity results for the discrete-time augmented Lagrangians in formulation II. These results lead directly to a mesh-invariant and λ-invariant AGP inner loop convergence rate estimate in closed convex feasible sets $\Omega_g \subset \mathbb{R}^{km} \oplus \mathbb{R}^{kn}$ prescribed by general unseparated continuous quasi-convex inequality constraint functions $\gamma(t,\cdot) : \mathbb{R}^m \oplus \mathbb{R}^n \to \mathbb{R}^r$.

LEMMA 8.3. *Assume that the discrete-time cost functions* $J(\cdot) : \mathbb{R}^{km} \to \mathbb{R}^1$ *defined by* (4.10*a*) *in formulation* II *are convex, with gradients that satisfy the* k*-uniform RMS strong monotonicity and Lipschitz continuity conditions* (8.6) *and* (8.7). *Then the corresponding augmented Lagrangians* $L(\lambda,\cdot) : \mathbb{R}^{km} \oplus \mathbb{R}^{kn} \to \mathbb{R}^1$ *in* (5.12c) *satisfy analogous conditions uniformly for* $k \geq 2$ *and* $\lambda \in \mathbb{R}^{kn}$. *More precisely, for all* $k \geq 2$, $\lambda \in \mathbb{R}^{kn}$, *and* w^a *and* w^b *in* $\mathbb{R}^{km} \oplus \mathbb{R}^{kn}$,

$$\langle \nabla L(\lambda, w^a) - \nabla L(\lambda, w^b), w^a - w^b \rangle_{rms} \geq \mu_L \, \|w^a - w^b\|^2_{rms} \tag{8.11a}$$

with

$$\mu_L = \frac{c\,\mu_J}{\mu_J + c\,(1 + b^2 \exp 2a)} \tag{8.11b}$$

and

$$\|\nabla L(\lambda, w^a) - \nabla L(\lambda, w^b)\|_{rms} \leq \Lambda_L \, \|w^a - w^b\|_{rms} \tag{8.12a}$$

with

$$\Lambda_L = \Lambda_J + c\,\sqrt{2 \max\{1, b^2 \exp 2a\}(1 + b^2 \exp 2a)}, \tag{8.12b}$$

where a *and* b *are the coefficient bounds in Lemma* 8.1.

Proof. Let $F(\lambda, u, v)$ denote the shifted penalty term in the augmented Lagrangians (5.12c), i.e.,

$$F(\lambda, u, v) = \frac{1}{2} c \, \left\| \frac{\lambda}{c} + \Theta u + \theta - v \right\|^2_{rms}. \tag{8.13}$$

Then for all $w = (u, v) \in \mathbb{R}^{km} \oplus \mathbb{R}^{kn}$,

$$\nabla_w L(\lambda, w) = (\nabla J(u) + \nabla_u F(\lambda, u, v) \, , \, \nabla_v F(\lambda, u, v)) \tag{8.14a}$$

with

$$\nabla_u F(\lambda, u, v) = c\,\Theta^* \, \left(\frac{\lambda}{c} + \Theta u + \theta - v \right), \tag{8.14b}$$

$$\nabla_v F(\lambda, u, v) = -c \, \left(\frac{\lambda}{c} + \Theta u + \theta - v \right). \tag{8.14c}$$

Hence,

$$\begin{aligned} &\langle \nabla_w L(\lambda, w^a) - \nabla_w L(\lambda, w^b), w^a - w^b \rangle_{rms} \geq \mu_J \|u^a - u^b\|^2_{rms} \\ &+c\,(\|\Theta(u^a - u^b)\|^2_{rms} - 2\|\Theta(u^a - u^b)\|_{rms}\|v^a - v^b\|_{rms} + \|v^a - v^b\|^2_{rms}). \end{aligned} \tag{8.15}$$

Let $\xi = u^a - u^b$ and $\eta = v^a - v^b$, and observe that $\|\Theta\xi\|_{rms} \leq \|\Theta\|_{rms}\|\xi\|_{rms}$. Then for any fixed $\epsilon > 0$, the right side of the inequality (8.15) is bounded below by

$$(\mu_J - \epsilon c\, \|\Theta\|^2_{rms})\|\xi\|^2_{rms} + c\,((1 + \epsilon)\|\Theta\xi\|^2_{rms} - 2\|\Theta\xi\|_{rms}\|\eta\|_{rms} + \|\eta\|^2_{rms}). \tag{8.16}$$

Consider that the polynomial $(1+\epsilon)z^2 - 2z + 1$ has the minimum value $\epsilon/(1+\epsilon)$ in $\mathbb{R}^1$. This fact and the k-uniform upper bound for $\|\Theta\|_{rms}$ in Lemma 8.1 proves that the expression in (8.16) is itself bounded below by

$$(\mu_J - \epsilon c\, b^2 \exp 2a)\|\xi\|_{rms}^2 + \frac{c\,\epsilon}{1+\epsilon}\|\eta\|_{rms}^2 \tag{8.17}$$

for all $\epsilon > 0$, $k \geq 2$ and all (ξ, η) in $\mathbb{R}^{km} \oplus \mathbb{R}^{kn}$. The estimate (8.11) is now obtained by verifying that

$$\max_{\epsilon>0}\ \min\left\{\mu_J - \epsilon c\, b^2 \exp 2a, \frac{c\,\epsilon}{1+\epsilon}\right\} = \frac{c\,\mu_J}{\frac{\mu_J + c\,(1+b^2\exp 2a)}{2} + \sqrt{(\frac{\mu_J + c\,(1+b^2\exp 2a)}{2})^2 - c\,\mu_J}}$$
$$\geq \frac{c\,\mu_J}{\mu_J + c\,(1+b^2\exp 2a)}.$$

To prove (8.12), note that by (8.14)

$$\begin{aligned}\|\nabla_w L(\lambda, w^a) - \nabla_w L(\lambda, w^b)\|_{rms} &\leq \Lambda_J \|u^a - u^b\|_{rms} \\ &\quad + c\,\sqrt{1 + \|\Theta^*\|_{rms}^2}\|\Theta(u^a - u^b) - (v^a - v^b)\|_{rms}.\end{aligned} \tag{8.18}$$

Moreover, since $\|\Theta^*\|_{rms} = \|\Theta\|_{rms}$ and

$$\|\Theta(u^a - u^b) - (v^a - v^b)\|_{rms}^2 \leq 2(\|\Theta\|_{rms}^2\|u^a - u^b\|_{rms}^2 + \|v^a - v^b\|_{rms}^2),$$

it now follows from Lemma 8.1 that the right side of the inequality (8.18) is bounded above by

$$[\Lambda_J + c\,\sqrt{2\max\{1, b^2\exp 2a\}(1 + b^2\exp 2a)}\ ]\ \|w^a - w^b\|_{rms}$$

for all $k \geq 2$, $\lambda \in \mathbb{R}^{kn}$, and w^a and w^b in $\mathbb{R}^{km} \oplus \mathbb{R}^{kn}$. □

THEOREM 8.4. *Assume that the discrete-time cost functions $J(\cdot) : \mathbb{R}^{km} \to \mathbb{R}^1$ defined by* (4.10a) *in formulation II are convex with gradients that satisfy the k-uniform RMS strong monotonicity and Lipschitz continuity conditions* (8.6) *and* (8.7). *Suppose that the inequality constraint functions $\gamma(t, \cdot) : \mathbb{R}^m \oplus \mathbb{R}^n \to \mathbb{R}^r$ that prescribe the relaxed feasible sets $\Omega_g \subset \mathbb{R}^{km} \oplus \mathbb{R}^{kn}$ in* (5.12b) *are continuous and quasi-convex for all $t \in [0,1]$. Let $L(\lambda, \cdot) : \mathbb{R}^{km} \oplus \mathbb{R}^{kn} \to \mathbb{R}^1$ denote the augmented Lagrangians in* (5.12c). *Then for each $k \geq 2$ and $\lambda \in \mathbb{R}^{kn}$, the relaxed NLP* (5.12) *has a corresponding unique solution $\overline{w}(\lambda)$. Moreover, let $\{w^{(j)}\}$ be any sequence in Ω_g generated by the AGP inner loop iteration* (5.6)–(5.7), *and let $r_j = L(\lambda, w^{(j)}) - L(\lambda, \overline{w}(\lambda))$. Then for all $k \geq 2$ and $\lambda \in \mathbb{R}^{kn}$, the sequence $\{r_j\}$ converges geometrically to zero with*

$$0 \leq \frac{r_j}{r_0} \leq (\nu_L)^j, \qquad j = 0, 1, 2, \ldots, \tag{8.19a}$$

where

$$\nu_L = 1 - \frac{2\mu_L\sigma_L\delta}{(\sqrt{1 + 2\mu_L\sigma_L} + 1)^2}, \tag{8.19b}$$

$$\sigma_L = \min\left\{\alpha, \frac{2\beta(1-\delta)}{\Lambda_L}\right\}, \tag{8.19c}$$

and μ_L and Λ_L are the (k, λ)-invariant numbers in Lemma 8.3.

Proof. Immediate from Lemma 8.3, and Theorems 6.1 and 6.2. □

8.2. Separated pointwise constraints. An RAGP counterpart of Theorem 8.4 is established below for problems with separated inequality constraint functions $\gamma(t,\cdot) = (\gamma^u(t,\cdot), \gamma^x(t,\cdot))$ that have continuous quasi-convex components $\gamma^u(t,\cdot) : \mathbb{R}^m \to \mathbb{R}^{r_u}$ and $\gamma^u(t,\cdot) : \mathbb{R}^n \to \mathbb{R}^{r_x}$ for $t \in [0,1]$. The following lemmas are needed in this development.

LEMMA 8.5. *Let Y be a nonempty convex set in $\mathbb{R}^N$ that is closed relative to a norm $\|\cdot\|$ induced by some inner product $\langle\cdot,\cdot\rangle$. Define $g : \mathbb{R}^N \oplus \mathbb{R}^N \to \mathbb{R}^1$ and $\hat{g} : \mathbb{R}^N \to \mathbb{R}^1$ by the rules*

$$g(x,y) = \frac{1}{2}\,\|x-y\|^2$$

and

$$\hat{g} = \min_{y\in Y}\; g(x,y),$$

and let $P_Y(x)$ denote the unique minimizer of $g(x,\cdot)$ in Y. Then $\hat{g}(\cdot)$ is Lipschitz continuously differentiable relative to $\|\cdot\|$, and the formula

$$\nabla\hat{g}(x) = \nabla_x g(x, P_Y) = x - P_Y(x) \tag{8.20}$$

prescribes the gradient of $\hat{g}(\cdot)$ relative to the inner product $\langle\cdot,\cdot\rangle$.

Proof. Since $\{\mathbb{R}^N, \langle\cdot,\cdot\rangle, \|\cdot\|\}$ is a Hilbert space, the projection theorem establishes the existence and uniqueness of the proximal point map P_Y. If the closed set Y is bounded and hence compact, then the assertions of the lemma follow directly from results in [9] and [10]. On the other hand, suppose that Y is not bounded. Fix $x_0 \in \mathbb{R}^N$ and $\delta_0 > 0$, and put $\rho_0 = \|x_0 - P_Y(x_0)\| + \delta_0$. By the triangle inequality and the nonexpansive property for P_Y [26], it follows that for $\|x - x_0\| < \delta_0$,

$$\begin{aligned}\|x_0 - P_Y(x)\| &\le \|x_0 - P_Y(x_0)\| + \|P_Y(x_0) - P_Y(x)\|\\ &\le \|x_0 - P_Y(x_0)\| + \|x_0 - x\|\\ &\le \rho_0\end{aligned}$$

and therefore

$$\hat{g}(x) = \min_{y\in Y(\rho_0)}\; g(x,y),$$

where $Y(\rho_0)$ is the compact set $\{y \in Y : \|y - x_0\| \le \rho_0\}$. But in this case, the assertions of the lemma follow again from trivial adjustments to the proofs in [9] and [10].

Finally, note that if (8.20) holds, then $\nabla\hat{g}(\cdot)$ is Lipschitz continuous with

$$\begin{aligned}\|\nabla\hat{g}(x_1) - \nabla\hat{g}(x_2)\| &\le \|x_1 - x_2\| + \|P_Y(x_1) - P_Y(x_2)\|\\ &\le 2\,\|x_1 - x_2\|\end{aligned} \tag{8.21}$$

for all x_1 and x_2. □

LEMMA 8.6. *Suppose that the relaxed feasible sets $\Omega_g \subset \mathbb{R}^{km} \oplus \mathbb{R}^{kn}$ are Cartesian products $\Omega_g^u \times \Omega_g^v$ prescribed by separated continuous quasi-convex inequality constraint functions in* (5.15). *Assume that the discrete-time cost functions $J(\cdot) : \mathbb{R}^{km} \to \mathbb{R}^1$ defined by equation* (4.10a) *in formulation* II *are convex with gradients that satisfy the k-uniform RMS strong monotonicity and Lipschitz continuity conditions* (8.6) *and*

(8.7). *Then the corresponding reduced augmented Lagrangians* $\hat{L}(\lambda,\cdot):\mathbb{R}^{km}\to\mathbb{R}^1$ *in* (5.17b) *satisfy analogous conditions uniformly for* $k\geq 2$ *and* $\lambda\in\mathbb{R}^{kn}$*. More precisely, for all* $k\geq 2$*,* $\lambda\in\mathbb{R}^{kn}$*, and* u^a *and* u^b *in* $\mathbb{R}^{km}$*,*

$$\langle\nabla\hat{L}(\lambda,u^a)-\nabla\hat{L}(\lambda,u^b),u^a-u^b\rangle_{rms}\geq\mu_{\hat{L}}\;\|u^a-u^b\|^2_{rms}\tag{8.22a}$$

with

$$\mu_{\hat{L}}=\mu_J\tag{8.22b}$$

and

$$\|\nabla\hat{L}(\lambda,u^a)-\nabla\hat{L}(\lambda,u^b)\|_{rms}\leq\Lambda_{\hat{L}}\;\|u^a-u^b\|_{rms}\tag{8.23a}$$

with

$$\Lambda_{\hat{L}}=\Lambda_J+\sqrt{2}cb^2\exp 2a.\tag{8.23b}$$

Proof. For $\lambda\in\mathbb{R}^{kn}$ and $u\in\mathbb{R}^{km}$, let

$$\hat{F}(\lambda,u)=\min_{v\in\Omega_g^v}\;F(\lambda,u,v),$$

where $F(\lambda,u,v)$ is the shifted penalty term (8.13) in the augmented Lagrangians (5.12c). Then

$$\hat{F}(\lambda,u)=J(u)+\hat{F}(\lambda,u)-\frac{1}{2}c\;\left\|\frac{\lambda}{c}\right\|^2_{rms}.\tag{8.24}$$

By Lemma 8.5 and the chain rule, $\hat{F}(\lambda,\cdot)$ is Lipschitz continuously differentiable with

$$\nabla_u\hat{F}(\lambda,u)=c\;\Theta^*\left(\frac{\lambda}{c}+\Theta u+\theta-P_{\Omega_g^v}\left(\frac{\lambda}{c}+\Theta u+\theta\right)\right).\tag{8.25}$$

For u^a and u^b in $\mathbb{R}^{km}$, put $x^a=\frac{\lambda}{c}+\Theta u^a+\theta$ and $x^b=\frac{\lambda}{c}+\Theta u^b+\theta$. Since $P_{\Omega_g^v}$ is nonexpansive and monotone nondecreasing [26], [32], and $\|\Theta^*\|_{rms}=\|\Theta\|_{rms}$, it follows from (8.25) and Lemma 8.1 that for all $k\geq 2$, $\lambda\in\mathbb{R}^{km}$, and u^a and u^b in $\mathbb{R}^{km}$,

$$\begin{aligned}&\langle\nabla_u\hat{F}(\lambda,u^a)-\nabla_u\hat{F}(\lambda,u^b),u^a-u^b\rangle_{rms}=c\;\|\Theta(u^a-u^b)\|^2_{rms}\\&-c\;\langle P_{\Omega_g^v}(x^a)-P_{\Omega_g^v}(x^b),\Theta(u^a-u^b)\rangle_{rms}\geq c\;\|\Theta(u^a-u^b)\|^2_{rms}\\&-\|P_{\Omega_g^v}(x^a)-P_{\Omega_g^v}(x^b)\|_{rms}\;\|\Theta(u^a-u^b)\|_{rms}\geq 0\end{aligned}\tag{8.26}$$

and

$$\begin{aligned}\|\nabla_u\hat{F}(\lambda,u^a)-\nabla_u\hat{F}(\lambda,u^b)\|^2_{rms}\leq{}&c^2\|\Theta^*\|^2_{rms}(\|\Theta(u^a-u^b)\|^2_{rms}\\&-2\langle x^a-x^b,P_{\Omega_g^v}(x^a)-P_{\Omega_g^v}(x^b)\rangle\\&+\|P_{\Omega_g^v}(x^a)-P_{\Omega_g^v}(x^b)\|^2_{rms})\\\leq{}&2c^2(b^2\exp 2a)^2\|u^a-u^b\|^2_{rms}.\end{aligned}\tag{8.27}$$

The (k,λ)-uniform estimates (8.22) and (8.23) are immediate consequences of (8.26), (8.27), (8.24), and the hypotheses for $J(\cdot)$. □

THEOREM 8.7. *Suppose that the relaxed feasible sets* $\Omega_g \subset \mathbb{R}^{km} \oplus \mathbb{R}^{kn}$ *are Cartesian products* $\Omega_g^u \times \Omega_g^v$ *prescribed by separated continuous quasi-convex inequality constraint functions in* (5.16)*. Assume that the discrete-time cost functions* $J(\cdot) : \mathbb{R}^{km} \to \mathbb{R}^1$ *defined by* (4.10a) *in formulation* II *are convex with gradients that satisfy the k-uniform RMS strong monotonicity and Lipschitz continuity conditions* (8.6) *and* (8.7)*. Let* $\hat{L}(\lambda, \cdot) : \mathbb{R}^{km} \to \mathbb{R}^1$ *denote the reduced augmented Lagrangians in* (5.17b)*. Then for each* $k \geq 2$ *and* $\lambda \in \mathbb{R}^{kn}$*, the reduced NLP* (5.17) *has a corresponding unique solution* $\overline{u}(\lambda)$*. Moreover, let* $\{u^{(j)}\}$ *be any sequence in* Ω_g^u *generated by the RAGP counterpart of the inner loop iteration* (5.6)–(5.7) *for* (5.17) *and let* $r_j = \hat{L}(\lambda, u^{(j)}) - \hat{L}(\lambda, \overline{u}(\lambda))$*. Then for all* $k \geq 2$ *and* $\lambda \in \mathbb{R}^{kn}$*, the sequence* $\{r_j\}$ *converges geometrically to zero with*

$$0 \leq \frac{r_j}{r_0} \leq (\nu_{\hat{L}})^j, \qquad j = 0, 1, 2, \ldots, \tag{8.28a}$$

where

$$\nu_{\hat{L}} = 1 - \frac{2\mu_{\hat{L}}\sigma_{\hat{L}}\delta}{(\sqrt{1 + 2\mu_{\hat{L}}\sigma_{\hat{L}}} + 1)^2}, \tag{8.28b}$$

$$\sigma_{\hat{L}} = \min\left\{\alpha, \frac{2\beta(1-\delta)}{\Lambda_{\hat{L}}}\right\}, \tag{8.28c}$$

and $\mu_{\hat{L}}$ *and* $\Lambda_{\hat{L}}$ *are the* (k, λ)*-invariant constants in Lemma* 8.6.

Proof. Immediate from Lemma 8.6 and Theorems 6.1 and 6.2. □

Note 7. As $c \to \infty$, the growth constants μ_L and $\mu_{\hat{L}}$ in Theorems 8.4 and 8.7 remain bounded, the Lipschitz constants Λ_L and $\Lambda_{\hat{L}}$ increase without bound, and the corresponding (k, λ)-invariant AGP and RAGP inner loop convergence rate factors ν_L and $\nu_{\hat{L}}$ approach 1. Moreover, this deterioration in convergence rate bounds is reflected in the actual convergence rate properties of the subject algorithms in formulation II. Augmented Lagrangian methods in formulation I are also adversely affected by large values of c, although the estimates in section 7 are too crude to show this. In fact, such large penalty constant ill-conditioning is a familiar feature of unscaled augmented Lagrangian methods applied to general NLPs, and preconditioning schemes designed specifically to address this issue for penalized objective functions have been proposed in [25], [20], and [21]. In the present optimal control context, the implementation of such methods remain to be explored. (See the final paragraph in section 1 for additional comments on the application of these preconditioners to control problems in formulation I.) Note that the estimates in Theorems 8.4 and 8.7 also predict deteriorating convergence rates for the AGP and RAGP methods when the state transition operator norms $\|\Theta\|_{rms}$ are large compared to 1.

Note 8. For optimal control problems with separated simple pointwise inequality constraints (e.g., bounds on the components of the state and control vectors), the RAGP iteration map typically costs significantly less to compute than its AGP counterpart. What is more, the estimates in Theorems 8.4 and 8.7 indicate that the RAGP method may have a convergence rate advantage over AGP as well, in cases where both methods are applicable. To see this, observe that

$$\begin{aligned}\Lambda_L &= \Lambda_J + c\sqrt{2\ \max\{1, b^2 \exp 2a\}(1 + b^2 \exp 2a)} \\ &\geq \Lambda_J + \sqrt{2}cb^2 \exp 2a \\ &= \Lambda_{\hat{L}},\end{aligned}$$

and therefore

$$\begin{aligned}\sigma_L &= \min\left\{\alpha, \frac{2\beta(1-\delta)}{\Lambda_L}\right\}\\ &\le \min\left\{\alpha, \frac{2\beta(1-\delta)}{\Lambda_{\hat{L}}}\right\}\\ &= \sigma_{\hat{L}}.\end{aligned}$$

Moreover,

$$\begin{aligned}\mu_L &= \frac{c\,\mu_J}{\mu_J + c\,(1 + b^2 \exp 2a)}\\ &< \frac{\mu_J}{1 + b^2 \exp 2a}\\ &= \frac{\mu_{\hat{L}}}{1 + b^2 \exp 2a}.\end{aligned}$$

Consequently,

$$\mu_L \sigma_L < \frac{\mu_{\hat{L}} \sigma_{\hat{L}}}{1 + b^2 \exp 2a},$$

and therefore $\nu_{\hat{L}} < \nu_L < 1$. It remains to be shown that this disparity in convergence rate *estimates* reflects the actual convergence behavior observed in numerical implementations of AGP and RAGP methods for the class of optimal control problems treated here.

Note 9. Reference [19] describes effective alternative primal/dual Lagrangian algorithms for specially structured fine-mesh ill-conditioned discrete-time optimal control problems.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Plenum Publishing Co., New York, 1987.

[2] E. L. Allgower, K. Böhmer, F. A. Potra, and W. C. Rheinboldt, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.

[3] M. Aljazzaf, *Multiplier Methods with Partial Elimination of Constraints for Nonlinear Programming,* Ph.D. thesis, North Carolina State University, Raleigh, NC, 1990.

[4] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automatic Control, 21 (1976), pp. 174–184.

[5] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods,* Academic Press, New York, 1982

[6] D. P. Bertsekas, *Nonlinear Programming,* Athena Scientific, Belmont, MA, 1995.

[7] F. E. Browder and W. V. Petryshyn, *Construction of fixed points of nonlinear mappings in Hilbert space*, J. Math Anal. Appl., 20 (1967), pp. 197-228.

[8] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations,* McGraw–Hill, New York, 1955.

[9] J. M. Danskin, *Theory of min-max with applications*, SIAM J. Appl. Math., 14 (1966), pp. 641–644.

[10] J. C. Dunn, *On the classification of singular and nonsingular extremals for the Pontryagin maximum principle*, J. Math. Anal. Appl., 17 (1967), pp. 1–36.

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

[12] J. C. Dunn, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–216.

[13] J. C. Dunn, *A projected Newton method for minimization problems with nonlinear inequality constraints*, Numer. Math., 53 (1988), pp. 377–409.

[14] J.C. Dunn and D. P. Bertsekas, *Efficient dynamic programming implementations of Newton's method for unconstrained optimal control problems*, J. Optim. Theory Appl., 63 (1989), pp. 23–38.

[15] J. C. Dunn *Formal augmented Newtonian projection methods for continuous-time optimal control problems*, in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, December 13–15, 1989, pp. 374–377.

[16] J. C. Dunn, *On* L^2 *sufficient conditions and the gradient projection method for optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1270–1290.

[17] J. C. Dunn, *Augmented gradient projection calculations for regulator problems with pointwise state and control constraints*, in Optimal Control: Theory, Algorithms, and Applications, W. W. Hager and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[18] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[19] W. W. Hager and G. D. Ianculescu, *Dual approximations in optimal control*, SIAM J. Control Optim., 22 (1984), pp. 423–465.

[20] W. W. Hager, *Dual techniques for constrained optimization*, J. Optim. Theory Appl., 55 (1987), pp. 37–71.

[21] W. W. Hager, *Analysis and implementation of a dual algorithm for constrained optimization*, J. Optim. Theory Appl., 79 (1993), pp. 427–462.

[22] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.

[23] C. T. Kelley and E. W. Sachs, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477–493.

[24] E. S. Levitin and B. T. Poljak, *Constrained minimization problems*, USSR Comp. Math. Phys., 6 (1966), pp. 1–50.

[25] D. G. Luenberger, Linear and Nonlinear Programming, 2nd ed., Addison-Wesley, Reading, MA, 1984.

[26] R. R. Phelps, *Convex sets and nearest points*, Proc. Amer. Math. Soc., 8 (1957), pp. 790–797.

[27] M. J. D. Powell, *A method for nonlinear constraints in minimizing problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.

[28] R. T. Rockafellar, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555–562.

[29] T. Tian and J. C. Dunn. *On the gradient projection method for optimal control problems with nonnegative* $\mathbb{L}^2$ *inputs*, SIAM J. Control Optim., 32 (1994), pp. 516-537.

[30] M. M. Vainberg, *Variational Method and Method of Monotone Operators in the Theory of Nonlinear Equations*, John Wiley and Sons, New York, 1973.

[31] E. H. Zarantonello, *Solving Functional Equations by Contractive Averaging*, Math. Research Center Tech. Report 160, Madison, WI, 1960.

[32] E. H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 237–341.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1112–1135

ON A CLASS OF DIFFEOMORPHIC MATCHING PROBLEMS IN ONE DIMENSION*

ALAIN TROUVÉ[†] AND LAURENT YOUNES[‡]

Abstract. We study a class of functional which can be used for matching objects which can be represented as mappings from a fixed interval, I, to some "feature space." This class of functionals corresponds to "elastic matching" in which a symmetry condition and a "focus invariance" are imposed. We provide sufficient conditions under which an optimal matching can be found between two such mappings, the optimal matching being a homeomorphism of the interval I. The differentiability of this matching is also studied, and an application to plane curve comparison is provided.

Key words. calculus of variations, shape representation and recognition, elastic matching, geodesic distance

AMS subject classifications. Primary, 49J45; Secondary, 68T10, 53A04

PII. S036301299934864X

1. Introduction.

In many applications, objects of interest can be represented as numerical functions θ which are defined on some interval $I \subset \mathbb{R}$ and take values in $\mathbb{R}^d$. Several examples may come from signal processing, in which measurements are made during a certain time interval (e.g., speech recognition), analysis of one-dimensional (1D) geological data (e.g., measurements in wells, I being a depth interval), or shape recognition, in which an object can represent a two-dimensional (2D) or three-dimensional (3D) curve.

A problem one typically has to face when dealing with such *functional objects* is to find ways to compare them. This comparison problem is most of the time posed as a *matching problem*, which may be described as "finding similar structures appearing at similar places (or similar times)." To be more explicit, given two "objects" θ and θ', expressed as functions defined on the same interval I, the issue is to find, for each $x \in I$, some $x' \in I$ such that $x \simeq x'$ and $\theta(x) \simeq \theta'(x')$. The matching is *consistent* if the correspondence $x \mapsto x'$ is one-to-one, i.e., there cannot be two distinct locations on θ which are associated to the same location on θ'; it is *complete* if each location in θ is matched to some location in θ', and *bicomplete* if, in addition, each location in θ' is matched to some location in θ. Consistent bicomplete matchings thus can be represented by bijections $\phi : I \to I$, and the matching problem can be formulated as finding such a ϕ such that $\phi \simeq \mathrm{id}$ (where id is the identity function $x \mapsto x$) and $\theta \simeq \theta' \circ \phi$ (in this last sentence, $\simeq$ must be understood as "as close as possible to").

A common approach to realize this program is to minimize some functional $L_{\theta,\theta'}(\phi)$ which is small when the requirements above are satisfied. One simple example is (letting $\dot\phi = \frac{d\phi}{dx}$)

$$L_{\theta,\theta'}(\phi) = \int_I (\dot\phi - 1)^2 dx + \lambda \int_I (\theta(x) - \theta' \circ \phi(x))^2 dx, \tag{1.1}$$

*Received by the editors March 3, 1999; accepted for publication (in revised form) May 23, 2000; published electronically November 15, 2000.
http://www.siam.org/journals/sicon/39-4/34864.html

[†] LAGA (UMR CNRS 7593), Institut Galilée Université Paris XIII, Av. J-B. Clément, F-93430 Villetaneuse, France (trouve@math.univ-paris13.fr).

[‡] CMLA (CNRS, UMR 8536), Ecole Normale Supérieure de Cachan, 61, Avenue du Président Wilson, F-94 235 Cachan CEDEX, France (younes@cmla.ens-cachan.fr).

and many functionals which are used in the literature fall into this category, with some variations. One of the drawbacks in this formulation is the lack of symmetry in θ and θ'. In general, matching θ to θ' or θ' to θ would yield distinct results. This is undesirable, unless there is some reason to privilege one object to the other, and symmetrical matching seems appealing in many contexts. A sufficient condition yielding symmetrical matching would be

$$L_{\theta,\theta'}(\phi) = L_{\theta',\theta}(\phi^{-1}),$$

which is true for functionals of the kind (see section 2)

$$L_{\theta,\theta'}(\phi) = \int_I F(\dot\phi, \theta, \theta' \circ \phi) dx \tag{1.2}$$

with $\xi F(1/\xi, v, u) = F(\xi, u, v)$.

One way to modify (1.1) in order to put it into the above form could be to set

$$F(\xi, u, v) = (\sqrt{\xi} - 1)^2 + \lambda\sqrt{\xi}(u - v)^2.$$

Note that, denoting by $|I|$ the length of the interval I, one has, in this case, (because ϕ is increasing from I onto I)

$$\int_I F(\dot\phi, \theta(x), \theta' \circ \phi(x))dx = 2|I| - \int_I \sqrt{\dot\phi}\left(2 - \lambda(\theta(x) - \theta' \circ \phi(x))^2\right) dx,$$

and the problem is equivalent to *maximizing* $\int_I \sqrt{\dot\phi}\left(2 - \lambda(\theta(x) - \theta' \circ \phi(x))^2\right) dx$.

The problem which has motivated this paper is the functional which has been designed in [9], [10]. In this case, F is given by

$$F(\xi, u, v) = \sqrt{\xi}\left|\cos \frac{u - v}{2}\right| . \tag{1.3}$$

The functional can be used to compare and match plane curves, θ being in this case the angle between the tangent and some reference axis. It has been shown that the minimum, over ϕ, of $\arccos L_{\theta,\theta'}(\phi)$ provides a distance between plane curves seen up to translation and scaling. (That is, it is not only symmetrical but also satisfies the triangular inequality.) In fact, this function F comes in a very natural way from the computation of paths of minimal energy in the space of plane curves.

In the present paper, we only consider the case when F can be written as $\sqrt{\xi}G(u, v)$, and one tries to maximize

$$\int_0^1 \sqrt{\dot\phi}G(\theta, \theta' \circ \phi)dx.$$

Some discussion on how this formulation can be seen as a consequence of some simple assumptions on the matching will be provided in section 2. We shall then study the existence and the properties of solutions of this type of variational problem. More precisely, we shall ask whether there exists some optimal matchings ϕ, *which are bijective.* When such a ϕ exists, we then want to discuss on its smoothness properties, and check, in particular, that the normalization by $\sqrt{\dot\phi}$ has not harmed too much of the smoothing properties of the initial formulation (1.1).

If we forget about θ and θ', we must thus deal with variational problems which fit into the following framework. Without loss of generality, we take $I = [0,1]$ for the rest of the paper and we introduce some notations.

Notation 1. Let Hom^+ be the set of increasing homeomorphisms on $[0,1]$, i.e., the set of continuous strictly increasing functions $\phi : [0,1] \to [0,1]$ such that $\phi(0) = 0$, $\phi(1) = 1$.

Since $\phi \in \mathrm{Hom}^+$ is continuous and increasing, ϕ is differentiable almost everywhere (a.e.). This derivative is denoted $\dot{\phi}$.

Given a measurable function $f : [0,1] \times [0,1] \to \mathbb{R}_+$, we define, for $\phi \in \mathrm{Hom}^+$,

$$U_f(\phi) = \int_0^1 \sqrt{\dot{\phi}} f(\phi(x), x) dx\,.$$

We also let $\tilde{f}(x,y) = f(y,x)$.

Most of the paper (sections 4, 5, and 6) will be devoted to proving the results which are stated in section 3. Some auxiliary results will also be given in section 7. We start with a discussion on which general form a matching function $F(\xi, u, v)$ may assume when some simple invariance properties are required.

2. Invariance properties of the matching. Let a function F be defined on $[0,+\infty[\times\mathbb{R}^d\times\mathbb{R}^d$, and specify the problem of optimal matching between two functions θ and θ', defined on $[0,1]$, and with values in $\mathbb{R}^d$ through the functional, defined for all ϕ, which are increasing diffeomorphisms of $[0,1]$,

$$L_{\theta,\theta'}(\phi) = \int_0^1 F(\dot{\phi}(x), \theta(x), \theta' \circ \phi(x)) dx.$$

As said before, one desirable property is symmetry: for any functions θ and θ', we want that

$$\phi = \mathrm{argmax} L_{\theta,\theta'} \Leftrightarrow \phi^{-1} = \mathrm{argmax} L_{\theta',\theta}.$$

Since

$$L_{\theta',\theta}(\phi^{-1}) = \int_0^1 \dot{\phi}(x) F\left(\frac{1}{\dot{\phi}(x)}, \theta' \circ \phi(c), \theta(x)\right) dx,$$

a sufficient condition for symmetry is the following.

[C1] For all $(\xi, u, v) \in [0,+\infty[\times\mathbb{R}^d \times \mathbb{R}^d$, one has $F(\xi,u,v) = \xi F(1/\xi, v, u)$.

Since we maximize $L_{\theta,\theta'}$, one should build F with suitable properties with respect to maximization. One such property is that F should be concave with respect to its first variable ξ. (To understand why concavity in the first variable is essential for this kind of problem, one can refer to [2].) This is stated in condition [C2].

[C2] For all u, v, $\xi \mapsto F(\xi, u, v)$ is concave on $[0, +\infty[$.

It is important to notice that this concavity assumption is consistent with the symmetry [C1] in the following sense. If [C1] is not true, it is natural to try to symmetrize F by replacing it by

$$F^s(\xi, u, v) = F(\xi, u, v) + \xi F\left(\frac{1}{\xi}, v, u\right).$$

It is easily shown, then, that condition [C2] is true for F^s as soon as it was originally true for F.

Another natural condition for the functional is that, when comparing a function θ with itself, the optimal ϕ is $\phi = \mathrm{id}$. In other terms, one should have, for all functions θ and all diffeomorphisms ϕ,

$$\int_0^1 F(\dot\phi, \theta, \theta \circ \phi)dx \leq \int_0^1 F(1, \theta, \theta)dx.$$

A sufficient condition for this can be that, for all ξ, u, v, $F(\xi, u, v) \leq F(1, u, u)$. If one takes into account the constraint $\int_0^1 \dot\phi = 1$, this can be weakened into the following condition (which can be shown to be necessary and sufficient [8]).

[C3] There exists a measurable function $\lambda : \mathbb{R}^d \to \mathbb{R}$ such that, for all $\xi > 0, u, v \in \mathbb{R}^d$,

$$F(\xi, u, v) \leq F(1, u, u) + \lambda(v)\xi - \lambda(u).$$

Indeed, assuming [C3], we have

$$\int_0^1 F(\dot\phi, \theta, \theta \circ \phi)dx \leq \int_0^1 F(1, \theta, \theta)dx + \int_0^1 \dot\phi\lambda \circ \theta \circ \phi dx - \int_0^1 \lambda \circ \theta dx,$$

and the last two integrals are equal by a change of variables.

Additional constraints may come from invariance properties which may be imposed on the matching. The first invariance property we consider will be called "focus invariance." Consider θ and θ' as signals, defined on $[0, 1]$, and assume that they have been matched by some function ϕ^*. Let $[a, b]$ be a subinterval of $[0, 1]$, and set $[a', b'] = [\phi^*(a), \phi^*(b)]$. We want to refocus the matching on these intervals. For this, we can rescale the functions θ and θ' on these intervals to get new signals defined on $[0, 1]$, and match the new signals. The question which arises then is whether this new matching is consistent with the one which has been obtained initially.

Let us be more precise. To rescale θ (resp., θ'), we define $\theta_{ab}(x) = \theta(a + (b - a)x)$ (resp., $\theta'_{a'b'}(x) = \theta'(a' + (b' - a')x)$), $x \in [0, 1]$. Comparing these signals with the functional F yields an optimal matching which, if it exists, maximizes

$$\int_0^1 F(\dot\phi(x), \theta_{a,b}(x), \theta'_{a',b'} \circ \phi(x))dx. \tag{2.1}$$

The optimal matching between the initial functions θ and θ' clearly maximizes

$$\int_a^b F(\dot\phi(y), \theta(y), \theta' \circ \phi(y))dy$$

with the constraints $\phi(a) = a'$ and $\phi(b) = b'$. Making the change of variables $y = a + (b - a)x$ and setting $\psi(x) = (\phi(y) - a')/(b' - a')$, this integral can be written as

$$(b - a) \int_0^1 F(\lambda\dot\psi(x), \theta_{a,b}(x), \theta'_{a',b'} \circ \psi(x))dx \tag{2.2}$$

with $\lambda = \frac{b'-a'}{b-a}$. We say that F satisfies a focus invariance propertyies if, for any θ and θ', the maximizer of (2.1) is the same as the maximizer of (2.2). One possible condition ensuring such a property is that F is itself (relatively) invariant under the transformation $(\xi, u, v) = (\lambda\xi, u, v)$.

[Focus] For some $\alpha > 0$, for all $\xi > 0$, $u, v \in \mathbb{R}^d$, $F(\lambda\xi, u, v) = \lambda^\alpha F(\xi, u, v)$.

This condition trivially implies that $F(\xi, u, v) = \xi^\alpha F(1, u, v)$. If condition [C1] is imposed, one sees that α has to be equal to $1/2$. We thus get the following.

The only matching functionals which satisfy [C1] *and* [Focus] *take the form*

$$F(\xi, u, v) = \sqrt{\xi} F_1(u, v) \tag{2.3}$$

with $F_1(u, v) = F_1(v, u)$.

These functionals satisfy [C2] as soon as $F_1 \geq 0$. For [C3], we must have, for all $u, v \in \mathbb{R}^d$,

$$F_1(u, v) \leq \sqrt{F_1(u, u) F_1(v, v)}. \tag{2.4}$$

Indeed, assuming [C3], we must have, for some function λ,

$$F_1(u, u) - \lambda(u) = \max_{v, \xi} \sqrt{\xi} F_1(u, v) - \lambda(v)\xi.$$

For a fixed v, $\sqrt{\xi} F_1(u, v) - \lambda(v)\xi$ has a finite maximum if $\lambda(v) > 0$, or if $F_1(u, v) = \lambda(v) = 0$. In the first case, the maximum is given by

$$\frac{F_1(u, v)^2}{4\lambda(v)}.$$

This implies that

$$F_1(u, u) - \lambda(u) = \max_v \frac{F_1(u, v)^2}{4\lambda(v)} \tag{2.5}$$

with the convention $0/0 = 0$. In particular, taking $v = u$, one has

$$F_1(u, u)^2 - 4\lambda(u) F_1(u, u) + 4\lambda(u)^2 \leq 0,$$

which is possible only if $F_1(u, u) = 2\lambda(u)$. Given this fact, (2.5) clearly implies (2.4). We thus have the following.

The only matching functionals which satisfy [C1]–[C3] *and* [Focus] *take the form*

$$F(\xi, u, v) = \sqrt{\xi} F_1(u, v) \tag{2.6}$$

with $F_1(u, v) = F_1(v, u)$, $F_1(u, v) \geq 0$, *and* $F_1(u, v) \leq \sqrt{F_1(u, u) F_1(v, v)}$.

The functional in (1.3) satisfies this property.

One must note, however, that focus invariance under the above form is not a suitable constraint for every matching problem. Let us restrict our attention to the comparison of plane curves, which has initially motivated the present paper. In this case, the functions θ are typically geometrical features computed along the curve and expressed as functions of the (euclidean) arc-length. In such a context, focusing should rather be interpreted from a geometrical point of view, as rescaling (a portion of) a plane curve to length 1. But, in this case, applying such a scale change may have some impact not only on the variable x (which here represents the length), but also on the *values* of the geometric features θ. In example (1.3), the geometric features were the orientation of the tangents, which are not affected by scale change, so that focus invariance is in this case equivalent to geometric scale invariance. Letting κ be the curvature computed along the curve, the same invariance would be true if we had taken $\theta = \kappa'/\kappa^2$ (which is the "curvature" which characterizes curves up to similitudes). But if we had chosen to compare precisely euclidean curvatures, the invariance constraints on the matching would be different. Since curvatures are scaled by λ^{-1} when a curve is scaled by λ, the correct condition should be (instead of [Focus])

$$F(\lambda\xi, \lambda u, v) = \lambda^{\alpha} F(\xi, u, v).$$

This comes from rescaling only the first curve. Rescaling the second curve yields

$$F(\lambda\xi, u, v/\lambda) = \lambda^{\beta} F(\xi, u, v).$$

Note that, if the symmetry condition is valid, we must have $\beta = 1 - \alpha$, which we assume hereafter.

One can solve this identity and compute all the (continuously differentiable) functions which satisfy it. This yields functions F of the kind

$$F(\xi, u, v) = H\left(\xi \frac{v}{u}\right) u^{\alpha} v^{\alpha - 1}.$$

Note that, since F should be concave as a function of ξ, H itself should be concave. The symmetry condition is ensured as soon as $xH(1/x) = H(x)$ for all x. One may set

$$F(\xi, u, v) = -|\xi v - u|,$$

which satisfies [C1]–[C3].

Many variations can be done on these computations. Inspiration on how devising functionals which satisfy given criteria of invariance can be obtained from the first chapters of [4].

We now return to our original problem, which contains, according to the above terminology, symmetrical, focus invariant matchings.

3. Main results.

3.1. Existence results.

Notation 2. Let a and b be two points in $[0,1]^2$ such that $a = (a_1, a_2)$ and $b = (b_1, b_2)$. We denote by $[a, b]$ the closed segment from a to b, i.e., the set of points $a + t(b - a)$, for $0 \leq t \leq 1$, and by $]a, b[$ the open segment from a to b defined by $]a, b[= [a, b] \setminus \{a, b\}$. Moreover, we say that such a segment is horizontal (resp., vertical) if $a_2 = b_2$ (resp., $a_1 = b_1$).

Notation 3. Denote by Δ_f the integral $\Delta_f = \int_0^1 f(x, x)dx$.

Let Ω_f be the set

$$\Omega_f = \left\{ (y, x) \in [0,1]^2 \mid |x - y| \leq \sqrt{1 - \left(\frac{\Delta_f}{\|f\|_\infty}\right)^2} \right\}.$$

THEOREM 3.1. *Assume that $f \geq 0$ is bounded and satisfies conditions* [H1] *and* [H2].

[H1] *There exists a finite family of closed segments $([a_j, b_j])_{j \in J}$ such that each of them is horizontal or vertical and f is continuous on $[0,1]^2 \setminus F$, where $F = \bigcup_{j \in J} [a_j, b_j]$.*

[H2] *Let f_s be defined by*

$$f_s(x) = \lim_{\delta \to 0} \left(\inf\{ f(u) \mid u \in [0,1]^2 \setminus F, \ |u - x| < \delta \} \right).$$

There does not exist any nonempty open vertical or horizontal segment $]a, b[$ *such that* $]a, b[\subset \Omega_f$ *and* f_s *vanishes on* $]a, b[$.

Then there exists $\phi^* \in Hom^+$ *such that* $U_f(\phi^*) = \max\{U_f(\phi), \phi \in Hom^+\}$. *Moreover, if* ϕ *is a maximizer of* U_f*, one has, for all* $x \in [0, 1]$, $(\phi(x), x) \in \Omega_f$.

3.2. Regularity of the optimal matching. We now give some conditions under which the optimal matching satisfies some smoothness properties.

DEFINITION 3.2. *We say that* $f : [0, 1]^2 \to \mathbb{R}$ *is Hölder continuous at* (y, x) *if there exist* $\alpha > 0$ *and* $C > 0$ *such that*

$$|f(y', x') - f(y, x)| \leq C \max(|y' - y|^\alpha, |x' - x|^\alpha) \tag{3.1}$$

for any $(y', x') \in [0, 1]^2$.

We say that f *is locally uniformly Hölder continuous at* (y_0, x_0) *if there exists a neighborhood* V *of* (y_0, x_0) *such that* f *is Hölder continuous at all* $(y, x) \in V$*, with constants* C *and* α*, which are uniform over* V.

THEOREM 3.3. *Let* f *be a nonnegative real-valued measurable function on* $[0, 1]^2$ *and assume that* U_f *reaches its maximal value on* Hom^+ *at* ϕ^*. *Then for any* $x_0 \in [0, 1]$*, if* $f(\phi(x_0), x_0) > 0$ *and if* f *is Hölder continuous at* $(\phi(x_0), x_0)$*, then* ϕ^* *is differentiable at* x_0 *with strictly positive derivative.*

Moreover, if f *is locally uniformly Hölder continuous, then* $\dot{\phi}^*$ *is continuous in a neighborhood of* x_0.

THEOREM 3.4. *Assume that* f *is continuously differentiable in both variables. Let* $\phi \in Hom^+$ *be such that* $U_f(\phi) = \max\{U_f(\psi) \mid \psi \in Hom^+\}$ *and that, for all* $x \in [0, 1]$*, one has* $f(\phi(x), x) > 0$. *Then,* ϕ *is twice continuously differentiable.*

4. Remarks.

4.1. Positivity of *f*. It is necessary to control the vanishing sets of the function f (as in condition [H2]) to obtain a homeomorphism. One simple example is when f vanishes on a square $S =]\frac{1}{2} - a, \frac{1}{2} + a[^2 \subset [0, 1]^2$, and $f = 1$ outside. (Such an f satisfies condition [H1].) Using the fact that (as a consequence of Lemma 5.16 below), ϕ must be linear on any section which does not encounter S, it is not very difficult to prove that the maximum is attained for ϕ which is *discontinuous* at $x = a + 1/2$, more precisely such that $\phi(x) = \frac{1}{2} - a$ and $\phi(x + 0) = \frac{1}{2} + a$.

4.2. Piecewise constant functions. A particular and important case in which condition [H1] is valid is the case of piecewise constant functions f. We state this as a corollary.

COROLLARY 4.1. *Assume that there exists* $0 = x_0 < x_1 < \cdots < x_m = 1$ *(resp.,* $0 = x'_0 < x'_1 < \cdots < x'_n = 1$*) and constants* $f_{kl} > 0$, $1 \leq k \leq m$, $1 \leq l \leq n$ *such that*

$$f(x, x') = \sum_{k=1}^{m} \sum_{l=1}^{n} f_{kl} \mathbf{1}_{[x_{k-1}, x_k[\times[x'_{l-1}, x'_l[}(x, x').$$

Then, there exists $\phi^* \in Hom^+$*, which is piecewise linear such that*

$$U_f(\phi^*) = \max\{U_f(\phi) \mid \phi \in Hom^+\}.$$

Remark. In fact, one needs to assume only that $f_{kl} > 0$ for k and l such that $[x_{k-1}, x_k[\times[x'_{l-1}, x'_l[$ is sufficiently close from the diagonal of the unit square, i.e., intersects the set Ω_f.

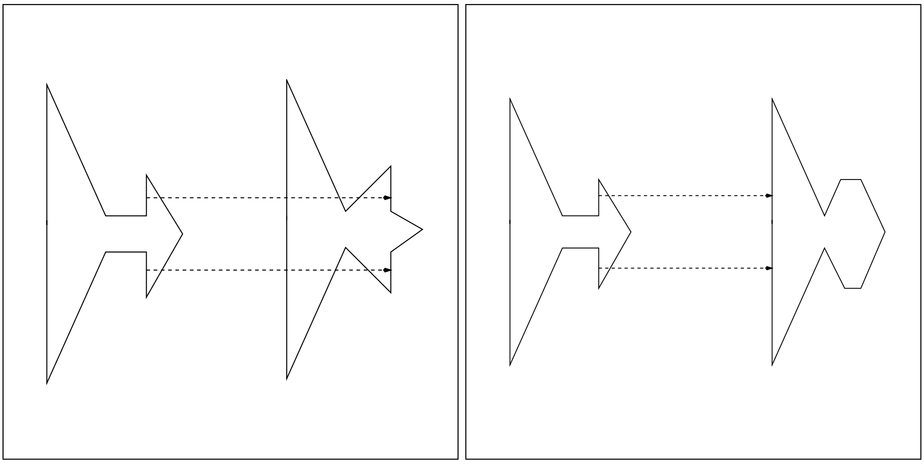


FIG. 4.1. *Matching of flat sections. Left: Situation in which* [H2] *is not true. A point on the first curve has a whole interval of possible matches on the second one with an opposite orientation of the tangent. Right: In this situation, there still exist portions in the second curve with opposite tangents, but because their respectve arc-lengths are far apart, they can be ignored (using the set* Ω_f*):* [H2] *is true.*

Proof. We need to show only that the optimal ϕ is piecewise linear. But we have

$$U_f(\phi) = \sum_{k=1}^{m} \sum_{l=1}^{n} f_{kl} \int_{x_{k-1}}^{x_k} \sqrt{\dot{\phi}(x)}\mathbf{1}_{[x_{l-1},x_l[}(\phi(x))dx.$$

Let $y_k = \phi(x_k)$, $k = 0, \dots, m$. Let $y'_l = \phi^{-1}(x'_l)$, $l = 0, \dots, n$. We have, writing $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$,

$$U_f(\phi) = \sum_{k=1}^{m} \sum_{l=1}^{n} f_{kl} \int_{x_{k-1} \vee y'_{l-1}}^{x_k \wedge y'_l} \sqrt{\dot{\phi}(x)}dx,$$

and Lemma 5.16 yields the result. □

4.3. Application to optimal matching of functions. Let us see what conditions [H1] and [H2] mean when f is of the kind

$$f(y, x) = F_1(\theta(y), \theta'(x))\,.$$

One of the examples we have in mind is the case when $\theta(y)$ and $\theta'(y)$ take values in $[0, 2\pi[$, and $F_1(u, v) = |\cos[(u - v)/2]|$. In this case, these functions, θ and θ', correspond to *rotation angles* of the unitary tangents to some plane curves, and matching is used to compare shapes on the basis of their silhouettes.

In this particular case, F_1 is continuous, but not continuously differentiable. It is, however, smooth enough to fit into the regularity condition [H1], so that the true constraint is on θ and θ'. Note that f is discontinuous on a horizontal (resp., vertical) segment as soon as θ (resp., θ') is discontinuous at the position of the segment in the horizontal (resp., vertical) axis. Thus [H1] implies that θ and θ' are continuous except at a finite number of points. Points of discontinuity of θ and θ' are angular points for the plane curves they represent; thus condition [H1] implies that one can safely perform a matching between shapes having a finite number of angular points, which is the case of most of the objects which can be observed in a standard environment. Note that, in this case, piecewise constant f corresponds to polygonal shapes, which is also an important example to deal with.

Condition [H2] essentially means that one cannot have intervals on which, for a given x_0, $F(\theta(.), \theta'(x_0)) = 0$. In the case of curve matching, when θ is an angle, and formula (1.3) is used, this means that one of the curves cannot have a flat portion which may be matched to a point of the other curve with opposite tangent. (Note that one can restrict this condition to points which are located at close enough positions on both curves; see Figure 4.1 for an illustration.) In particular, the condition is always true if the compared curves contain no flat sections.

5. Proof of the existence.

5.1. Sketch of the proof. In the next section, we will introduce a compact set $\mathcal{D}^*$, containing Hom^+, and extend the functional U_f to this space. We first prove the existence of the maximum for this extended functional through the following proposition.

PROPOSITION 5.1. *If f satisfies condition* [H1]*, then U_f is upper-semicontinuous on $\mathcal{D}^*$. Since $\mathcal{D}^*$ is compact, there exists $\phi^* \in \mathcal{D}^*$ such that*

$$U_f(\phi^*) = \sup\{U_f(\psi) \,|\, \psi \in \mathcal{D}^*\}.$$

Moreover, if ϕ is a maximizer of U_f, one has, for all $x \in [0,1]$, $(\phi(x), x) \in \Omega_f$.

Theorem 3.1 will then be a consequence of the following proposition.

PROPOSITION 5.2. *If f satisfies condition* [H2] *in Theorem* 3.1 *and $\phi^* \in \mathcal{D}^*$ is such that*

$$U_f(\phi^*) = \sup\{U_f(\psi) \,|\, \psi \in \mathcal{D}^*\},$$

then $\phi \in Hom^+$.

Before proving these propositions, we introduce $\mathcal{D}^*$.

5.2. The set $\mathcal{D}^*$.

5.2.1. Definition. Let $\mathcal{M}_1$ be the set of the positive Radon measures μ on $[0,1]$ such that $\mu([0,1]) = 1$. Let $\mathcal{M}$ be the set of measures on $[0,1]$ such that $\mu([0,1]) \leq 1$. We let $\mathcal{D}^*$ be the set of all ϕ which can be written as

$$\phi(s) = \mu([0,s[)$$

for some $\mu \in \mathcal{M}_1$. Such ϕ are nondecreasing, left continuous, and satisfy $\phi(0) = 0$ and $\phi(1) \leq 1$. Conversely, any ϕ satisfying these properties is in $\mathcal{D}^*$, and the associated measure is unique and will be denoted by μ_ϕ. Note that $\mu_\phi(\{1\}) = 1 - \phi(1)$ for all $x \in [0,1[$, $\mu_\phi(\{x\}) = \phi(x+0) - \phi(x)$, where $\phi(x+0)$ is the right limit of ϕ at x.

Any $\mu \in \mathcal{M}_1$ can be written in a unique way under the form

$$\mu = \omega dx + \nu,$$

where dx is the Lebesgue measure on $[0,1]$, ω is a measurable, nonnegative, function on $[0,1]$, and ν is singular. (There exists a set E of Lebesgue measure 0 such that $\nu(A) = \nu(A \cap E)$ for every Borel set $A \subset [0,1]$.) For $\phi \in \mathcal{D}^*$, we take the notation

$$d\mu_\phi = \omega_\phi ds + d\nu_\phi.$$

DEFINITION 5.3. *For a function $\phi \in [0,1]$, we let $\dot{\phi}(x)$ be the limit when $\epsilon \to 0$ of $\frac{\phi(x+\epsilon)-\phi(x)}{\epsilon}$ when this limit exists and $\dot{\phi}(x) = 0$ otherwise. If $\phi \in \mathcal{D}^*$, one has $\dot{\phi} = \omega_\phi$ a.e.* [7, Theorem 8.18].

Following [6], we extend the functional U_f to $\mathcal{D}^*$ by letting

$$U_f(\phi) = \int_0^1 \sqrt{\dot{\phi}(x)} f(\phi(x), x) ds\,.$$

We also denote by $\mathcal{D}^*_+$ the set of functions $\phi \in \mathcal{D}^*$ for which $\int_a^b \dot{\phi} > 0$ for any $0 \leq a < b \leq 1$. We have $\phi \in \mathrm{Hom}^+$ if $\phi \in \mathcal{D}^*_+$, and ν_ϕ is diffuse, i.e., $\nu_\phi(\{x\}) = 0$ for any $x \in [0,1]$. (Note that this is not a necessary condition: there exists functions in Hom^+ such that $\dot{\phi} = 0$ a.e. See [3, example 18–8] .

5.2.2. Weak convergence in $\mathcal{D}^*$.

Measures $\mu_n \in \mathcal{M}, n \geq 0$ are said to converge for the weak*-topology to a limit μ if, for every continuous function F on $[0,1]$, one has

$$\lim_{n\to\infty} \int_0^1 F d\mu_n = \int_0^1 F d\mu.$$

Since this is the only kind of convergence we use on $\mathcal{M}$ and $\mathcal{M}_1$, the statement "μ_n converges to μ" will always mean convergence in the weak*-topology. We say that $\phi_n \in \mathcal{D}^*$ weakly converges to $\phi \in \mathcal{D}^*$ if μ_{ϕ_n} converges to μ_ϕ.

We list some results related to this convergence.

PROPOSITION 5.4 (see [5]). *The sets $\mathcal{M}$ and $\mathcal{M}_1$ are compact for the weak*-topology.*

If ϕ_n weakly converges to ϕ, then, for all $x \in [0,1]$ such that ϕ is continuous at x, one has $\phi_n(x) \to \phi(x)$.

Note that, since ϕ is increasing, its discontinuity set is at most countable.

PROPOSITION 5.5. *Let ϕ_n be a sequence in $\mathcal{D}^*$, such that $\dot{\phi}_n dx$ and ν_{ϕ_n} both converge in $\mathcal{M}$, respectively, to $\alpha dx + \rho$ and $\beta dx + \tau$. Then μ_{ϕ_n} converges to $\mu \in \mathcal{M}_1$ such that $\mu = (\alpha + \beta)dx + (\rho + \tau)$.*

This is obvious. Note that, by compactness of $\mathcal{M}$, from any sequence ϕ_n one can extract a subsequence such that both $\dot{\phi}_n dx$ and ν_{ϕ_n} converge.

We introduce, here and for what follows, a mollifier g, i.e., an infinitely differentiable mapping $g : \mathbb{R} \to \mathbb{R}$, with compact support included in $]-1,1[$, such that $\int_{-1}^1 g(x)dx = 1$. For $\epsilon > 0$, we let $g_\epsilon(x) = g(x/\epsilon)/\epsilon$. One has the following lemma.

LEMMA 5.6 (see [6]). *Let $\mu_n = \omega_n dx$ be a sequence of absolutely continuous measures in $\mathcal{M}$ which converges to $\alpha dx + \rho$. Then, for any $\epsilon > 0$, one has*

$$\lim_{n\to\infty} \int_0^1 |\omega_n * g_\epsilon - \alpha * g_\epsilon - \rho * g_\epsilon|\, dx = 0.$$

Here, $\mu * g_\epsilon$ denotes (as usually) the convolution of μ by g_ϵ,

$$\mu * g_\epsilon(x) = \int_0^1 g_\epsilon(x-y) d\mu(y)\,.$$

5.2.3. A symmetry property of U_f.

For $\phi \in \mathcal{D}^*$, we define, for $y \in [0,1]$,

$$\phi^-(y) = \sup\{x \in [0,1] \,|\, \phi(x) < y\},$$

with the convention that $\sup\emptyset = 0$.

Our purpose is to prove the following proposition (recall that we have denoted $\tilde{f}(x,y) = f(y,x)$).

PROPOSITION 5.7. *For all $\phi \in \mathcal{D}^*$, one has*

$$U_f(\phi) = U_{\tilde{f}}(\phi^-).$$

The proof will be carried on with several lemmas.

LEMMA 5.8. *Let $\phi \in \mathcal{D}^*$.*

1. *We have $\phi^- \in \mathcal{D}^*$ and $(\phi^-)^- = \phi$.*
2. *For any $x \in [0,1]$,*

$$\phi^-(\phi(x)) \le x \le \phi^-(\phi(x)+0) \tag{5.1}$$

so that $\phi^- \circ \phi(x) = x$ as soon as ϕ^- is continuous at $\phi(x)$.

3. *For any $y \in [0,1]$,*

$$\phi(\phi^-(y)) \le y \le \phi(\phi^-(y)+0) \tag{5.2}$$

so that $\phi \circ \phi^-(y) = y$ as soon as ϕ is continuous at $\phi^-(y)$.

Moreover, if $\phi \in \mathcal{D}^*_+$, then ϕ^- is continuous.

Proof. ϕ^- is nondecreasing, $\phi^-(0) = 0$, $\phi^-(1) \le 1$, and ϕ^- is left continuous, since

$$\phi^-(y) = \sup_{h>0}\left(\sup\{x, \phi(x) < y-h\}\right) = \sup_{h>0}\phi^-(y-h)$$

so that $\phi^- \in \mathcal{D}^*$. Now, let us show that for any $(x,y) \in [0,1]^2$,

$$\phi(x) > y \Rightarrow \phi^-(y) < x. \tag{5.3}$$

Indeed (assume $x \ne 0$—otherwise the result is trivial), if $\phi(x) > y$, since ϕ is left continuous, there exists $h > 0$ such that $\phi(x-h) > y$ so that $\phi^-(y) \le x - h < x$. Moreover, we deduce from the definition of ϕ^- that for any $x, y \in [0,1]$,

$$\phi^-(y) < x \Rightarrow \phi(x) \ge y. \tag{5.4}$$

Now, since $\phi(x) = \sup\{\ y \in [0,1] \mid y < \phi(x)\ \}$, using (5.3) and (5.4), we get

$$\phi(x) \le \sup\{y \in [0,1] | \ \phi^-(y) < x\} = (\phi^-)^-(x) \le \sup\{y \in [0,1] | \phi(x) \ge y\} = \phi(x)$$

so that 1 is proved.

From 1, we deduce that $2 \iff 3$ so that it is sufficient to prove 3. For any y, there exists an increasing sequence x_n which converges to $\phi^-(y)$ such that $\phi(x_n) < y$. Since ϕ is left continuous, this yields $\phi \circ \phi^-(y) \le y$. Moreover, for all $h > 0$, one has $\phi(\phi^-(y)+h) \ge y$ so that $y \le \phi(\phi^-(y)+0)$.

Now, assume that $\phi \in \mathcal{D}^*_+$ and assume that ϕ^- is discontinuous at $y_0 \in [0,1[$. Then $\phi^-(y_0+0) > \phi^-(y_0)$ so that $(\phi^-)^-$ has the constant value y_0 on $]\phi^-(y_0), \phi^-(y_0+0)]$. Since $(\phi^-)^- = \phi$, we get a contradiction with the fact that ϕ is strictly increasing. □

Note that ϕ is continuous at $\phi^-(y)$ if and only if $\mu_\phi(\{\phi^-(y)\}) = \nu_\phi(\{\phi^-(y)\}) = 0$.

LEMMA 5.9. *Let $\phi \in \mathcal{D}^*$. If ϕ is derivable at x, then*

$$\lim_{h\to 0}\frac{\phi(x+h+0) - \phi(x)}{h} = \dot{\phi}(x).$$

Proof. Let us show that, for any sequence x_n which converges to x, $x_n \neq x$, one has $\frac{\phi(x_n+0)-\phi(x)}{x_n-x} \to \dot{\phi}(x)$. Since ϕ is increasing, the limit is clearly larger than $\dot{\phi}(x)$. Letting $\epsilon_n = (x_n - x)^2 > 0$, we have

$$\frac{\phi(x_n+0)-\phi(x)}{x_n-x} \leq \frac{\phi(x_n+\epsilon_n)-\phi(x)}{x_n-x} = \frac{\phi(x_n+\epsilon_n)-\phi(x)}{x_n+\epsilon_n-x}(1+|x_n-x|),$$

and the last term converges to $\dot{\phi}(x)$. □

Define $P_\phi = \{\, x \in [0,1] \mid \dot{\phi}(x) > 0 \,\}$ to be the set of locations where the derivative of ϕ exists and is strictly positive (see Definition 5.3). One has the following lemma.

Lemma 5.10. *For any $x_0 \in [0,1]$, $x_0 \in P_\phi \iff \phi(x_0) \in P_{\phi^-}$. Hence,*

$$\dot{\phi}(x) = \frac{1}{\dot{\phi^-}(\phi(x))}\mathbf{1}_{P_{\phi^-}}(\phi(x))$$

(with the convention $0/0 = 0$).

Proof. ($\Leftarrow$) Assume that $\phi(x_0) \in P_{\phi^-}$; then ϕ is continuous at x_0. Indeed, if $\phi(x_0) < \phi(x_0+0)$, then ϕ^- is constant on $]\phi(x_0), \phi(x_0+0)[$ so that $\dot{\phi^-}(\phi(x_0)) = 0$ and $\phi(x_0) \notin P_{\phi^-}$ (which is a contradiction). Moreover, since ϕ^- is continuous at $\phi(x_0)$, we deduce from Lemma 5.8 that $\phi^-(y_0) = x_0$, where $y_0 = \phi(x_0)$. Now, noting that for any $h \in \mathbb{R}^*$ such that $x_0 + h \in [0,1]$, $\phi(x_0+h) \neq \phi(x_0)$ (otherwise ϕ^- should be discontinuous at $y_0 = \phi(x_0)$), we get using (5.1) and the fact that $\phi^-(\phi(x_0)) = x_0$

$$\frac{\phi(x_0+h)-\phi(x_0)}{\phi^-[\phi(x_0+h)+0]-\phi^-[\phi(x_0)]} \leq \frac{\phi(x_0+h)-\phi(x_0)}{h} \leq \frac{\phi(x_0+h)-\phi(x_0)}{\phi^-[\phi(x_0+h)]-\phi^-[\phi(x_0)]}.$$

Since ϕ is continuous at x_0, Lemma 5.9 applied to ϕ^- implies that $[\phi(x_0+h)-\phi(x_0)]/h$ converges to $(\dot{\phi^-} \circ \phi(x_0))^{-1} > 0$ when h tends to 0 so that $x_0 \in P_\phi$.

($\Rightarrow$) Now, assume that $x_0 \in P_\phi$. Then if $y_0 = \phi(x_0)$, ϕ^- is continuous at y_0. Indeed, if $\phi^-(y_0) < \phi^-(y_0+0)$, then since $\phi = (\phi^-)^-$, ϕ is constant on $]\phi^-(y_0), \phi^-(y_0+0)]$. However, from (5.1), we get that $x_0 \in [\phi^-(y_0), \phi^-(y_0+0)]$ so that $\dot{\phi}(x_0) = 0$ (which is a contradiction). Hence, ϕ^- is continuous at $y_0 = \phi(x_0)$ and $\phi^-(y_0) = x_0$. Now, using the part ($\Leftarrow$) for ϕ^-, we deduce that $y_0 = \phi^-(y_0) \in P_\phi = P_{(\phi^-)^-}$ implies $y_0 \in P_{\phi^-}$, i.e., $\phi(x_0) \in P_{\phi^-}$ so that the proof is finished. □

Lemma 5.11. *If $\phi \in \mathcal{D}^*$ and g is a measurable function, one has, for all $x \in [0,1]$,*

$$\int_0^{\phi(x)} g \circ \phi^-(v)dv = \int_0^x g(u)\dot{\phi}(u)du + \int \mathbf{1}_{[0,x[}(u)g(u)d\nu_\phi(u). \tag{5.5}$$

Proof. This lemma can be proved first for $g = \mathbf{1}_{[0,b[}$ and extended to any g in a standard way. □

We can now prove Proposition 5.7. Applying Lemma 5.10 to ϕ^- instead of ϕ, we get

$$\begin{aligned} U_{\tilde{f}}(\phi^-) &= \int_0^1 \frac{\mathbf{1}_{P_\phi}(\phi^-(u))}{\sqrt{\dot{\phi} \circ \phi^-(u)}} f(u, \phi^-(u))du \\ &= \int_0^{\phi(1)} \frac{\mathbf{1}_{P_\phi}(\phi^-(u))}{\sqrt{\dot{\phi} \circ \phi^-(u)}} f(\phi \circ \phi^-(u), \phi^-(u))du. \end{aligned}$$

To justify this equality, we must show that the replacements of 1 by $\phi(1)$ and of u by $\phi \circ \phi^-(u)$ are valid. Assume that $\phi(1) < 1$. This implies that $\nu_\phi(\{1\}) > 0$ and thus that $1 \notin P_\phi$; for $u > \phi(1)$, one has $\phi^-(u) = 1$, so that $\mathbf{1}_{P_\phi}(\phi^-(u)) = 0$, which justifies the first replacement. For the second one, one has $\phi \circ \phi^-(x) \neq x$ only if ϕ is discontinuous at $\phi^-(x)$ and so not differentiable at $\phi^-(x)$, so that $\phi^-(x) \notin P_\phi$.

Now, Lemma 5.11 implies $U_{\tilde{f}}(\phi^-) = U_f(\phi)$ since ϕ is not derivable (hence, with our convention $\dot{\phi} = 0$) ν_ϕ a.e. [7, Theorem 8.11].

5.3. Proof of Proposition 5.1. We prove that U_f is upper-semicontinuous on $\mathcal{D}^*$. Let us consider the following lemma.

LEMMA 5.12. *Let f be a nonnegative function on $[0,1]^2$ which satisfies* [H1]. *Then, there exists a sequence $(f_n)_{n\geq 0}$ of continuous and nonnegative functions on $[0,1]^2$ such that for all $\phi \in \mathcal{D}^*$*

$$U_f(\phi) = \inf_{n\geq 0} U_{f_n}(\phi).$$

Proof. Let $F = \bigcup_{j\in J}[a_j, b_j]$ be the compact set defined in Theorem 3.1, let $M = \|f\|_\infty$, and consider the sequence $(f_n)_{n\geq 0}$ of nonnegative functions defined by

$$f_n(x) = (1 - \alpha_n(x))M + \alpha_n(x)f(x),$$

where $\alpha_n(x) = (Cd(x,F))^{1/n}$ and $d(x,F)$ is the usual distance from x to F and $0 < C < 1/\sqrt{2}$. One easily shows that f_n is continuous on $[0,1]$ and that $(f_n(x))_{n\geq 0}$ is a decreasing sequence converging to $f(x)$ for any $x \in [0,1]^2 \setminus F$.

Now consider $\phi \in \mathcal{D}^*$. By definition of the f_n's, $U_{f_n}(\phi)$ is a decreasing sequence. To show the result, it is sufficient to prove that

$$\sqrt{\dot{\phi}} f_n(\phi(x), x) \to \sqrt{\dot{\phi}} f(\phi(x), x) \text{ a.e.}$$

The result is obviously true for $x \in \{\ z \in [0,1] \mid \dot{\phi}(z) = 0 \ \text{ or } (\phi(z), z) \in [0,1]^2 \setminus F\}$. However, the set $\mathcal{F} = \{\ z \in [0,1] \mid \dot{\phi}(z) > 0,\ (\phi(z), z) \in F\ \}$ contains only isolated points, so that the result is proved: indeed, by contradiction assume that there exist $x \in \mathcal{F}$ and a sequence $(x_n)_{n\geq 0}$ of points of $\mathcal{F} \setminus \{x\}$ converging to x. Since F contains only a finite number of segments, there exists $j_0 \in J$ such that (up to the extraction of a subsequence) $x_n \in [a_{j_0}, b_{j_0}]$ for all $n \geq 0$. Moreover, since there exist n and n' such that $x_n \neq x_{n'}$, the segment $[a_{j_0}, b_{j_0}]$ is vertical so that $\phi(x_n)$ has a constant value and $\dot{\phi}(x) = 0$, which is a contradiction. □

Using Lemma 5.12, we deduce that if Theorem 3.1 is proved for nonnegative and continuous f, then, using the fact that the infimum of a family of upper-semicontinuous functions is upper-semicontinuous, we will get the result for any f nonnegative and satisfying condition [H1]. Hence, we can assume that f is continuous and nonnegative and prove that U_f is upper-semicontinuous.

For this, we consider a sequence $\phi_n \in \mathcal{D}^*$ such that ϕ_n weakly converges to ϕ in $\mathcal{D}^*$ (i.e., μ_{ϕ_n} converges to μ_ϕ in $\mathcal{M}$). Replacing, if needed, ϕ_n by a subsequence, we assume that both $\dot{\phi}_n dx$ and ν_{ϕ_n} converge in $\mathcal{M}$, respectively to $\alpha dx + \rho$ and $\beta dx + \tau$. By Proposition 5.5, ϕ_n weakly converges to $\phi \in \mathcal{D}^*$ with $\mu_\phi = (\alpha + \beta)dx + (\rho + \tau)$. We shall show that $\limsup U_f(\phi_n) \leq U_f(\phi)$. We have

$$U_f(\phi) - U_f(\phi_n) = \int_0^1 \left(\sqrt{\dot{\phi}(x)} - \sqrt{\dot{\phi}_n(x)}\right) f(\phi(x), x)dx + \int_0^1 \sqrt{\dot{\phi}_n(x)}(f(\phi(x), x) - f(\phi_n(x), x))dx.$$

Moreover,

$$\begin{aligned}
&\left|\int_0^1 \sqrt{\dot{\phi}_n}(x)(f(\phi(x),x) - f(\phi_n(x),x))dx\right| \\
&\quad\le \int_0^1 \sqrt{\dot{\phi}_n}(x)\left|(f(\phi(x),x) - f(\phi_n(x),x))\right| dx \\
&\quad\le \left[\int_0^1 \dot{\phi}_n(x)dx\right]^{1/2} \left[\int_0^1 \left|(f(\phi(x),x) - f(\phi_n(x),x))\right|^2 dx\right]^{1/2} \\
&\quad= \left[\int_0^1 \left|(f(\phi(x),x) - f(\phi_n(x),x))\right|^2 dx\right]^{1/2}.
\end{aligned}$$

This last integral tends to 0 by dominated convergence, since ϕ_n converges to ϕ a.e. (Proposition 5.4) and f is continuous. We thus have

$$\lim_{n\to\infty} \int_0^1 \sqrt{\dot{\phi}_n}(x)(f(\phi(x),x) - f(\phi_n(x),x))dx = 0. \tag{5.6}$$

We now study

$$\int_0^1 \left(\sqrt{\dot{\phi}}(x) - \sqrt{\dot{\phi}_n}(x)\right) f(\phi(x),x)dx.$$

We show that

$$\limsup \int_0^1 \sqrt{\dot{\phi}_n}(x) f(\phi(x),x)dx$$

is smaller than

$$\int_0^1 \sqrt{\alpha(x)} f(\phi(x),x)dx.$$

Since

$$\int_0^1 \sqrt{\dot{\phi}}(x) f(\phi(x),x)dx = \int_0^1 \sqrt{\alpha(x)+\beta(x)} f(\phi(x),x)dx,$$

this will prove Proposition 5.1.

We follow the method of [6], using the mollifier g_ϵ. We first prove the following lemmas.

Lemma 5.13. *For any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \int_0^1 \sqrt{\dot{\phi}_n * g_\epsilon(x)} f(\phi(x),x)dx = \int_0^1 \sqrt{\alpha * g_\epsilon(x) + \rho * g_\epsilon(x)} f(\phi(x),x)dx. \tag{5.7}$$

Proof. Indeed, we have

$$\begin{aligned}
&\int_0^1 \left|\sqrt{\dot{\phi}_n * g_\epsilon(x)} - \sqrt{\alpha * g_\epsilon(x) + \rho * g_\epsilon(x)}\right| f(\phi(x),x)dx \\
&\quad\le \|f\|_\infty \int_0^1 \left|\sqrt{\dot{\phi}_n * g_\epsilon(x)} - \sqrt{\alpha * g_\epsilon(x) + \rho * g_\epsilon(x)}\right| dx \\
&\quad\le \|f\|_\infty \left[\int_0^1 |\dot{\phi}_n * g_\epsilon(x) - (\alpha * g_\epsilon(x) + \rho * g_\epsilon(x))|dx\right]^{1/2} \\
&\quad\le \|f\|_\infty \left[\int_{-\infty}^{\infty} |\dot{\phi}_n * g_\epsilon(x) - (\alpha * g_\epsilon(x) + \rho * g_\epsilon(x))|dx\right]^{1/2},
\end{aligned}$$

which tends to 0 by Lemma 5.6. We have used the inequality $|\sqrt{a}-\sqrt{b}|^2 \leq |a-b|$, which is true for all $a, b \geq 0$. □

Lemma 5.14. *We have*

$$\lim_{\epsilon\to 0}\sup_n \left|\int_0^1 \sqrt{\dot\phi_n} * g_\epsilon(x) f(\phi(x),x)dx - \int_0^1 \sqrt{\dot\phi_n} f(\phi(x),x)dx\right|.$$

Proof. For any $\eta > 0$ one can find a continuous function F_η on $[0,1]$ such that

$$\int_0^1 |F_\eta(x) - f(\phi(x),x)|^2 dx < \eta^2 .$$

Fixing such an η, one has

$$\int_0^1 \sqrt{\dot\phi_n} * g_\epsilon(x)\,|f(\phi(x),x) - F_\eta(x)|\,dx \leq \eta\left[\int_0^1 \dot\phi_n * g_\epsilon(x)dx\right]^{1/2} \leq \eta,$$

where we have used the fact that, by Jensen's inequality, we have for all $n \geq 0$

$$\sqrt{\dot\phi_n} * g_\epsilon(x) \leq \sqrt{\dot\phi_n * g_\epsilon}.$$

Similarly,

$$\int_0^1 \sqrt{\dot\phi_n(x)}\,|f(\phi(x),x) - F_\eta(x)|\,dx \leq \eta.$$

We have

$$\begin{aligned}
&\left|\int_0^1 \sqrt{\dot\phi_n} * g_\epsilon(x) F_\eta(x)dx - \int_0^1 \sqrt{\dot\phi_n(y)} F_\eta(y)dy\right| \\
&\leq \int_{-\epsilon}^{1+\epsilon} dx \int_0^1 dy \sqrt{\dot\phi_n}(y) g_\epsilon(x-y)|F_\eta(x) - F_\eta(y)| \\
&\leq K_\eta(\epsilon)\int_0^1 \sqrt{\dot\phi_n}(y)dy \int_{-\epsilon}^{1+\epsilon} g_\epsilon(x-y)dx \\
&\leq K_\eta(\epsilon)\int_0^1 \sqrt{\dot\phi_n}(y)dy \leq K_\eta(\epsilon),
\end{aligned}$$

where $K_\eta(\epsilon) = \sup_{|y-x|\leq\epsilon}(|F_\eta(x) - F_\eta(y)|)$: we have used the fact that, for all y,

$$\int_{-\epsilon}^{1+\epsilon} g_\epsilon(x-y)dx \leq \int_{-\infty}^{\infty} g_\epsilon(x-y)dx = \int_{-\infty}^{\infty} g_\epsilon(x)dx = 1$$

and

$$\int_0^1 \sqrt{\dot\phi_n}(y)dy \leq 1.$$

Hence, we deduce that, for all n,

$$\left|\int_0^1 \left(\sqrt{\dot\phi_n} * g_\epsilon(x) - \sqrt{\dot\phi_n(x)}\right) f(\phi(x),x)dx\right| \leq 2\eta + K_\eta(\epsilon).$$

Since for $\eta > 0$, $K_\eta(\epsilon) \to 0$ when ϵ vanishes, we get the result. □

We end with the following lemma.

LEMMA 5.15. *We have*

$$\lim_{\epsilon\to 0}\int_0^1 \sqrt{\alpha * g_\epsilon(x) + \rho * g_\epsilon(x)}f(\phi(x),x)dx = \int_0^1 \sqrt{\alpha(x)}f(\phi(x),x)dx. \tag{5.8}$$

Proof. Indeed,

$$\begin{aligned}
&\left|\int_0^1 \left[\sqrt{\alpha * g_\epsilon(x) + \rho * g_\epsilon(x)} - \sqrt{\alpha(x)}\right] f(\phi(x),x)dx\right| \\
&\quad\le \left|\int_0^1 \left[\sqrt{\alpha * g_\epsilon(x) + \rho * g_\epsilon(x)} - \sqrt{\alpha * g_\epsilon(x)}\right] f(\phi(x),x)dx\right| \\
&\quad + \left|\int_0^1 \left[\sqrt{\alpha * g_\epsilon(x)} - \sqrt{\alpha(x)}\right] f(\phi(x),x)dx\right| \\
&\quad\le \|f\|_\infty \left|\int_0^1 \sqrt{\rho * g_\epsilon(x)}dx\right| + \|f\|_\infty \left|\int_0^1 \left[\sqrt{\alpha * g_\epsilon(x)} - \sqrt{\alpha(x)}\right] dx\right|.
\end{aligned}$$

One has $\rho * g_\epsilon(x) = \int_0^1 g_\epsilon(x-y)d\rho(y) \le \frac{1}{\epsilon}\rho(]x-\epsilon, x+\epsilon[)$. Since ρ is singular, this upper-bound tends to 0 a.e. (cf., for example, [7, Theorem 8.6]). Thus

$$\begin{aligned}
\int_0^1 \sqrt{\rho * g_\epsilon(x)}dx &\le \int_0^1 \sqrt{\rho * g_\epsilon(x)}\mathbf{1}_{\rho * g_\epsilon(x)\le 1}dx + \int_0^1 \sqrt{\rho * g_\epsilon(x)}\mathbf{1}_{\rho * g_\epsilon(x)> 1}dx \\
&\le \int_0^1 \sqrt{\rho * g_\epsilon(x)}\mathbf{1}_{\rho * g_\epsilon(x)\le 1}dx \\
&\quad + \left[\int_0^1 \rho * g_\epsilon(x)dx\right]^{1/2}\left[\int_0^1 \mathbf{1}_{\rho * g_\epsilon(x)> 1}dx\right]^{1/2} \\
&\le \int_0^1 \sqrt{\rho * g_\epsilon(x)}\mathbf{1}_{\rho * g_\epsilon(x)\le 1}dx + \left[\int_0^1 \mathbf{1}_{\rho * g_\epsilon(x)> 1}dx\right]^{1/2},
\end{aligned}$$

which tends to 0 by dominated convergence. We have used the fact that $\int_0^1 \rho * g_\epsilon(x)dx \le 1$.

On the other hand,

$$\left|\int_0^1 \left[\sqrt{\alpha * g_\epsilon(x)} - \sqrt{\alpha(x)}\right] dx\right|^2 \le \int_0^1 |\alpha * g_\epsilon(x) - \alpha(x)|dx,$$

which tends to 0 by [1, Theorem IV.22] . □

We can now end the proof of Proposition 5.1. For any $\eta > 0$, we deduce from Lemmas 5.15 and 5.14 that there exists $\epsilon > 0$ so that

$$\int_0^1 \sqrt{\alpha * g_\epsilon(x) + \rho * \epsilon(x)}f(\phi(x),x)dx \le \int_0^1 \sqrt{\alpha(x)}f(\phi(x),x)dx + \eta,$$

and for all $n \ge 0$

$$\int_0^1 \sqrt{\dot\phi_n}f(\phi(x),x)dx \le \int_0^1 \sqrt{\dot\phi_n} * g_\epsilon(x)f(\phi(x),x)dx + \eta.$$

Now, using Lemma 5.13, we deduce that for n sufficiently large, we have

$$\int_0^1 \sqrt{\dot\phi_n * g_\epsilon(x)} f(\phi(x), x) dx \le \int_0^1 \sqrt{\alpha * g_\epsilon(x) + \rho * \epsilon(x)} f(\phi(x), x) dx + \eta.$$

Moreover, by Jensen's inequality, we have, for all n, $\sqrt{\dot\phi_n} * g_\epsilon(x) \le \sqrt{\dot\phi_n * g_\epsilon}$ so that, using the previous inequalities, we get for sufficiently large n

$$\int_0^1 \sqrt{\dot\phi_n} f(\phi(x), x) dx \le \int_0^1 \sqrt{\alpha(x)} f(\phi(x), x) dx + 3\eta.$$

Taking the lim sup and since η is arbitrary, we get the result.

We now prove the last statement of Proposition 5.1, that is, the fact that if ϕ is a miximizer of U_f, then, for all $x \in [0, 1]$, one has $(\phi(x), x) \in \Omega_f$. We start with a simple fact.

LEMMA 5.16. *Let* $[a, b] \subset [0, 1]$ *and* $[\tilde a, \tilde b] \subset [0, 1]$. *Then*

$$\max \left\{ \int_a^b \sqrt{\dot\phi(x)} dx \mid \phi \in \mathcal{D}^*, \phi(a) = \tilde a, \phi(b) = \tilde b \right\} = \sqrt{b - a}\sqrt{\tilde b - \tilde a},$$

and the maximum is attained for ϕ *linear between* a *and* b.

Proof. Indeed,

$$\left[\int_a^b \sqrt{\dot\phi} \right]^2 \le (b - a) \int_a^b \dot\phi \le (b - a)(\tilde b - \tilde a)$$

with equality if ϕ is linear. □

LEMMA 5.17.

$$U_f(\phi) \le \|f\|_\infty \sqrt{1 - \|\phi - \mathrm{id}\|_\infty^2}.$$

Proof. Take $x \in [0, 1]$ and set $M = |\phi(x) - x|$. Assume first that $\phi(x) = x + M$. Applying Lemma 5.16 between 0 and x and between x and 1, we get

$$U_f(\phi) \le \|f\|_\infty \int_0^1 \sqrt{\dot\phi} dx \le \|f\|_\infty \left(\sqrt{x}\sqrt{x + M} + \sqrt{1 - x}\sqrt{1 - x - M} \right),$$

and elementary calculus yields that the right-hand side is always smaller than

$$\|f\|_\infty \sqrt{1 - M^2}\,.$$

The case $\phi(x) = x - M$ is handled similarly and yields the same upper-bound.

We thus have that, for all $x \in [0, 1]$,

$$U_f(\phi) \le \|f\|_\infty \sqrt{1 - |\phi(x) - x|^2},$$

and taking the infimum of the upper-bound over all x yields the conclusion of the lemma. □

Now, if $U_f^* = \sup\{U_f(\phi), \phi \in \mathcal{D}^*\}$, we always have

$$U_f^* \ge U_f(\mathrm{id}) = \int_0^1 f(x, x) dx = \Delta_f\,.$$

Thus, if $U_f(\phi) = U_f^*$, we have

$$\Delta_f \leq U_f(\phi) \leq \|f\|_\infty \sqrt{1 - \|\phi - \mathrm{id}\,\|_\infty^2}\,;$$

that is,

$$\|\phi - \mathrm{id}\,\|_\infty \leq \sqrt{1 - \left(\frac{\Delta_f}{\|f\|_\infty}\right)^2}\,,$$

which concludes the proof of Proposition 5.1.

5.4. Proof of Proposition 5.2. Let $\phi \in \mathcal{D}^*$ such that $U_f(\phi)$ is maximal. We denote by m the Lebesgue's measure on $[0,1]$. Proposition 5.2 is an obvious consequence of the two following lemmas.

Lemma 5.18. *For any* $0 \leq a < b \leq 1$*, we have* $\int_a^b \dot\phi(x)dx > 0$*, i.e.,* $\phi \in \mathcal{D}_+^*$*.*

Proof. Let us assume that Ω_f has a nonempty interior; that is, $\Delta_f < \|f\|_\infty$. (If this is not the case, Proposition 5.1 implies that the only maximizer is $\phi = \mathrm{id}\,$, which trivially belongs to $\mathcal{D}_+^*$.)

First, let us prove that $U_f(\phi) > 0$. Indeed, from condition [H2], we get that there exists a point (y_0, x_0) in the interior of $[0,1]^2$ such that $f_s(y_0, x_0) > 0$. Since f_s is lower-semicontinuous, f_s is strictly positive in a small neighborhood of (y_0, x_0). Now, define $\tilde\phi$ such that $\tilde\phi(0) = 0$, and $\tilde\phi(1) = 1$, $\tilde\phi(x_0) = y_0$, and $\tilde\phi$ is linear on $[0, x_0]$ and $[x_0, 1]$ (by Lemma 5.16). Since $\tilde\phi$ is strictly increasing on $[0,1]$, we deduce from [H1] that except possibly for a finite number of x, $(\phi(x), x) \notin F$ so that $f(\phi(x), x) = f_s(\phi(x), x)$. Hence, $U_f(\phi) \geq U_f(\tilde\phi) = U_{f_s}(\tilde\phi) > 0$.

Now assume that there exists $0 \leq a < b \leq 1$ such that $\int_a^b \dot\phi(x)dx = 0$. Let $a' = \inf\{z \in [0,a] \mid \int_z^b \dot\phi(x)dx = 0\}$ and $b' = \sup\{\, z \in [b,1] \mid \int_b^z \dot\phi(x)dx = 0\,\}$. We have $\int_{a'}^{b'} \dot\phi(x)dx = 0$, and since $U_f(\phi) > 0$, we have $a' > 0$ or $b' < 1$. Assume that $b' < 1$. (The case $a' > 0$ can be handled similarly.)

Now, for any $\eta > 0$ such that $b' + \eta \leq 1$, and any $\alpha \in]0,1[$, let us define

$$\omega^{\alpha,\eta}(x) = \dot\phi(x)\mathbf{1}_{x \notin [a', b'+\eta]} + (1-\alpha)\dot\phi(x)\mathbf{1}_{x \in]b', b'+\eta]} + \alpha\frac{K_\phi^\eta}{b'-a'}\mathbf{1}_{[a',b']},$$

where $K_\phi^\eta = \int_{a'}^{b'+\eta} \dot\phi(x)dx = \int_{b'}^{b'+\eta} \dot\phi(x)dx$. Let $\mu_{\alpha,\eta} = \omega^{\alpha,\eta}m + \nu_\phi$. One has $\mu_{\alpha,\eta} \in \mathcal{M}_1$, and, letting $\phi_{\alpha,\eta}(x) = \mu_{\alpha,\eta}([0,x[)$, one has

$$|\phi_{\alpha,\eta}(x) - \phi(x)| = \alpha\left|\int_0^x \left(\frac{K_\phi^\eta}{b'-a'}\mathbf{1}_{[a',b']} - \dot\phi\mathbf{1}_{]b',b'+\eta]}\right)\right| \leq \alpha \int_0^1 \dot\phi \leq \alpha,$$

so that $\|\phi_{\alpha,\eta} - \phi\|_\infty \leq \alpha$.

Let us show that $\phi(a') = \phi(b')$. Let R be the rectangle containing the points (y, x) such that $x \in]a', b'[$ and $y \in]\phi(a'), \phi(b')[$. If $\phi(a') < \phi(b')$, then this rectangle has a nonempty interior. Since $(\phi(a'), a') \in \Omega_f$ and $(\phi(b'), b') \in \Omega_f$, the intersection $R \cap \Omega_f$ also has a nonempty interior and, in particular, contains horizontal segments on which f cannot identically vanish. Thus, if $\phi(a') < \phi(b')$, there exist $x_0 \in]a', b'[$ and $y_0 \in\,]\phi(a'), \phi(b')[$ such that $f_s(y_0, x_0) > 0$, and this in turn implies that f_s is strictly positive in a small neighborhood of (y_0, x_0). Now considering $\tilde\phi$ such that $\tilde\phi(x) = \phi(x)$ on $[0, a'] \cup [b', 1]$, $\tilde\phi(x_0) = y_0$, and $\tilde\phi$ is linear on $[a', x_0]$ and on $[x_0, b']$, we have (using the same argument as in the beginning of the proof) $U_f(\phi) < U_f(\tilde{\phi})$ with $\tilde{\phi} \in \mathcal{D}^*$, which is a contradiction.

Since $\phi(a') = \phi(b')$, the segment with end points $(\phi(a'), a')$ and $(\phi(b'), b')$ is vertical and clearly lies in Ω_f, so that by condition [H2] there exists $x_0 \in]a', b'[$ such that $f_s(\phi(x_0), x_0) > 0$. Using the fact that $\|\phi_{\alpha,\eta} - \phi\|_\infty \leq \alpha$, we deduce that there exist $\delta > 0$ and $c > 0$ such that for any sufficiently small α and any $x \in [0,1]$, except eventually a finite number (we use here the fact that $\phi_{\alpha,\eta}$ is strictly increasing on $[a', b']$ and condition [H1]), if $|x - x_0| < \delta$, we have $f(\phi_{\alpha,\eta}(x), x) = f_s(\phi_{\alpha,\eta}(x), x) \geq c$.

Now we have

$$\begin{aligned} U_f(\phi_{\alpha,\eta}) - U_f(\phi) = &\int_{a'}^{b'} \sqrt{\frac{\alpha K_\phi^\eta}{b' - a'}} f(\phi_{\alpha,\eta}(x), x) dx \\ &+ \int_{b'}^{b'+\eta} \left[\sqrt{(1-\alpha)\dot{\phi}(x)} f(\phi_{\alpha,\eta}(x), x) - \sqrt{\dot{\phi}(x)} f(\phi(x), x) \right] dx. \end{aligned}$$

However, for α sufficiently small,

$$\int_{a'}^{b'} \sqrt{\frac{\alpha K_\phi^\eta}{b' - a'}} f(\phi_{\alpha,\eta}(x), x) dx \geq \left(\frac{2c\delta}{\sqrt{b' - a'}} \right) \sqrt{\alpha K_\phi^\eta},$$

and

$$\begin{aligned} &\left| \int_{b'}^{b'+\eta} \sqrt{(1-\alpha)\dot{\phi}(x)} f(\phi_{\alpha,\eta}(x), x) - \sqrt{\dot{\phi}(x)} f(\phi(x), x) dx \right| \\ &\qquad \leq 2\|f\|_\infty \int_{b'}^{b'+\eta} \sqrt{\dot{\phi}(x)} dx \leq 2\sqrt{\eta}\|f\|_\infty \sqrt{K_\phi^\eta}, \end{aligned}$$

so that

$$U_f(\phi_{\alpha,\eta}) - U_f(\phi) \geq \sqrt{K_\phi^\eta} \left(\frac{2c\delta}{\sqrt{b' - a'}} \sqrt{\alpha} - 2\sqrt{\eta}\|f\|_\infty \right).$$

Hence, choosing η sufficiently small, say $\eta < \frac{4\eta^2\delta^2\alpha}{(b'-a')\|f\|_\infty}$, we get $U_f(\phi_{\alpha,\eta}) - U_f(\phi) > 0$, which is a contradiction with the definition of ϕ as a maximizer. □

LEMMA 5.19. *For any* $a \in [0,1]$, $\nu_\phi(\{a\}) = 0$.

Proof. Let $\phi \in \mathcal{D}_+^*$ such that $U_f(\phi) = \max_{\mathcal{D}^*} U_f$. Proposition 5.7 implies that $U_{\tilde{f}}(\phi^-) = \max_{\mathcal{D}^*} U_{\tilde{f}}$. But if f satisfies conditions [H1] and [H2], so does $\tilde{f}$, and thus, one has $\phi^- \in \mathcal{D}_+^*$. Lemma 5.8 now implies that $\phi = (\phi^-)^-$ is continuous, which concludes the proof. □

6. Proof of the regularity results.

Proof of Theorem 3.3. The idea of the proof is to use the fact that after a proper change of variable and rescaling, ϕ^* is the solution of a local variational problem around any point x. Hence, the behavior of ϕ^* at x depends only on the properties of the locally optimal solutions involving the values of f in a small neighborhood of $(\phi^*(x), x)$.

Let $0 \leq a < b \leq 1$, and define for any $\phi \in \mathrm{Hom}^+$ the new "focusing" functions $\phi_{a,b} \in \mathrm{Hom}^+$ and $f_{a,b}^\phi : [0,1]^2 \to \mathbb{R}$ by

$$\phi_{a,b}(x') = \frac{\phi(a + x'(b-a)) - \phi(a)}{\phi(b) - \phi(a)} \; \forall x' \in [0,1],$$

and

$$f^{\phi}_{a,b}(y',x') = f(\phi(a) + y'(\phi(b) - \phi(a)), a + x'(b-a)).$$

One has, after a simple computation,

$$\int_a^b \sqrt{\dot\phi(x)} f(\phi(x),x)dx = \sqrt{(b-a)(\phi(b)-\phi(a))}U_{f^{\phi}_{a,b}}(\phi_{a,b})$$

so that

$$U_{f^{\phi^*}_{a,b}}(\phi^*_{a,b}) = \max_{\phi\in\mathrm{Hom}^+} U_{f^{\phi^*}_{a,b}}(\phi).$$

Let $\delta_{a,b} = \|f^{\phi^*}_{a,b}\|_\infty - \int_0^1 f^{\phi^*}_{a,b}(x',x')dx'$ and assume that $\|f^{\phi^*}_{a,b}\|_\infty > 0$; then we deduce from Theorem 3.1 that

$$\|\phi^*_{a,b} - \mathrm{Id}\|_\infty \le \sqrt{1 - \left(\frac{\|f^{\phi^*}_{a,b}\|_\infty - \delta_{a,b}}{\|f^{\phi^*}_{a,b}\|_\infty}\right)^2} \le \sqrt{2\frac{\delta_{a,b}}{\|f^{\phi^*}_{a,b}\|_\infty}}. \tag{6.1}$$

Hence, if $x_0 \in [a,b]$ and f is Hölder continuous with parameter $\alpha > 0$ at $(\phi^*(x_0),x_0)$ and if $f(\phi^*(x_0),x_0) > 0$, then we get easily that there exists a constant k_{x_0} (depending only on $(\phi^*(x_0),x_0)$) such that

$$\delta_{a,b} \le k_{x_0}\max(|\phi^*(b) - \phi^*(a)|^\alpha, (b-a)^\alpha)\,.$$

Using the fact that $\|f^{\phi^*}_{a,b}\|_\infty \ge f(\phi^*(x_0),x_0) > 0$, we deduce from inequality (6.1) that

$$\|\phi^*_{a,b} - \mathrm{Id}\|_\infty \le C_{x_0}(b-a)^{\alpha/2}\max\left(\left(\frac{|\phi^*(b)-\phi^*(a)|}{b-a}\right)^{\alpha/2}, 1\right), \tag{6.2}$$

with $C_{x_0} = \sqrt{\frac{2k_{x_0}}{f(\phi^*(x_0),x_0)}}$. Choosing $a = x_0$ and $b = x_0 + h$ with any $h > 0$ such that $x_0 + h \le 1$, we deduce from the previous inequality that, for all $u \in]0,1]$,

$$|\Delta\phi^*(x_0,uh) - \Delta\phi^*(x_0,h)| \le \Delta\phi^*(x_0,h)C_{x_0}\frac{h^{\alpha/2}}{u}\max(\Delta\phi^*(x_0,h)^{\alpha/2},1), \tag{6.3}$$

where for any $h' > 0$ we have $\Delta\phi^*(x_0,h') = (\phi^*(x_0+h') - \phi^*(x_0))/h'$. The fact that

$$\lim_{h\to 0, h'>0} \Delta\phi^*(x_0,h')$$

exists and is positive is a consequence of the following lemma, applied to $F(h) = \Delta\phi^*(x_0,h)$. (Note that $hF(h) \to 0$ if $h \to 0$, since ϕ^* is continuous.)

LEMMA 6.1. *Let $F > 0$ be a function defined on $]0,\beta]$ (for some $\beta > 0$) and such that, for all $h \in]0,\beta]$ and for all $u \in]0,1]$, and for some constants $K > 0$, $\rho > 0$, and $\beta > 0$,*

$$|F(uh) - F(h)| \le K.F(h)(1 + F(h)^\rho)\frac{h^\rho}{u}\,. \tag{6.4}$$

Assume, moreover, that $\lim_{h\to 0} hF(h) = 0$. Then, $\lim_{h\to 0} F(h)$ exists and is strictly positive.

Proof of Lemma 6.1. Let $h_0 \in]0, \beta]$, and let, for $n \geq 1$, $v_n = F(h_0 2^{-n})$. From (6.4), we get, for some constant K',

$$|v_{n+1} - v_n| \leq K' h_0^\rho v_n (1 + v_n^\rho) 2^{-\rho n} \,. \tag{6.5}$$

Clearly, to prove that v_n converges, it suffices to prove that it is bounded. In fact, it merely suffices to prove that $v_n \leq C.2^{\gamma n}$ for some $\gamma < 1$, since, in this case, (6.5) yields an inequality of the kind

$$|v_{n+1} - v_n| \leq K'' v_n 2^{-\rho' n}$$

for some constants K'' and $\rho' > 0$, which implies in turn that v_n is bounded, since $\prod_{k=0}^{\infty} (1 + K'' 2^{-\rho' k}) < \infty$.

So, fix $\gamma < 1$ and let us prove that, if h_0 is taken to be small enough, one has, for all n, $v_n \leq F(h_0) 2^{\gamma n}$. Assuming that this is true for $n \geq 0$ (recall that $v_0 = F(h_0)$, so that it is true for $n = 0$), we show that this is true for $n + 1$. We have

$$v_{n+1} \leq F(h_0).2^{\gamma n} (1 + K' h_0^\rho (1 + F(h_0)^\rho 2^{\rho \gamma n}) 2^{-\rho n}) \leq F(h_0) 2^{\gamma n} (1 + K' h_0^\rho (1 + F(h_0)^\rho))$$

so that it suffices to take h_0 such that $1 + K' h_0^\rho (1 + F(h_0)^\rho) < 2^\gamma$ to get the desired conclusion.

Thus, v_n converges to a limit v. But since (6.5) implies that

$$v_{n+1} \geq v_n (1 - K' h_0^\rho 2^{-\rho n}),$$

we have, letting n_0 such that $K' h_0^\rho 2^{-\rho n_0} < 1$, for all $n \geq n_0$,

$$0 < v_{n_0} \prod_{k=n_0}^{\infty} (1 - K' h_0^\rho 2^{-\rho k}) \leq v_n, \tag{6.6}$$

which imples that $v > 0$.

Now, if h_n is any sequence which tends to 0 from above, one can find, for all n, an integer k_n such that $h_0 2^{-k_n - 1} < h_n \leq h_0 2^{-k_n}$, and (6.4) implies that

$$|F(h_n) - v_{k_n}| \leq K' |v_{k_n}| 2^{-\rho k_n}$$

so that, since k_n tends to infinity, $F(h_n)$ tends to v, which proves Lemma 6.1. □

We thus have proved that ϕ^* has a strictly positive right derivative denoted $\dot{\phi}_r^*(x_0)$. In the same way, we can prove that the left derivative denoted $\dot{\phi}_l^*(x_0)$ exists and is stricly positive. It thus remains to prove that both derivatives coincide. In fact, relation (6.2) with $a = x_0 - h$ and $b = x_0 + h$ yields

$$\lim_{h \to 0} \left| \frac{\phi^*(x_0) - \phi^*(x_0 - h)}{\phi^*(x_0 + h) - \phi^*(x_0 - h)} - \frac{1}{2} \right| = 0 \,.$$

Since the left-hand part of the inequality also tends to $\frac{\dot{\phi}_l^*}{\dot{\phi}_l^* + \dot{\phi}_r^*} - \frac{1}{2}$, we get the result. Hence the first part of the theorem is proved.

Now, if f is locally uniformly Hölder continuous at x_0, there exists (since ϕ^* is continuous) an $\epsilon > 0$ in $[0, 1]$ such that (3.1) holds at point $(\phi(x), x)$ for all x such that $|x - x_0| < \epsilon$. As a consequence, ϕ^* will thus be differentiable at all such $x \in]x_0 - \epsilon, x_0 + \epsilon[$ and the increments $\Delta\phi^*(x, h)$ will converge, as $h \to 0$, uniformly to $\dot{\phi}^*(x)$. Since these increments are continuous, $\dot{\phi}^*$ is also continuous on $]x_0 - \epsilon, x_0 + \epsilon[$. □

Proof of Theorem 3.4. By Theorem 3.3, ϕ is continuously differentiable. Moreover, one has, for any ψ smooth diffeomorphism of $[0, 1]$,

$$\int_0^1 \sqrt{\dot{\phi}}(x)\sqrt{\dot{\psi}}(x) f(\phi(x), \psi(x))dx \leq \int_0^1 \sqrt{\dot{\phi}}(x) f(\phi(x), x)dx \tag{6.7}$$

(simply using that the left-hand term is $U_f(\phi \circ \psi^{-1})$). If h is any smooth function in $[0, 1]$ such that $h(0) = h(1) = 0$, there exists a small enough t such that $\psi(x) = x + th(x)$ is a diffeomorphism, and, after computation of the first variation in the left-hand term of (6.7), one gets that, for all smooth h with $h(0) = h(1) = 0$,

$$\int_0^1 \sqrt{\dot{\phi}}(x)\dot{h}(x) f(\phi(x), x) = -2\int_0^1 \sqrt{\dot{\phi}}(x)\frac{\partial f}{\partial y}(\phi(x), x).h(x)dx.$$

So, letting $q(x) = \sqrt{\dot{\phi}}(x) f(\phi(x), x)$, one has

$$q(x) = q(0) + 2\int_0^x \sqrt{\dot{\phi}}(u)\frac{\partial f}{\partial y}(\phi(u), u)du$$

so that q is differentiable in the ordinary sense, with derivative $2\sqrt{\dot{\phi}}\frac{\partial f}{\partial y}(\phi(.), .)$, which is continuous. Since

$$\dot{\phi}(x) = \frac{q^2(x)}{f(\phi(x), x)},$$

the numerator being positive and continuously differentiable, we get the fact that $\dot{\phi}$ is continuously differentiable. □

7. Auxiliary results.

We conclude this paper with two simple results which have important practical applications. The first one validates the possibility of implementing a matching combined with the fitting of some registration parameters. This enables us to recover some invariance properties which have not directly been incorporated in F.

The second result provides an approximation scheme, which permits us to work safely with discretized versions of a signal. It also naturally yields consistent multiscale minimization procedures, which is important for efficiency of numerical implementations.

7.1. Handling additional parameters.

In many practical situations, a matching is searched for up to some given finite-dimensional parameter which performs some registration between the two quantities which are compared. For example, in the formulation $f(\phi(x), x) = F(\theta \circ \phi(x), \theta'(x))$, one may consider that the functions θ should be identified to $\theta + b$ for any $b \in \mathbb{R}$ (in order to get a translation invariant matching), so that the complete problem becomes maximizing

$$\int_0^1 \sqrt{\dot{\phi}} F(\theta \circ \phi(x) + b, \theta'(x))dx$$

over all ϕ and b. For example, in [9], translation on θ represented rotations of plane curves, rotation-invariant comparison being a desirable feature for shape comparison.

More generally, we shall deal in this section with a function f, which depends on an additional extraneous parameter $\lambda \in \mathbb{R}^d$, and we shall try to find ϕ^* and λ^* which maximize

$$V_f(\phi, \lambda) = \int_0^1 \sqrt{\dot{\phi}(x) f(\phi(x), x, \lambda)} dx$$

for $\phi \in \mathcal{D}$ and $\lambda \in K$, where K is a compact subset of $\mathbb{R}^2$. Sufficient conditions for existence are provided in the following theorem. We let f_λ be the function $(x, y) \mapsto f(x, y, \lambda)$.

THEOREM 7.1. *We assume that f is continuous in $\lambda \in K$, uniformly in (x, y), and that, for all λ, the function f_λ satisfies conditions* [H1] *and* [H2] *of Theorem* 3.1. *Then, there exist $\lambda^* \in K$ and $\phi^* \in Hom^+$ such that*

$$V_f(\phi^*, \lambda^*) = \max\{V_f(\phi, \lambda) \mid \phi \in Hom^+, \lambda \in K\}.$$

Proof. By Theorem 3.1, for all $\lambda \in K$, the functional $\phi \mapsto V_f(\phi, \lambda)$ is upper-semicontinuous in $\phi \in \mathcal{D}^*$. Moreover, it is uniformly continuous in λ, since

$$\begin{aligned}|V_f(\phi, \lambda) - V_f(\phi, \lambda')| &\leq \int_0^1 \sqrt{\dot{\phi}} |f(\phi(x), x, \lambda) - f(\phi(x), x, \lambda')| dx \\ &\leq \sup_{x,y} |f(x, y, \lambda) - f(x, y, \lambda')|,\end{aligned}$$

which tends to 0 if λ tends to λ'. This implies that U is upper-semicontinuous as a function of the two variables ϕ and λ, and thus that there exists a maximizer $(\phi^*, \lambda^*) \in \mathcal{D}^* \times K$. Now, since ϕ^* is a maximizer of $U_f(., \lambda^*)$ over $\mathcal{D}^*$, Proposition 5.2 implies that $\phi^* \in \mathrm{Hom}^+$. □

7.2. Approximation schemes.

THEOREM 7.2. *Let $(f_n, n \geq 0)$ and f be functions defined on $[0, 1]^2 \times K$ such that*

$$\lim_{n \to \infty} \sup_{x,y,\lambda} |f_n(x, y, \lambda) - f(x, y, \lambda)| = 0.$$

Assume that all f_n and f satisfy the conditions of Theorem 7.1. *Let (ϕ_n^*, λ_n^*) be maximizers of V_{f_n} over $\mathcal{D}^* \times K$; then, there exists a subsequence of (ϕ_n^*, λ_n^*) which converges in $\mathcal{D}^* \times K$ to a maximizer (ϕ^*, λ^*) of V_f.*

Proof. Indeed, the hypotheses trivially imply that U_n converges to U uniformly on $\mathcal{D}^* \times K$, so that $\max U_n \to \max U$ and

$$\lim_{n \to \infty} U(\phi_n^*, \lambda_n^*) = \max U\,.$$

Now from (ϕ_n^*, λ_n^*) one can extract a converging subsequence in the compact space $\mathcal{D}^* \times K$ to a limit denoted (ϕ, λ), and one must have $U(\phi, \lambda) = \max U$ because U is upper-semicontinuous. □

One application of the theorem is the following. Assume, for example, that f is continuous and f_λ satisfies condition [H2] of Theorem 3.1 for all $\lambda \in K$. Let $g_n(x, y, \lambda)$ be a piecewise constant approximation of f, and let $f_n = \epsilon_n + g_n$, where ϵ_n is a sequence which tends to 0. Then, each f_n satisfies the conditions of Theorem 7.1 so that Theorem 7.2 applies. Such a situation is typical in numerical procedures.

REFERENCES

[1] B. Brezis, *Analyse fonctionnelle, théorie et applications*, Masson, Paris, 1983 (English translation: Springer-Verlag).

[2] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1989.

[3] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.

[4] P. Olver, *Equivalence, Invariants, and Symmetry*, Cambridge University Press, Cambridge, UK, 1995.

[5] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

[6] M. Piccioni, S. Scarlatti, and A. Trouvé, *A variational problem arising from speech recognition*, SIAM J. Appl. Math., 58 (1998), pp. 753–771.

[7] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, Toronto, London, 1966.

[8] A. Trouvé and L. Younes, *Diffeomorphic matching problems in one dimension: Designing and minimizing matching functionals*, in Proceedings of the 6th European Conference on Computer Vision, Trinity College, Dublin, Ireland, 2000, pp. 573–587.

[9] L. Younes, *Computable elastic distances between shapes*, SIAM J. Appl. Math, 58 (1998), pp. 565–586.

[10] L. Younes, *Optimal matching between shapes via elastic deformations*, Image and Vision Computing Journal, 17 (1999), pp. 381–389.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1136–1159

PROXIMAL POINT APPROACH AND APPROXIMATION OF VARIATIONAL INEQUALITIES*

A. KAPLAN† AND R. TICHATSCHKE†

Abstract. A general approach to analyze convergence of the proximal-like methods for variational inequalities with set-valued maximal monotone operators is developed. It is oriented to methods coupling successive approximation of the variational inequality with the proximal point algorithm as well as to related methods using regularization on a subspace and weak regularization. This approach also covers so-called multistep regularization methods, in which the number of proximal iterations in the approximated problems is controlled by a criterion characterizing these iterations as to be effective. The conditions on convergence require control of the exactness of the approximation only in a certain region of the original space. Conditions ensuring linear convergence of the methods are established.

Key words. variational inequalities, monotone operators, convex programming, proximal point methods, weak regularization

AMS subject classifications. 47H05, 47H19, 65J20

PII. S0363012998333116

1. Introduction.

Variational inequalities with maximal monotone operators include convex programs, convex-concave saddle point problems, equations, and inclusions with maximal monotone operators and a series of other problems.

The proximal point method, introduced by Martinet [28] and later investigated in a more general setting by Rockafellar [34], has initiated a couple of new algorithms for solving these problems. We refer to [2], [14], [35], [42] for algorithms based on multiplier methods and to [1], [4], [15], [16], [17] for modifications of the penalty technique. Proximal variants of bundle methods for nonsmooth convex optimization problems are given in [3], [22], [25], [29], and for partial inverses of monotone operators as well as for decomposition and parallel optimization methods; see [5], [6], [9], [39], and [40]. The number of papers dealing with the proximal point approach is growing fast and we can cite here only a few of them. Moreover, it is known (see [16], [24], [34]) that some classical numerical methods can be interpreted as special applications of this approach. In [8] this fact was established concerning the *Douglas–Rachford splitting method* for finding a zero of the sum of two monotone operators. This points out new applications of the proximal technique, in particular, to problems in mathematical physics; for such problems, see, for instance, [26].

The basic results of Rockafellar [34] on convergence of the proximal point method for solving variational inequalities with maximal monotone operators were generalized in [27] concerning the rate of convergence, and in [11] a similar analysis was made for methods using the proximal technique on a subspace. More precisely, in the papers mentioned the methods were studied for the equivalent problem of finding a zero of a maximal monotone operator under the assumption that the proximal iterations can be performed inexactly. However, an approximation of the data of the problem was not considered there.

*Received by the editors March 16, 1998; accepted for publication (in revised form) June 19, 2000; published electronically December 5, 2000.

http://www.siam.org/journals/sicon/39-4/33311.html

†Department of Mathematics, University of Trier, D-54286 Trier, Germany (a.a.kaplan@tiscalinet.de, tichat@uni-trier.de).

In this paper the convergence analysis covers methods which couple a successive approximation of the variational inequality with the proximal point approach, as well as related methods using a regularization on a subspace or a weak regularization.

The problem under consideration is the following variational inequality:

$$\text{(1.1)} \qquad \text{find } u \in K \text{ such that } \exists\, y \in \mathcal{T}u: \quad \langle y, v-u\rangle \geq 0 \;\; \forall v \in K,$$

with K a convex, closed subset of a Hilbert space V, $\mathcal{T} : V \to 2^{V'}$ a monotone operator, $D(\mathcal{T}) \equiv \{v \in V : \mathcal{T}v \neq \emptyset\} \supset K$, V' the dual space of V, and $\langle\cdot,\cdot\rangle$ the duality pairing between V and V'.

We recall that an operator $\mathcal{A} : V \to 2^{V'}$ is said to be *monotone* if

$$\langle w-z, u-v\rangle \geq 0 \;\text{ whenever }\; w \in \mathcal{A}u,\; z \in \mathcal{A}v,$$

and *strongly monotone* if

$$\langle w-z, u-v\rangle \geq \gamma\|u-v\|^2 \;\text{ with some } \gamma > 0.$$

A monotone operator $\mathcal{A}$ is called *maximal monotone* if its graph is not properly contained in the graph of any other monotone operator $\mathcal{A}^1 : V \to 2^{V'}$.

Throughout the whole paper it is supposed that H is a given Hilbert space such that V can be continuously embedded into H, V_1 is a given closed subspace of V, and $\mathcal{P} : V \to V_1$ is the orthogonal projection operator (orthoprojector). If V_1 is also closed in H, then $\mathcal{P}$ can be defined as the orthoprojector according to the norm of H. In the framework of the approach considered, the solution of (1.1) is obtained by solving approximately the sequence of variational inequalities

$$\text{(1.2)} \qquad \begin{aligned} &u \in K_i : \langle \mathcal{T}_i u, v-u\rangle + \chi_i(\mathcal{P}u - \mathcal{P}u^{i,s-1}, \mathcal{P}v - \mathcal{P}u)_H \geq 0 \;\; \forall v \in K_i,\\ &\qquad s = 1,\dots,s(i);\; i = 1,2,\dots, \end{aligned}$$

with $\mathcal{T}_i : V \to V'$, and K_i certain approximations for $\mathcal{T}$ and K, respectively, $(\cdot,\cdot)_H$ the inner product in H, $u^{i,s-1}$ a solution of the previous problem ($u^{i,0} \equiv u^{i-1,s(i-1)}$), $0 < \chi_i \leq \bar{\chi}$.

For example, treating variational problems in mechanics, often K_i is an internal approximation of K constructed by means of a finite element method, and $\mathcal{T}_i$ is the gradient of a smoothened energy functional (cf. the approximation of the linear model of elasticity with friction in [10, section 4.2] that corresponds formally to (1.2) with $\chi_i \equiv 0$; by the way, we avoid the identification of V and V' taking into account the standard estimation technique in finite element methods).

The choice of the space H as well as of the subspace $V_1 \subset V$ depends on specific properties of the problem under consideration. In particular, this choice has to ensure the strong monotonicity of the operators $\mathcal{T}_i + \chi_i\mathcal{M}$, where $\mathcal{M} : V \to V'$ is defined by $\langle \mathcal{M}u, v\rangle = (\mathcal{P}u, \mathcal{P}v)_H \; \forall u, v \in V$. For variational inequalities studied in [19], the space V is of the type $H^1(\Omega)$ and the operator $\mathcal{T}$ is not strongly monotone in V. However, (weak) regularization with $V_1 = V$, $H = L_2(\Omega)$ ensures strong monotonicity of the operators $\mathcal{T} + \chi_i\mathcal{M}$, $\mathcal{T}_i + \chi_i\mathcal{M}$ and weak convergence of the iterates in V, i.e., it leads to the same quality of convergence as in the case of applying the standard (strong) proximal regularization with $V_1 = V$, $H = V$.

The situation, that a subspace $V_1 \neq V$ with appropriate properties is known, is typical for problems in mathematical physics. Referring again to [19], for a two-body contact problem considered there, V_1 is a known 3-dimensional subspace of the rigid displacements of one of the bodies.

A suitable choice of V_1 and H can accelerate essentially the numerical calculation (see the analysis of some examples in [16], [18], and numerical experiments with real-life problems in [36], [37].)

The strong formalization of method (1.2) is described in section 2 as *multistep regularization method* (MSR-method). The notion "multistep regularization" reflects the presence of an inner cycle (with respect to s) of proximal iterations.

The case that $\mathcal{T}$ is a subdifferential of a convex functional, $H = V = V_1$, and $s(i) = 1\ \forall i$, includes several "diagonal" variants of the proximal point method for convex optimization (cf. [16, sections 9 and 12], [23] and the references therein).

There are a lot of papers concerning the *diagonal approximation* of the ill-posed problem (1.1) with the use of the *Browder–Tikhonov regularization.* In this case the auxiliary problems have the form

$$u \in K_i \ :\quad \langle \mathcal{T}_i u + \chi_i \mathcal{M}_0(u - \hat{u}), v - u\rangle \geq 0 \quad \forall v \in K_i,$$

where $\mathcal{M}_0 : V \to V'$ is a strongly monotone operator, $\hat{u} \in V$ is a fixed element, and $\chi_i > 0, \lim\ \chi_i = 0$. Fundamental results in this direction were obtained by Mosco [30]; for concrete algorithms see Vasil'ev [41].

In as much as $\chi_i \to 0$ is not necessary for the convergence of proximal methods, they possess a better stability and provide a better efficiency of fast convergent algorithms solving the regularized auxiliary problems.

In methods of type (1.2) the iterations with respect to s (for a fixed approximation level i) continue "as long as they remain effective" (see Remark 3.12 below). The MSR-methods were developed in [16] for convex variational problems, i.e., in the case that $\mathcal{T}$ is a subdifferential of a convex functional. In [19], [12], [13], and [20] they were adapted to some problems in elasticity theory and optimal control with PDEs. In comparison with diagonal proximal processes, the MSR-methods allow better approximations of a sought solution of the original problem under the same approximate data $\mathcal{T}_i,\ K_i$ and the same χ_i, ϵ_i, so that the numerical expense can be reduced. This has been confirmed, in particular, by numerical experiments for Bingham problems in [37] and for optimal control problems governed by elliptic equations in [36].

Here we investigate the convergence of such methods in the more general setting that the operator $\mathcal{T}$ may be nonpotential. This permits consideration of new applications, in particular, saddle point problems and complementarity problems. The principal distinction to the convergence analysis made in [16] is caused by the fact that the condition about the uniform approximation of the objective functional used in [16] cannot be adapted in order to describe the closeness between $\mathcal{T}$ and $\mathcal{T}_i$.

In comparison to the present paper, in [21] the convergence (without rate of convergence) of scheme (1.2) is investigated for the particular case $H = V = V_1$ (hence, $\mathcal{P}$ is the identity operator) and under more restrictive assumptions concerning the approximation of the operator $\mathcal{T}$ and the set K; cf. conditions (c$'$), (d$'$), (e$'$) below in Remark 3.5(ii) and Assumptions 3.4(c), (d), (e).

The paper is organized as follows. Section 2 contains the description of a generalized MSR-scheme. Its convergence is studied in section 3 and estimates of the rate of convergence for the basic variant (with $H = V = V_1$) are given in section 4.

We do not consider here the adaptation of the generalized MSR-scheme to particular problems. This will be the object of a forthcoming paper. For some applications of the basic variant to variational inequalities with potential and nonpotential operators $\mathcal{T}$ we refer to [21].

2. Multistep proximal regularization scheme. We denote by $\|\cdot\|, \|\cdot\|_{V'}$, and $\|\cdot\|_H$ the norms in V, V', and H, respectively. With z, C, D from V (respectively, from V'), let

$$dist(z,D) = \inf_{w\in D} \|z-w\|, \quad dist(C,D) = \sup_{z\in C} dist(z,D),$$
$$(\text{resp., } dist_{V'}(z,D) = \inf_{w\in D} \|z-w\|_{V'},\ dist_{V'}(C,D) = \sup_{z\in C} dist_{V'}(z,D)).$$

The following assumption concerns the chosen triple H, V_1, and $\mathcal{P}$.

ASSUMPTION 2.1. *For a given linear, continuous, and monotone operator $\mathcal{B} : V \to V'$ with the symmetry property*

$$\langle \mathcal{B}u, v\rangle = \langle \mathcal{B}v, u\rangle \quad \forall u, v \in V$$

and a given $\tilde{\chi} > 0$ there exists $\beta_0 > 0$ such that

$$\tilde{\chi}\langle \mathcal{B}u, u\rangle + \|\mathcal{P}u\|_H^2 \geq \beta_0\|u\|^2 \quad \forall u \in V. \tag{2.1}$$

If this assumption is valid, one can introduce on V a new norm $|||\cdot|||$ by

$$|||u|||^2 = \tilde{\chi}\langle \mathcal{B}u, u\rangle + \|\mathcal{P}u\|_H^2, \tag{2.2}$$

which is equivalent to the original one $\|\cdot\|$ because of

$$\beta_0\|u\|^2 \leq |||u|||^2 \leq (\beta_1\tilde{\chi} + \beta_2^2)\|u\|^2, \tag{2.3}$$

with $\beta_1 = \sup_{u\neq 0} \frac{\|\mathcal{B}u\|_{V'}}{\|u\|}$ and $\beta_2 : \|u\|_H \leq \beta_2\|u\| \quad \forall u \in V$.

Denote $S_\tau = \{u \in V : |||u||| \leq \tau\}$. Let $\{\mathcal{T}_i\}, \mathcal{T}_i : V \to V'$, be a sequence of monotone and hemicontinuous (i.e., weak continuous along each line segment in V) operators, approximating $\mathcal{T}$, and $\{K_i\}$, $K_i \subset D(\mathcal{T}_i) \subset V$, be a sequence of convex closed sets, approximating K.

Henceforth it is supposed that the solution set U^* of problem (1.1) is nonempty, Assumption 2.1 is fulfilled, and radii r and r^* are chosen such that

$$U^* \cap S_{r^*/8} \neq \emptyset \quad \text{and} \quad r \geq r^*. \tag{2.4}$$

MULTISTEP REGULARIZATION METHOD.
Let $\{\chi_i\}, \{\epsilon_i\}, \{\delta_i\}$ be positive controlling sequences with

$$\sup_i \ \chi_i \leq \bar{\chi} < \infty, \quad \lim_{i\to\infty} \ \epsilon_i = 0,$$

and $u^0 \in S_{r^*/4}$.
Step i: *Given u^{i-1}.*
(a) *Set $u^{i,0} := u^{i-1}$, $s := 1$.*
(b) *Given $u^{i,s-1}$, define*

$$u^{i,s} : \quad \|u^{i,s} - \bar{u}^{i,s}\| \leq \epsilon_i, \tag{2.5}$$

where $\bar{u}^{i,s}$ is the exact solution of the variational inequality

$$u \in K_i : \langle \mathcal{T}_i u, v - u\rangle + \chi_i(\mathcal{P}u - \mathcal{P}u^{i,s-1}, \mathcal{P}v - \mathcal{P}u)_H \geq 0 \quad \forall v \in K_i. \tag{2.6}$$

(c) *If* $\|\mathcal{P}u^{i,s} - \mathcal{P}u^{i,s-1}\|_H > \delta_i$, *then set* $s := s+1$ *and repeat* (b).
Otherwise, set $s(i) := s, u^i := u^{i,s}, i := i+1$ *and repeat Step* i.

The choice $H = V = V_1$ corresponds to the *basic variant of the MSR-method* [16], and $V_1 \neq V$ reflects the *(MSR-) method with regularization on a subspace* (cf. [16], [12], [13], [20]). If $V_1 = V$ and $\| \cdot \|_H$ is weaker than $\| \cdot \|$, one deals with the *(MSR-) method with a weak regularization* (see [16], [19], [37]). For the diagonal method with regularization on a subspace we refer to [11] and to Rockafellar's interpretation for the multiplier method in [35].

Of course, condition (2.5) is not a practicable criterion. However, it is convenient to unify the investigation of algorithms with different stopping rules for the auxiliary problems: in fact, (2.6) is a variational inequality with a strongly monotone operator (see Remark 3.5(i)), and therefore, criterion (2.5) can be satisfied by means of the own stopping rule of an algorithm inserted into the MSR-method to solve the auxiliary problems (2.6). In Remark 3.14 we discuss the use of the criterion

$$\begin{aligned}
&u^{i,s} \in K_i : \\
&\langle \mathcal{T}_i u^{i,s}, v - u^{i,s}\rangle + \chi_i \left(\mathcal{P}u^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}v - \mathcal{P}u^{i,s}\right)_H \geq -\epsilon_i' \|v - u^{i,s}\| \\
&\forall v \in K_i.
\end{aligned}$$

One should refer also to [7], [38] for practicable notions of inexact proximal iterations.

3. Convergence analysis. According to Riesz's representation theorem, the variational inequality (2.6) can be rewritten in the form

$$u \in K_i : \langle \mathcal{T}_i u + \chi_i \mathcal{M}(u - u^{i,s-1}), v - u\rangle \geq 0 \quad \forall v \in K_i, \tag{3.1}$$

where the linear operator $\mathcal{M} : V \to V'$ is defined by

$$\langle \mathcal{M}u, v\rangle = (\mathcal{P}u, \mathcal{P}v)_H \quad \forall u, v \in V.$$

In order to prove convergence of this method, we need some auxiliary statements.

LEMMA 3.1. *Let $G \subset V$ be a closed convex set, and let $\mathcal{A} : V \to V'$ be a single-valued operator with $D(\mathcal{A}) \supset G$. Suppose that the inequality*

$$\langle \mathcal{A}u - \mathcal{A}v, u - v\rangle \geq \langle \mathcal{B}u - \mathcal{B}v, u - v\rangle \quad \forall u, v \in G \tag{3.2}$$

is valid with $\mathcal{B}$ as in Assumption 2.1. *Moreover, for arbitrary $a^0 \in V$ and $\chi \in (0, 2\tilde{\chi}^{-1}]$, let a^1 be a solution of the variational inequality*

$$u \in G \ : \ \langle \mathcal{A}u, v - u\rangle + \chi(\mathcal{P}u - \mathcal{P}a^0, \mathcal{P}v - \mathcal{P}u)_H \geq 0 \quad \forall v \in G. \tag{3.3}$$

Then the inequality

$$|||a^1 - v|||^2 - |||a^0 - v|||^2 \leq -\|\mathcal{P}a^1 - \mathcal{P}a^0\|_H^2 + \frac{2}{\chi}\langle \mathcal{A}v, v - a^1\rangle \tag{3.4}$$

is true for each $v \in G$.

Proof. From the definition of $||| \cdot |||$ we obtain

$$\begin{aligned}
&|||a^1 - v|||^2 - |||a^0 - v|||^2 = -\|\mathcal{P}a^1 - \mathcal{P}a^0\|_H^2 \\
&-2(\mathcal{P}a^1 - \mathcal{P}a^0, \mathcal{P}v - \mathcal{P}a^1)_H \\
&+\tilde{\chi}\langle \mathcal{B}(a^1 - v), a^1 - v\rangle - \tilde{\chi}\langle \mathcal{B}(a^0 - v), a^0 - v\rangle.
\end{aligned}$$

Using this inequality together with (3.3) one can conclude that

$$\begin{aligned} &|||a^1 - v|||^2 - |||a^0 - v|||^2 \le -\|\mathcal{P}a^1 - \mathcal{P}a^0\|_H^2 + \frac{2}{\chi}\langle \mathcal{A}a^1, v - a^1\rangle \\ &+\tilde{\chi}\langle \mathcal{B}(a^1 - v), a^1 - v\rangle - \tilde{\chi}\langle \mathcal{B}(a^0 - v), a^0 - v\rangle, \end{aligned} \tag{3.5}$$

and, due to (3.2) and $0 < \chi \le 2\tilde{\chi}^{-1}$, this yields

$$\begin{aligned} |||a^1 - v|||^2 - |||a^0 - v|||^2 &\le -\|\mathcal{P}a^1 - \mathcal{P}a^0\|_H^2 - \frac{2}{\chi}\langle \mathcal{A}v - \mathcal{A}a^1, v - a^1\rangle \\ &+\frac{2}{\chi}\langle \mathcal{A}v, v - a^1\rangle + \tilde{\chi}\langle \mathcal{B}(v - a^1), v - a^1\rangle - \tilde{\chi}\langle \mathcal{B}(a^0 - v), a^0 - v\rangle \\ &\le -\|\mathcal{P}a^1 - \mathcal{P}a^0\|_H^2 + \frac{2}{\chi}\langle \mathcal{A}v, v - a^1\rangle. \qquad \square \end{aligned}$$

LEMMA 3.2. *Let $G \subset V$, $G_1 \subset V$, and G be a convex closed set. For a given $y \in Y'$, let a solution $\bar{u}$ of the problem*

$$\textit{find} \quad u \in G \cap G_1 \ : \ \langle y, v - u\rangle \ge 0 \quad \forall v \in G \cap G_1 \tag{3.6}$$

belong to $\mathrm{int}G_1$. *Then $\bar{u}$ is also a solution of the problem*

$$\textit{find} \quad u \in G \ : \ \langle y, v - u\rangle \ge 0 \quad \forall v \in G. \tag{3.7}$$

Proof. Take an arbitrary $w \in G\backslash G_1$. Because $\bar{u} \in \mathrm{int}G_1$, there exists $\lambda = \lambda(w) \in (0,1)$ such that $\bar{u} + \lambda(w - \bar{u}) \in G_1$; hence $\bar{u} + \lambda(w - \bar{u}) \in G \cap G_1$. Therefore, we get from (3.6) that

$$\langle y, \bar{u} + \lambda(w - \bar{u}) - \bar{u}\rangle \ge 0,$$

thus

$$\langle y, w - \bar{u}\rangle \ge 0 \quad \forall w \in G\backslash G_1, \tag{3.8}$$

and combining (3.6), (3.8) we obtain that $\bar{u}$ solves (3.7). $\square$

LEMMA 3.3. *Let $G \subset V$ be a closed convex set, and let $\mathcal{A} : V \to 2^{V'}$ be a maximal monotone operator with $D(\mathcal{A}) \supset G$, $G \cap \mathrm{int}D(\mathcal{A}) \ne \emptyset$. Assume that*

$$\sup_{v\in G} \sup_{y\in\mathcal{A}v} \|y\|_{V'} < \infty, \tag{3.9}$$

and for some $u \in G$ and each $v \in G$, there exists $y(v) \in \mathcal{A}v$ satisfying

$$\langle y(v), v - u\rangle \ge 0. \tag{3.10}$$

Then, with some $y \in \mathcal{A}u$, the inequality

$$\langle y, v - u\rangle \ge 0$$

holds for each $v \in G$.

Proof. Let $\mathcal{A}_1 v = \mathcal{A}v + \mathcal{I}(v - u)$, with $\mathcal{I} : V \to V'$ the operator of the canonical isometry. The operator $\mathcal{A}_1$ is strongly monotone, and due to Theorem 1 in [33], it is maximal monotone.

For fixed $q(u) \in \mathcal{A}u$, using the inequality

$$\begin{aligned}\langle q(v) + \mathcal{I}(v-u), v-u\rangle &\geq \langle q(u), v-u\rangle + \|v-u\|^2 \\ &\geq \|v-u\| \left(\|v-u\| - \|q(u)\|_{V'}\right),\end{aligned}$$

which holds for any $q(v) \in \mathcal{A}v$, one can conclude that

$$\langle q(v) + \mathcal{I}(v-u), v-u\rangle \geq 0 \tag{3.11}$$

if

$$\|v\| \geq \|u\| + \|q(u)\|_{V'}.$$

Taking into account (3.11) and $G \cap \mathrm{int} D(\mathcal{A}) \neq \emptyset$, the straightforward application[1] of Theorem 5 in [33] guarantees that the variational inequality

$$\text{find} \quad w \in G \quad \text{such that} \quad \exists\, y_1(w) \in \mathcal{A}_1 w \ : \ \langle y_1(w), v-w\rangle \geq 0 \quad \forall v \in G$$

is solvable. If $w = u$, then of course $y_1(w) \in \mathcal{A}u$; hence the statement of the lemma is true.

Otherwise, we use the relation

$$\langle \bar{y}(v), v-u\rangle \geq 0 \quad \forall v \in G, \tag{3.12}$$

which follows from (3.10) with $\bar{y}(v) = y(v) + \mathcal{I}(v-u) \in \mathcal{A}_1 v$. Let $w_\lambda = u + \lambda(w-u)$ for $\lambda \in (0,1]$. Obviously, $w_\lambda \in G$, and according to (3.12), there exists $\bar{y}(w_\lambda) \in \mathcal{A}_1 w_\lambda$ such that

$$\langle \bar{y}(w_\lambda), w-u\rangle \geq 0.$$

Due to (3.9), the set $\{\bar{y}(w_\lambda) : \lambda \in (0,1]\}$ is bounded in V'. Hence, if λ tends to 0 in a suitable manner, the corresponding sequence $\{\bar{y}(w_\lambda)\}$ converges weakly in V' to some $\bar{y}$. Because also $w_\lambda \to u$ in V, the maximal monotonicity of $\mathcal{A}_1$ yields $\bar{y} \in \mathcal{A}_1 u$ and

$$0 \leq \lim \langle \bar{y}(w_\lambda), w-u\rangle = \langle \bar{y}, w-u\rangle.$$

From the last inequality and the inequality $\langle y_1(w), v-w\rangle \geq 0$, given with $v = u$, we obtain

$$\langle \bar{y} - y_1(w), u-w\rangle \leq 0,$$

but this contradicts the strong monotonicity of $\mathcal{A}_1$. □

Let us recall that Assumption 2.1 is supposed to be fulfilled and r, r^* are chosen according to (2.4). Set

$$Q_i = K_i \cap S_r, \qquad Q = K \cap S_r, \qquad Q^* = U^* \cap S_{r^*}. \tag{3.13}$$

To investigate the convergence of the MSR-method for problem (1.1) we make the following general assumption.

ASSUMPTION 3.4.

(a) $\sup_{u\in Q} \ \sup_{y\in\mathcal{T}u} \ \|y\|_{V'} \leq \mu(r) < \infty$ *and* $Q \cap \mathrm{int} D(\mathcal{T}) \neq \emptyset$*;*

[1] Condition $G \cap \mathrm{int} D(\mathcal{A}) \neq \emptyset$ corresponds to the case (b) in the theorem mentioned, and (3.11) ensures that $\langle q_1, v-u\rangle \geq 0$ whenever $v \in D(\mathcal{A}_1) \cap G$, $\|v\| > \alpha \equiv \|u\| + \|q(u)\|_{V'}$, and $q_1 \in \mathcal{A}_1(v)$.

(b) *with* $\mathcal{B}$ *and* $\tilde{\chi}$ *from Assumption* 2.1, *for each* i, *the relation*

$$\langle \mathcal{T}_i u - \mathcal{T}_i v, u - v\rangle \geq \langle \mathcal{B}u - \mathcal{B}v, u - v\rangle \quad \forall u, v \in Q_i$$

is valid, and $\tilde{\chi} \leq 2\bar{\chi}^{-1}$;

(c) *for each* $z \in Q$ *there exist points* $v_i(z) \in Q_i$ *and a compact set* $\Lambda(z) \subset \mathcal{T}z$ *such that*

$$\lim_{i\to\infty} \langle y, v_i(z) - z\rangle = 0 \ \forall y \in \Lambda(z), \quad \lim_{i\to\infty} dist_{V'}(\mathcal{T}_i v_i(z), \Lambda(z)) = 0$$

($\Lambda(z)$ may depend on $\{v_i(z)\}$).

For given sequences $\{\varphi_i\}, \{\sigma_i\}$, *such that* $\lim_{i\to\infty}\varphi_i = \lim_{i\to\infty}\sigma_i = 0$, *it holds that*

(d) *for each* $u \in Q^*$ *there exist points* $w_i \in Q_i$, *satisfying*

$$\|u - w_i\| \leq \varphi_i, \quad dist_{V'}(\mathcal{T}_i w_i, \hat{\Lambda}(u)) < \sqrt{\beta_0}\sigma_i, \ i = 1, 2, \ldots,$$

where $\hat{\Lambda}(u) = \{y \in \mathcal{T}u : \langle y, v - u\rangle \geq 0 \ \ \forall v \in Q\}$;

(e) *for each triple* $u \in Q^*, \ y \in \hat{\Lambda}(u)$, *and* $v_i \in Q_i$ *there exists* $w_i \in Q$ *such that*

$$\langle y, w_i - v_i\rangle \leq \varphi_i \|y\|_{V'} \quad (i = 1, 2, \ldots);$$

(f) *all weak limit points of an arbitrary sequence* $\{v_i\}, \ v_i \in Q_i$, *belong to* Q.

Remark 3.5.

(i) According to Assumption 3.4(b), the operator $\mathcal{T}_i + \chi_i\mathcal{M}$ is strongly monotone on Q_i.

(ii) If $V_1 = V = H$, then Assumption 2.1 is fulfilled with $\mathcal{B} = 0$. Therefore, $|||u||| = \|u\| \ \forall u \in V$, and Assumption 3.4(b) follows from the monotonicity of $\mathcal{T}_i$. In this case all results given below hold also true with $\beta_1 = 0, \ \beta_2 = 1$ if, instead of Assumptions 3.4(c), (d), and (e), the following conditions are valid (cf. [21]):

(c′) for each $z \in Q$ there exist points $v_i(z) \in Q_i$ and a compact set $\Lambda(z) \subset \mathcal{T}z$ such that

$$\lim_{i\to\infty} \|v_i(z) - z\| = 0, \quad \lim_{i\to\infty} dist_{V'}(\mathcal{T}_i v_i(z), \Lambda(z)) = 0;$$

(d′) $dist(Q^*, Q_i) \leq \varphi_i, \quad i = 1, 2, \ldots;$

(e′) for each i and $u_i \in Q_i$ there exists $v_i \in Q$, satisfying

$$\|u_i - v_i\| \leq \varphi_i, \quad dist_{V'}(\mathcal{T}_i u_i, \mathcal{T}v_i) < \sigma_i.$$

We underline that estimate (3.19) and relation (3.34) below remain true also. In section 4 we shall use the modified Assumption 3.4 with (d′) and (e′) instead of (d) and (e).

LEMMA 3.6. *Let Assumptions* 3.4(a), (b), (d), (e) *be fulfilled, and let the relations*

$$\frac{1}{4r}\left(\frac{4}{\chi_i}\left(\mu(r)\varphi_i + r\sigma_i\right) - \tilde{\epsilon}_i^2\right) + \beta_3\epsilon_i < 0, \qquad \tilde{\epsilon}_i = \delta_i - \beta_2\epsilon_i > 0 \tag{3.14}$$

hold for $i \leq i_0$, *with* $\beta_3 = \left(2\beta_1\bar{\chi}^{-1} + \beta_2^2\right)^{1/2}$. *Moreover, assume that in the MSR-method the relations* $s(i) < \infty$ *for* $i < i_0$, *and* $|||u^{i_0,s}||| < r^*$, $|||\bar{u}^{i_0,s}||| < r^*$ *for each* s *are valid. Then* $s(i_0) < \infty$ *is true.*

Proof. Fix $u^{**} \in U^* \cap S_{r^*}$. Due to Assumptions 3.4(d), (e), one can choose $v^{i_0} \in Q_{i_0},\ v^{i_0,s} \in Q$ such that

$$\|u^{**} - v^{i_0}\| \le \varphi_{i_0},$$

and, for some $y \in \hat{\Lambda}(u^{**})$,

$$\langle y, v^{i_0,s} - \bar{u}^{i_0,s}\rangle \le \varphi_i \|y\|_{V'}, \qquad \left\|\mathcal{T}_{i_0} v^{i_0} - y\right\|_{V'} \le \sqrt{\beta_0}\,\sigma_{i_0}.$$

Then, regarding Assumption 3.4(a) and (2.3),

$$\begin{aligned}
&\left\langle \mathcal{T}_{i_0} v^{i_0}, v^{i_0} - \bar{u}^{i_0,s}\right\rangle \\
&= \left\langle y, v^{i_0} - \bar{u}^{i_0,s}\right\rangle + \left\langle \mathcal{T}_{i_0} v^{i_0} - y, v^{i_0} - \bar{u}^{i_0,s}\right\rangle \\
&= \left\langle y, u^{**} - v^{i_0,s}\right\rangle + \left\langle y, v^{i_0} - u^{**}\right\rangle \\
&\quad + \left\langle y, v^{i_0,s} - \bar{u}^{i_0,s}\right\rangle + \left\langle \mathcal{T}_{i_0} v^{i_0} - y, v^{i_0} - \bar{u}^{i_0,s}\right\rangle \\
&\le \left\langle y, u^{**} - v^{i_0,s}\right\rangle + 2r\sigma_{i_0} + 2\mu(r)\varphi_{i_0}.
\end{aligned}$$

However, from the definition of $\hat{\Lambda}(u)$ and $v^{i_0,s} \in Q$, one can conclude that

$$\langle y, u^{**} - v^{i_0,s}\rangle \le 0;$$

hence,

$$\left\langle \mathcal{T}_{i_0} v^{i_0}, v^{i_0} - \bar{u}^{i_0,s}\right\rangle \le 2r\sigma_{i_0} + 2\mu(r)\varphi_{i_0}. \tag{3.15}$$

Now, using Assumption 3.4(b) and Lemma 3.1 with $\mathcal{A} = \mathcal{T}_{i_0},\ G = K_{i_0},\ \chi = \chi_{i_0},\ v = v^{i_0}$, and $a^0 = u^{i_0,s-1}$, in view of (2.6) and (3.15) we obtain

$$\begin{aligned}
&|||\bar{u}^{i_0,s} - v^{i_0}|||^2 - |||u^{i_0,s-1} - v^{i_0}|||^2 \\
&\le -\|\mathcal{P}\bar{u}^{i_0,s} - \mathcal{P}u^{i_0,s-1}\|_H^2 + \frac{2}{\chi_{i_0}}\langle \mathcal{T}_{i_0} v^{i_0}, v^{i_0} - \bar{u}^{i_0,s}\rangle \\
&\le -\|\mathcal{P}\bar{u}^{i_0,s} - \mathcal{P}u^{i_0,s-1}\|_H^2 + \frac{4}{\chi_{i_0}}\left(r\sigma_{i_0} + \mu(r)\varphi_{i_0}\right).
\end{aligned} \tag{3.16}$$

With regard to (2.5) and $\|\mathcal{P}u\|_H \le \beta_2\|u\|\ \ \forall u \in V$, the second inequality in (3.14) ensures

$$\begin{aligned}
&\|\mathcal{P}\bar{u}^{i_0,s} - \mathcal{P}u^{i_0,s-1}\|_H \ge \|\mathcal{P}u^{i_0,s} - \mathcal{P}u^{i_0,s-1}\|_H - \|\mathcal{P}u^{i_0,s} - \mathcal{P}\bar{u}^{i_0,s}\|_H \\
&\ge \|\mathcal{P}u^{i_0,s} - \mathcal{P}u^{i_0,s-1}\|_H - \beta_2\|u^{i_0,s} - \bar{u}^{i_0,s}\| > \delta_{i_0} - \beta_2\epsilon_{i_0} > 0
\end{aligned}$$

for $1 \le s < s(i_0)$. Together with (3.16) this yields

$$\begin{aligned}
&|||\bar{u}^{i_0,s} - v^{i_0}|||^2 - |||u^{i_0,s-1} - v^{i_0}|||^2 \\
&< -\tilde{\epsilon}_{i_0}^2 + \frac{4}{\chi_{i_0}}\left(r\sigma_{i_0} + \mu(r)\varphi_{i_0}\right),
\end{aligned} \tag{3.17}$$

and taking into account the first inequality in (3.14), we have

$$|||\bar{u}^{i_0,s} - v^{i_0}||| < |||u^{i_0,s-1} - v^{i_0}|||.$$

Also, the straightforward application of (2.3), (2.5), and $\tilde{\chi} \le 2\bar{\chi}^{-1}$ gives

$$|||\bar{u}^{i_0,s} - u^{i_0,s}||| \le \beta_3\epsilon_{i_0}.$$

Moreover, from the last inequality, (3.14), and (3.17), due to $|||u^{i_0,s-1}||| < r^*$, $|||\bar{u}^{i_0,s}||| < r^*$, $|||v^{i_0}||| \le r$, $r^* \le r$, we obtain immediately that

$$\begin{aligned} &|||u^{i_0,s} - v^{i_0}||| - |||u^{i_0,s-1} - v^{i_0}||| \\ &< \frac{1}{4r}\left[\frac{4}{\chi_{i_0}}\left(r\sigma_{i_0} + \mu(r)\varphi_{i_0}\right) - \tilde{\epsilon}_{i_0}^2\right] + \beta_3\epsilon_{i_0} < 0 \end{aligned} \tag{3.18}$$

is valid for $1 \le s < s(i_0)$.

Summing up the inequalities (3.18) for $s = 1, \dots, \bar{s}$, with $\bar{s} < s(i_0)$ arbitrarily chosen, one gets

$$|||u^{i_0,\bar{s}} - v^{i_0}||| < |||u^{i_0,0} - v^{i_0}||| + \bar{s}\left[\frac{1}{4r}\left(\frac{4}{\chi_{i_0}}\left(r\sigma_{i_0} + \mu(r)\varphi_{i_0}\right) - \tilde{\epsilon}_{i_0}^2\right) + \beta_3\epsilon_{i_0}\right];$$

therefore,

$$\bar{s} < |||u^{i_0,0} - v^{i_0}||| \left[\frac{1}{4r}\left(\tilde{\epsilon}_{i_0}^2 - \frac{4}{\chi_{i_0}}\left(r\sigma_{i_0} + \mu(r)\varphi_{i_0}\right)\right) - \beta_3\epsilon_{i_0}\right]^{-1},$$

and

$$s(i_0) < |||u^{i_0,0} - v^{i_0}||| \left[\frac{1}{4r}\left(\tilde{\epsilon}_{i_0}^2 - \frac{4}{\chi_{i_0}}\left(r\sigma_{i_0} + \mu(r)\varphi_{i_0}\right)\right) - \beta_3\epsilon_{i_0}\right]^{-1} + 1 \tag{3.19}$$

proves the lemma. □

COROLLARY 3.7. *Let Assumptions* 3.4(a), (b), (d), (e), *and (for each i) the relations* (3.14) *be satisfied. Moreover, assume that the inequalities* $|||\bar{u}^{i,s}||| < r^*$, $|||u^{i,s}||| < r^*$ *are fulfilled step by step. Then the relation* $s(i) < \infty$ *holds for each* i, *i.e., the method suggested is well defined.*

LEMMA 3.8. *Let Assumptions* 3.4(a), (b), (d), (e) *be fulfilled, and let the controlling parameters* ϵ_i, χ_i, δ_i, σ_i, *and* φ_i *satisfy*

$$\frac{1}{4r}\left(\frac{4}{\chi_i}\left(\mu(r)\varphi_i + r\sigma_i\right) - (\delta_i - \beta_2\epsilon_i)^2\right) + \beta_3\epsilon_i < 0 \tag{3.20}$$

and

$$\sum_{i=1}^{\infty}\left(2\left(\frac{\mu(r)\varphi_i + r\sigma_i}{\chi_i}\right)^{1/2} + \beta_3(\epsilon_i + 2\varphi_i)\right) < \frac{1}{2}r^*, \tag{3.21}$$

with $\beta_2 : \ \|u\|_H \le \beta_2\|u\| \ \forall u \in V$ *and* β_3 *defined in Lemma* 3.6. *Then,* (3.14) *holds and in the MSR-method* $s(i) < \infty$ *is valid for each* i*; variational inequality* (2.6) *is uniquely solvable, and* $|||u^{i,s}||| < r^*$, $|||\bar{u}^{i,s}||| < r^*$ *are true for all* (i, s).

Proof. The inequality $\delta_i - \beta_2\epsilon_i > 0$ follows immediately from (3.20), (3.21).

Indeed, (3.21) yields $\beta_3\epsilon_i < \frac{1}{2}r^*$, and due to (3.20), $\beta_3\epsilon_i < \frac{1}{4r}(\delta_i - \beta_2\epsilon_i)^2$. If $\delta_i \le \beta_2\epsilon_i$, then $(\delta_i - \beta_2\epsilon_i)^2 \le (\beta_2\epsilon_i)^2$; hence,

$$\beta_3\epsilon_i < \frac{1}{4r}(\beta_2\epsilon_i)^2.$$

In view of $\beta_3 = \left(2\beta_1\bar{\chi}^{-1} + \beta_2^2\right)^{1/2} \ge \beta_2$, we have $\beta_2\epsilon_i \le \beta_3\epsilon_i < \frac{1}{2}r^*$; therefore

$$\beta_3\epsilon_i < \frac{1}{4r}(\beta_2\epsilon_i)^2 < \frac{r^*}{8r}\beta_2\epsilon_i.$$

However, this contradicts $\beta_3 \ge \beta_2, r^* < r$.

Suppose that i_0 and s_0 are kept fixed, $s(i) < \infty$ for $i < i_0$, and that $0 \leq s_0 < s(i_0)$. Denote

$$\Theta_0 = \{(i, s) \ : \ i < i_0,\ 0 < s \leq s(i) \text{ and } \ i = i_0,\ 0 < s \leq s_0\}.$$

Assume, moreover, that for all $(i, s) \in \Theta_0$ the variational inequality (2.6) is uniquely solvable[2] and the inequalities $|||u^{i,s}||| < r^*$, $|||\bar{u}^{i,s}||| < r^*$ are true.

For fixed $u^{**} \in S_{r^*/8} \cap U^*$ and each i, due to Assumption 3.4(d), one can choose $v^i \in Q_i$ and $y^i \in \hat{\Lambda}(u^{**})$ such that

$$\|v^i - u^{**}\| \leq \varphi_i, \quad \|\mathcal{T}_i v^i - y^i\|_{V'} \leq \sqrt{\beta_0}\sigma_i. \tag{3.22}$$

Then, as in the proof of Lemma 3.6, we obtain (cf. (3.18)) for $i < i_0,\ 0 < s < s(i)$, and $i = i_0,\ 0 < s \leq s_0$ that

$$|||u^{i,s} - v^i||| - |||u^{i,s-1} - v^i||| \leq \frac{1}{4r}\left(\tau_i - \tilde{\epsilon}_i^2\right) + \beta_3\epsilon_i < 0, \tag{3.23}$$

with $\tau_i = \frac{4}{\chi_i}\left(\mu(r)\varphi_i + r\sigma_i\right),\ \tilde{\epsilon}_i = \delta_i - \beta_2\epsilon_i$.

Also for $i < i_0,\ s = s(i)$, similarly to (3.16), one can conclude that

$$|||\bar{u}^{i,s(i)} - v^i|||^2 - |||u^{i,s(i)-1} - v^i|||^2 \leq \tau_i,$$

and regarding $|||\bar{u}^{i,s(i)} - u^{i,s(i)}||| < \beta_3\epsilon_i$ and the implication $a^2 - b^2 < c^2 \ \Rightarrow\ a - |b| < |c|$, it follows that

$$|||u^{i,s(i)} - v^i||| - |||u^{i,s(i)-1} - v^i||| \leq \sqrt{\tau_i} + \beta_3\epsilon_i. \tag{3.24}$$

Summing up the inequalities (3.23), where $i < i_0$ is fixed and $0 < s < s(i)$, together with (3.24), we get

$$|||u^{i,s(i)} - v^i||| - |||u^{i,0} - v^i||| \leq \sqrt{\tau_i} + \beta_3\epsilon_i.$$

Due to (2.3) and (3.22), this yields

$$|||u^{i,s(i)} - u^{**}||| - |||u^{i,0} - u^{**}||| \leq \sqrt{\tau_i} + \beta_3(\epsilon_i + 2\varphi_i). \tag{3.25}$$

Because the mapping $u \to \mathcal{T}_{i_0}u + \chi_{i_0}\mathcal{M}(u - u^{i_0,s_0})$ is monotone and hemicontinuous on Q_{i_0} and Q_{i_0} is a convex closed and bounded set, by Theorem 5 in [33], Remark 3.5(i), and (3.1) the variational inequality

$$u \in Q_{i_0} \ : \ \langle \mathcal{T}_{i_0}u, v - u\rangle + \chi_{i_0}\left(\mathcal{P}u - \mathcal{P}u^{i_0,s_0}, \mathcal{P}v - \mathcal{P}u\right)_H \geq 0 \ \ \forall v \in Q_{i_0} \tag{3.26}$$

is uniquely solvable. We denote its solution by $\hat{u}^{i_0,s_0+1}$. The use of Lemma 3.1 (as in the proof of estimate (3.16)) leads to

$$|||\hat{u}^{i_0,s_0+1} - v^{i_0}||| \leq |||u^{i_0,s_0} - v^{i_0}||| + \sqrt{\tau_{i_0}}. \tag{3.27}$$

Observing (3.22), (3.23) (taken with $i = i_0,\ 0 < s \leq s_0$), and (3.27), the inequality

$$|||\hat{u}^{i_0,s_0+1} - u^{**}||| \leq |||u^{i_0,0} - u^{**}||| + \sqrt{\tau_{i_0}} + 2\beta_3\varphi_{i_0} \tag{3.28}$$

[2] This can be proved directly. We shall get solvability of (2.6) from the solvability of (3.26) below.

can be obtained similarly to (3.25). Regarding that $u^{i+1,0} := u^{i,s(i)}$ (cf. Step (c) of the MSR-method), the inequalities (3.25) and (3.28) imply

$$\begin{aligned}|||\hat{u}^{i_0,s_0+1} - u^{**}||| &\leq |||u^{1,0} - u^{**}||| \\ &+ \sum_{k=1}^{i_0-1} (\sqrt{\tau_k} + \beta_3(\epsilon_k + 2\varphi_k)) + \sqrt{\tau_{i_0}} + 2\beta_3\varphi_{i_0}.\end{aligned} \tag{3.29}$$

Due to (3.21) and the choice of u^{**} and $u^{1,0}$, estimate (3.29) yields

$$|||\hat{u}^{i_0,s_0+1}||| < r^* - \beta_3\epsilon_{i_0}. \tag{3.30}$$

Thus, we are in the situation of Lemma 3.2 (for $y = \mathcal{T}_{i_0}\hat{u}^{i_0,s_0+1} + \chi_{i_0}\mathcal{M}(\hat{u}^{i_0,s_0+1} - u^{i_0,s_0})$, $G = Q_{i_0}$, $G_1 = S_{r^*}$) and Remark 3.5(i), and hence, the variational inequality (2.6) with $i = i_0$ and $s = s_0 + 1$ is uniquely solvable, $\bar{u}^{i_0,s_0+1} = \hat{u}^{i_0,s_0+1}$, and

$$|||u^{i_0,s_0+1}||| < r^*. \tag{3.31}$$

This enables us to conclude that

$$|||\bar{u}^{i_0,s}||| < r^*, \ \ |||u^{i_0,s}||| < r^* \ \forall s,$$

and Lemma 3.6 ensures that $s(i_0) < \infty$. To complete the induction, note that the unique solvability of (2.6) with $i = 1, s = 1$, as well as the inequalities

$$|||\bar{u}^{1,1}||| < r^*, \ \ |||u^{1,1}||| < r^*$$

and the finiteness of $s(1)$, can be established quite analogously. □

Notice for the future that, using (3.21) and (3.29), one can conclude that $\sup_{i,s} |||\bar{u}^{i,s}||| < r^*$, too.

THEOREM 3.9. *Let $\mathcal{T}$ be a maximal monotone operator, Assumption* 3.4 *be fulfilled, and suppose that the points $u^{i,s}$, generated by the MSR-method, as well as the points $\bar{u}^{i,s}$ which solve the variational inequalities* (2.6)*, satisfy the relations*

$$|||\bar{u}^{i,s}||| < r^* \text{ and } |||u^{i,s}||| < r^* \ \forall i, s.$$

Moreover, assume that the controlling parameters satisfy (3.14) *and*

$$\sum_{i=1}^{\infty} \sqrt{\frac{\sigma_i}{\chi_i}} < \infty, \quad \sum_{i=1}^{\infty} \sqrt{\frac{\varphi_i}{\chi_i}} < \infty, \quad \sum_{i=1}^{\infty} \epsilon_i < \infty. \tag{3.32}$$

Then, $s(i) < \infty$ holds for each i and the sequences $\{u^{i,s}\}$, $\{\bar{u}^{i,s}\}$ converge weakly in V to an element $u^* \in U^*$.

Proof. Corollary 3.7 ensures that $s(i) < \infty$ for each i. With an arbitrary $w \in U^* \cap S_{r^*}$, the inequality

$$|||u^{i+1,0} - w||| - |||u^{i,0} - w||| \leq \sqrt{\tau_i} + \beta_3(\epsilon_i + 2\varphi_i)$$

with the same $\tau_i = \frac{4}{\chi_i}(\mu(r)\varphi_i + r\sigma_i)$, $\beta_3 = (2\beta_1\bar{\chi}^{-1} + \beta_2^2)^{1/2}$ can be established similarly to (3.25). Therefore, condition (3.32) and Lemma 2.2.2 in [32] imply the convergence of the sequence $\{|||u^{i,0} - w|||\}$.

Note that the inequalities (3.23), (3.24) remain true with a "new" $v^i \in Q_i$ which is chosen according to Assumption 3.4(d) such that

$$\|v^i - w\| \le \varphi_i, \quad \inf_{y \in \hat{\Lambda}(w)} \|\mathcal{T}_i v^i - y\|_{V'} < \sqrt{\beta_0}\sigma_i.$$

For $1 \le s < s(i)$ this leads to the relation

$$\begin{aligned} -(\sqrt{\tau_i} + \beta_3(\epsilon_i + 2\varphi_i)) + |||u^{i+1,0} - w||| &\le |||u^{i,s} - w||| \\ &< |||u^{i,0} - w||| + 2\beta_3\varphi_i. \end{aligned} \tag{3.33}$$

Indeed, from (3.23) we get immediately

$$|||u^{i,s} - v^i||| < |||u^{i,0} - v^i||| \text{ if } 1 \le s < s(i),$$

and from (3.24) it follows that

$$-\sqrt{\tau_i} - \beta_3\epsilon_i + |||u^{i+1,0} - v^i||| \le |||u^{i,s} - v^i||| < |||u^{i,0} - v^i|||.$$

The latter inequality together with $|||v^i - w||| \le \beta_3\varphi_i$ proves (3.33).

Hence, the sequences $\{|||u^{i,s} - w|||\}$ and $\{|||\bar{u}^{i,s} - w|||\}$ converge to the same limit as $\{|||u^{i,0} - w|||\}$.

Now, from inequality (3.16), which can be extended to each i and $1 \le s \le s(i)$, taking into account that

$$|||v^i - w||| \le \beta_3\varphi_i, \quad |||\bar{u}^{i,s}||| < r^*, \quad |||u^{i,s-1}||| < r^*, \quad |||v^i||| \le r,$$

one can conclude that

$$\begin{aligned} &\|\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}\|_H^2 \\ &\le -|||\bar{u}^{i,s} - v^i|||^2 + |||u^{i,s-1} - v^i|||^2 + \tau_i \\ &\le 4r\left|-|||\bar{u}^{i,s} - v^i||| + |||u^{i,s-1} - v^i|||\right| + \tau_i \\ &\le 4r\left(\left|-|||\bar{u}^{i,s} - w||| + |||u^{i,s-1} - w|||\right| + 2\beta_3\varphi_i\right) + \tau_i. \end{aligned}$$

Therefore,

$$\lim_{i\to\infty} \max_{1 \le s \le s(i)} \|\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}\|_H = 0. \tag{3.34}$$

Using the definition of $\bar{u}^{i,s}$, we get

$$\langle \mathcal{T}_i\bar{u}^{i,s}, v - \bar{u}^{i,s}\rangle + \chi_i\left(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}v - \mathcal{P}\bar{u}^{i,s}\right)_H \ge 0 \quad \forall v \in Q_i, \tag{3.35}$$

and, in view of the monotonicity of $\mathcal{T}_i$, this yields

$$\langle \mathcal{T}_i v, v - \bar{u}^{i,s}\rangle + \chi_i\left(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}v - \mathcal{P}\bar{u}^{i,s}\right)_H \ge 0 \quad \forall v \in Q_i. \tag{3.36}$$

For an arbitrary $z \in Q$, let us choose $v^i(z) \in Q_i$ and $y^i \in \Lambda(z)$ such that

$$\lim_{i\to\infty} \langle y, v^i(z) - z\rangle = 0 \ \forall y \in \Lambda(z), \quad \lim_{i\to\infty} \|\mathcal{T}_i v^i(z) - y^i\|_{V'} = 0. \tag{3.37}$$

This choice is possible due to Assumption 3.4(c). Then,

$$\begin{aligned} &\langle y^i, z - \bar{u}^{i,s}\rangle + \chi_i\left(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}z - \mathcal{P}\bar{u}^{i,s}\right)_H \\ &= \langle y^i, z - v^i(z)\rangle + \chi_i\left(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}z - \mathcal{P}v^i(z)\right)_H \\ &+ \langle y^i - \mathcal{T}_i v^i(z), v^i(z) - \bar{u}^{i,s}\rangle \\ &+ \left[\langle \mathcal{T}_i v^i(z), v^i(z) - \bar{u}^{i,s}\rangle + \chi_i\left(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}v^i(z) - \mathcal{P}\bar{u}^{i,s}\right)_H\right], \end{aligned} \tag{3.38}$$

and in view of (3.36) the term in the square brackets is nonnegative.

Let $\left\{\bar{u}^{i,s}\right\}_{(i,s)\in\Gamma}$ be a subsequence of $\left\{\bar{u}^{i,s}\right\}$, which converges weakly in V to some u^*. With regard to $\sup_{i,s}|||\bar{u}^{i,s}||| < r^*$ and Assumption 3.4(f) one can easily show that $u^* \in K \cap \mathrm{int} S_r$. Now, we choose from $\left\{\bar{u}^{i,s}\right\}_{(i,s)\in\Gamma}$ an infinite subsequence $\left\{\bar{u}^{i,s_i}\right\}_{i\in I}$ and, without loss of generality, suppose that $\{y^i\}_{i\in I}$ converges in V' to an element y (according to Assumption 3.4(c), $\{y^i\}$ belongs to the compact set $\Lambda(z)$). Maximal monotonicity of $\mathcal{T}$ yields $y \in \mathcal{T}z$. Passing to the limit with $s = s_i,\ i \to \infty,\ i \in I$ in equality (3.38), in view of the relations (3.34), (3.37), and the boundedness of the sequences $\{\chi_i\}, \{\bar{u}^{i,s}\}$, we obtain

$$\langle y, z - u^*\rangle \geq 0.$$

Now, on using Assumption 3.4(a), Lemma 3.3 permits us to conclude that

$$\langle y^*, z - u^*\rangle \geq 0 \text{ for some } y^* \in \mathcal{T}u^* \text{ and all } z \in Q.$$

Since $u^* \in \text{int } S_r$, Lemma 3.2 guarantees that $\langle y^*, z-u^*\rangle \geq 0\ \forall z \in K$, hence $u^* \in U^*$. Now, due to Lemma 1 in [31], both sequences $\left\{\bar{u}^{i,s}\right\}$ and $\left\{u^{i,s}\right\}$ converge weakly in V to u^*. □

Remark 3.10. Theorem 3.9 establishes convergence of the MSR-method if the set K is bounded and $dist_V(K_i, K) \leq \sup_i \varphi_i$. Indeed, if r^* is chosen such that $K \subset S_\tau$ holds with some $\tau < r^* - \beta_3(\sup_i\ \varphi_i + \sup_i\ \epsilon_i)$, we get

$$|||\bar{u}^{i,s}||| < r^*,\ \ |||u^{i,s}||| < r^*\ \forall i, s.$$

Thus, the conditions (3.14) (used for all i) and (3.32) ensure weak convergence of $\left\{\bar{u}^{i,s}\right\}$ and $\left\{u^{i,s}\right\}$ to an element $u^* \in U^*$.

If boundedness of K is not supposed, compiling Lemma 3.8 and Theorem 3.9, the following statement on convergence of the MSR-method can be obtained immediately.

THEOREM 3.11. *Let $\mathcal{T}$ be a maximal monotone operator and let Assumption* 3.4 *as well as the conditions* (3.20), (3.21) *be fulfilled. Then, the MSR-method started with arbitrary $u^0 \in S_{r^*/4}$ has the following properties:*

(i) $s(i) < \infty\ \ \forall i$*;*

(ii) $|||u^{i,s}||| < r^*,\ \ |||\bar{u}^{i,s}||| < r^*$ *for each* (i, s);

(iii) *both sequences $\left\{u^{i,s}\right\}$ and $\left\{\bar{u}^{i,s}\right\}$ converge weakly in V to an element $u^* \in U^*$.*

Remark 3.12. Now we are ready to specify the notion “effective iteration” used in the introduction. The (i, s)th iteration is considered to be effective if $\|\mathcal{P}u^{i,s} - \mathcal{P}u^{i,s-1}\|_H > \delta_i$, where δ_i and the other controlling parameters satisfy the conditions (3.20), (3.21) (or (3.14), (3.32) in case K is bounded). To fulfill condition (3.14), (3.32), the parameters can be chosen a priori, for instance in the following way:

$$\chi_i \equiv \chi \in (0, \bar{\chi}],\quad \varphi_i = \sigma_i = \frac{1}{(a+i)^{2+\alpha}},\quad \epsilon_i = \frac{1}{(a+i)^{1+\alpha}},$$

with α a small positive number and an arbitrary $a \geq 0$. Then, the values δ_i can be easily calculated from (3.14). In case the conditions (3.20), (3.21) are used, the choice of $a \geq 0$ in the ratios above has to ensure (3.21), and then δ_i has to be calculated from (3.20). In both cases, under appropriate φ_i, σ_i, and ϵ_i, it makes sense to take δ_i as small as possible.

Of course, a faster decrease of $\varphi_i,\ \sigma_i,\ \epsilon_i$ (and hence, δ_i) is admissible, too.

Remark 3.13. If δ_i is chosen sufficiently large (for instance, $\delta_i > 2r^*\ \ \forall i$), then from Theorem 3.11 (also from Remark 3.10) it follows that $s(i) = 1$ for each i. Indeed, successively for $i = 1, 2, \ldots$ we obtain (without the use of (3.20)) that $|||u^{i,0}||| < r^*$, $|||u^{i,1}||| < r^*$, and regarding (2.2),

$$\|\mathcal{P}u^{i,1} - \mathcal{P}u^{i,0}\|_H \le |||u^{i,1} - u^{i,0}||| < 2r^*$$

holds true, and thus $\|\mathcal{P}u^{i,1} - \mathcal{P}u^{i,0}\|_H < \delta_i$.

This means that the MSR-method passes over to the usual (one-step) diagonal method. Obviously, in this case condition (3.20) can be omitted in Lemma 3.8 and Theorem 3.11.

Remark 3.14. Now let us return to the stopping rule for the auxiliary problems (2.6). Suppose that $\tilde{\chi} \le 2\bar{\chi}^{-1}$,

$$\langle \mathcal{T}_i u - \mathcal{T}_i v, u - v\rangle \ge \langle \mathcal{B}u - \mathcal{B}v, u - v\rangle \;\forall u, v \in K_i,$$

(a modification of Assumption 3.4(b)) and that Assumption 2.1 is valid also. Moreover, instead of (2.5), let $u^{i,s} \in K_i$ be defined by

$$\begin{aligned} &\langle \mathcal{T}_i u^{i,s}, v - u^{i,s}\rangle \\ &\quad + \chi_i(\mathcal{P}u^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}v - \mathcal{P}u^{i,s})_H \ge -\epsilon_i' \|v - u^{i,s}\| \quad \forall v \in K_i. \end{aligned} \tag{3.39}$$

Then, inserting in this inequality $v = \bar{u}^{i,s}$ and summing up the result with the obvious inequality

$$\langle \mathcal{T}_i \bar{u}^{i,s}, u^{i,s} - \bar{u}^{i,s}\rangle + \chi_i(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}u^{i,s} - \mathcal{P}\bar{u}^{i,s})_H \ge 0,$$

we obtain

$$\langle \mathcal{T}_i \bar{u}^{i,s} - \mathcal{T}_i u^{i,s}, \bar{u}^{i,s} - u^{i,s}\rangle + \chi_i(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s}, \mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s})_H \le \epsilon_i' \|\bar{u}^{i,s} - u^{i,s}\|.$$

With regard to this inequality, the modified Assumption 3.4(b) and formula (2.1) lead to

$$\frac{\beta_0}{2}\|\bar{u}^{i,s} - u^{i,s}\|^2 \le \frac{\epsilon_i'}{\chi_i}\|\bar{u}^{i,s} - u^{i,s}\|;$$

hence $\|\bar{u}^{i,s} - u^{i,s}\| \le \frac{2\epsilon_i'}{\beta_0 \chi_i}$.

Thus, under the mentioned modification of Assumption 3.4(b) the convergence results above remain true if $u^{i,s}$ is defined by criterion (3.39) with $\epsilon_i' \le \frac{1}{2}\beta_0 \epsilon_i \chi_i$.

THEOREM 3.15. *Let the conditions of Theorem* 3.9 *or Theorem* 3.11 *be fulfilled and let*

$$\lim_{i\to\infty} \max_{1\le s\le s(i)} \|\mathcal{P}u^{i,s} - \mathcal{P}u^*\|_H = 0, \tag{3.40}$$

with u^* a weak limit of $\{u^{i,s}\}$. Then both sequences $\{u^{i,s}\}$ and $\{\bar{u}^{i,s}\}$ converge to u^* (strongly) in V.

In particular, if $\dim V_1 < \infty$, condition (3.40) follows from the weak convergence of $\{u^{i,s}\}$, taking into account that the orthoprojector is a continuous operator and in the finite dimensional space weak and strong convergence coincides. Also, if the embedding of V into H is compact, (3.40) is obviously fulfilled.

Proof. Due to (2.5) and the continuous embedding of V into H, relation (3.40) leads to

$$\lim_{i\to\infty} \max_{1\le s\le s(i)} \|\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^*\|_H = 0. \tag{3.41}$$

Assumptions 3.4(d), (e) permit one to choose $v^i \in Q_i, y^i \in \hat{\Lambda}(u^*)$, and $v^{i,s} \in Q$ such that

$$\|u^* - v^i\| \le \varphi_i, \qquad \|\mathcal{T}_i v^i - y^i\|_{V'} \le \sqrt{\beta_0}\sigma_i, \qquad i = 1, 2, \dots, \tag{3.42}$$

and

$$\langle y^i, v^{i,s} - \bar{u}^{i,s}\rangle \le \varphi_i \|y^i\|_{V'} \quad \text{for each } i \text{ and } 1 \le s \le s(i). \tag{3.43}$$

On account of Assumption 3.4(b) and the symmetry of the operator $\mathcal{B}$, one gets

$$\begin{aligned}
&\langle \mathcal{B}\bar{u}^{i,s} - \mathcal{B}u^*, \bar{u}^{i,s} - u^*\rangle \\
&= \langle \mathcal{B}\bar{u}^{i,s} - \mathcal{B}v^i, \bar{u}^{i,s} - v^i + v^i - u^*\rangle + \langle \mathcal{B}v^i - \mathcal{B}u^*, \bar{u}^{i,s} - u^*\rangle \\
&= \langle \mathcal{B}\bar{u}^{i,s} - \mathcal{B}v^i, \bar{u}^{i,s} - v^i\rangle - \langle \mathcal{B}u^* - \mathcal{B}v^i, \bar{u}^{i,s} - u^*\rangle - \langle \mathcal{B}\bar{u}^{i,s} - \mathcal{B}v^i, u^* - v^i\rangle \\
&\le \langle \mathcal{T}_i\bar{u}^{i,s} - \mathcal{T}_i v^i, \bar{u}^{i,s} - v^i\rangle - \langle \mathcal{B}\bar{u}^{i,s} - \mathcal{B}u^*, u^* - v^i\rangle - \langle \mathcal{B}\bar{u}^{i,s} - \mathcal{B}v^i, u^* - v^i\rangle,
\end{aligned}$$

and the use of inequality (3.35) with $v = v^i$ yields

$$\begin{aligned}
&\langle \mathcal{B}\bar{u}^{i,s} - \mathcal{B}u^*, \bar{u}^{i,s} - u^*\rangle \\
&\le \langle \mathcal{T}_i v^i, v^i - \bar{u}^{i,s}\rangle + \chi_i \left(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}v^i - \mathcal{P}\bar{u}^{i,s}\right)_H \\
&\quad + \langle \mathcal{B}u^* + \mathcal{B}v^i - 2\mathcal{B}\bar{u}^{i,s}, u^* - v^i\rangle \\
&= \langle \mathcal{T}_i v^i - y^i, v^i - \bar{u}^{i,s}\rangle + \langle y^i, v^i - u^*\rangle + \langle y^i, v^{i,s} - \bar{u}^{i,s}\rangle + \langle y^i, u^* - v^{i,s}\rangle \\
&\quad + \chi_i \left(\mathcal{P}\bar{u}^{i,s} - \mathcal{P}u^{i,s-1}, \mathcal{P}v^i - \mathcal{P}\bar{u}^{i,s}\right)_H \\
&\quad + \langle \mathcal{B}u^* + \mathcal{B}v^i - 2\mathcal{B}\bar{u}^{i,s}, u^* - v^i\rangle.
\end{aligned} \tag{3.44}$$

Now, in view of $\langle y^i, u^* - v^{i,s}\rangle \le 0$, the boundedness of $\{\chi_i\}$ and $\{y^i\}$ (see the definition of $\hat{\Lambda}(u)$ and Assumption 3.4(a)) and of the monotonicity of $\mathcal{B}$, from (3.34), (3.42), and (3.43) we obtain

$$\lim_{i\to\infty} \max_{1\le s\le s(i)} \langle \mathcal{B}\bar{u}^{i,s} - \mathcal{B}u^*, \bar{u}^{i,s} - u^*\rangle = 0.$$

The last relation together with (3.41) leads to

$$\lim_{i\to\infty} \max_{1\le s\le s(i)} |||\bar{u}^{i,s} - u^*||| = 0,$$

and all we have to do is to use (2.3) and (2.5). □

The proximal point method studied in [34], applied to the variational inequality (1.1), corresponds formally to the MSR-scheme with $H = V = V_1$, $\mathcal{T}_i = \mathcal{T}$, $K_i = K$, $s(i) = 1$. However, because we have supposed above that $\mathcal{T}_i$ belongs to the class of hemicontinuous operators, convergence of the method in [34] does not follow from our Theorem 3.11.

Remark 3.16. Lemma 3.1 remains true without any change in the proofs if we suppose that $\mathcal{A}$ is a multivalued operator and inequality (3.2) is fulfilled for any pair of elements belonging to $\mathcal{A}u$ and $\mathcal{A}v$, respectively.

Lemmas 3.6 and 3.8 and Theorems 3.9, 3.11, and 3.15 remain true with minor (and straightforward) modifications in the proofs if, instead of the hemicontinuity of $\mathcal{T}_i$, we suppose that $\mathcal{T}_i : V \to 2^{V'}$ are maximal monotone operators and the following alterations in Assumption 3.4 are performed:

- replace $\lim_{i\to\infty}\ dist_{V'}\ (\mathcal{T}_i v_i(z), \Lambda(z)) = 0$ by $\lim_{i\to\infty}\ \inf_{\eta\in\mathcal{T}_i v_i(z)}\ dist_{V'}\ (\eta, \Lambda(z)) = 0$ in (c);
- replace $dist_{V'}\ (\mathcal{T}_i w_i, \hat{\Lambda}(u)) < \sqrt{\beta_0}\sigma_i$ by $\inf_{\eta\in\mathcal{T}_i w_i(z)}\ dist_{V'}\ (\eta, \hat{\Lambda}(u)) < \sqrt{\beta_0}\sigma_i$ in (d);
- add any condition which provides the solvability of (3.26).
  Taking into account that in this case the operator $\mathcal{T}_i + \chi_i\mathcal{M}$ is also maximal monotone (cf. [33, Theorem 1]) and Q_i is a bounded set, it suffices to suppose that $K_i \cap \mathrm{int}D(\mathcal{T}_i) \neq \emptyset$ or $\mathcal{T}_i$ is locally bounded at some $u \in K_i$ (Theorem 5 in [33]).

The inequality

$$\langle \mathcal{T}_i u - \mathcal{T}_i v, u - v\rangle \geq \langle \mathcal{B}u - \mathcal{B}v, u - v\rangle$$

in Assumption 3.4(b) means in this case that

$$\langle \eta(u) - \eta(v), u - v\rangle \geq \langle \mathcal{B}u - \mathcal{B}v, u - v\rangle \quad \forall \eta(u) \in \mathcal{T}_i u, \forall \eta(v) \in \mathcal{T}_i v.$$

4. Rate of convergence. In this section we investigate the rate of convergence of the basic variant of the MSR-method that corresponds to $H = V = V_1, \mathcal{B} = 0$, and hence, the norms $|||\cdot|||$ and $\|\cdot\|$ coincide.

Let $\delta \in (0, 2r^*), l > 0$, be fixed and denote $U_\delta = \{u \in Q : dist(u, Q^*) \leq \delta\}$. We need the following additional assumption.

ASSUMPTION 4.1. *There exists a constant $d_0 > 0$ such that, for each $u \in U_\delta$ and each $y \in \mathcal{T}u$, the inequality*

$$\inf_{v\in Q^*}\ \langle y, u - v\rangle \geq d_0\|u - u^*(u)\|^l \tag{4.1}$$

holds with $u^*(u) = \arg\min_{w\in Q^*}\ \|u - w\|$.

Assumption 4.1 supposes implicitly that $r^* = r$ or $U^* = Q^*$ has to be.

LEMMA 4.2. *Let Assumption 4.1 be fulfilled. Then, for each $u \in Q$ and each $y \in \mathcal{T}u$, the inequality*

$$\inf_{v\in Q^*}\ \langle y, u - v\rangle \geq d\|u - u^*(u)\|^l \tag{4.2}$$

is valid with $d = (\frac{\delta}{r+r^*})^l d_0$.

Proof. Consider the nontrivial case $Q\backslash U_\delta \neq \emptyset$ and let $u \in Q\backslash U_\delta,\ v \in Q^*$ be chosen arbitrarily. Define $\lambda = \lambda(u, v) \in (0, 1)$ such that $\tilde{u} = \lambda u + (1-\lambda)v \in \partial U_\delta$ (∂U_δ is the boundary of U_δ). Obviously, $\tilde{u} - v = \lambda(u - v),\ \frac{1-\lambda}{\lambda}(\tilde{u} - v) = u - \tilde{u}$, and regarding the monotonicity of $\mathcal{T}$, we obtain that

$$\frac{1-\lambda}{\lambda}\langle y(u) - y(\tilde{u}), \tilde{u} - v\rangle = \langle y(u) - y(\tilde{u}), u - \tilde{u}\rangle \geq 0$$

for any $y(u) \in \mathcal{T}u,\ y(\tilde{u}) \in \mathcal{T}\tilde{u}$.

Hence, due to Assumption 4.1,

$$\langle y(u), \tilde{u} - v\rangle \geq \langle y(\tilde{u}), \tilde{u} - v\rangle \geq d_0\|\tilde{u} - u^*(\tilde{u})\|^l$$

and

$$\langle y(u), u - v\rangle \geq \frac{d_0}{\lambda}\|\tilde{u} - u^*(\tilde{u})\|^l > d_0\|\tilde{u} - u^*(\tilde{u})\|^l.$$

But $\|\tilde{u} - u^*(\tilde{u})\| = \delta,\ \|u - u^*(u)\| < r + r^*$; therefore

$$\langle y(u), u - v\rangle > \left(\frac{\delta}{r + r^*}\right)^l d_0\|u - u^*(u)\|^l.$$

Because v is an arbitrary point in Q^*, this leads to (4.2). But if $u \in U_\delta$, then (4.2) follows immediately from (4.1) and $\delta < 2r^*$. □

Let

$$\begin{aligned}
\rho_{i,s} &= dist^2\left(u^{i,s}, Q^*\right),\\
\gamma_i &= 4r(\epsilon_i + \varphi_i) + \frac{4}{\chi_i}\left[r\sigma_i + (\mu(r) + 2dr)\,\varphi_i + 2dr\epsilon_i\right],\\
c_i &= \rho_{1,0}\prod_{k=1}^{i}\left(1 + \frac{2d}{\chi_k}\right)^{-s(k)+\frac{1}{2}} \quad \text{for } i \geq 1, \quad c_0 = \rho_{1,0},
\end{aligned}$$

and define by $\bar{s}(i)$ the largest integer not exceeding $\sqrt{c_{i-1}}\varphi_i^{-1} + 2$.

Theorem 4.3. *Let Assumptions* 3.4(a), (d′), (e′) *(see Remark* 3.5(ii)*) be fulfilled. Moreover, suppose that the controlling parameters of the MSR-method satisfy the relations* (3.21)*.*

(i) *If Assumption* 4.1 *with* $l = 2$ *is valid, as well as*

$$\gamma_i \leq c_{i-1}\left(1 + \frac{2d}{\chi_i}\right)^{-\bar{s}(i)+1}\left[\left(1 + \frac{2d}{\chi_i}\right)^{1/2} - 1\right] \tag{4.3}$$

and

$$\begin{aligned}
&-(\delta_i - \epsilon_i)^2 + \frac{4}{\chi_i}\left(r\sigma_i + \mu(r)\varphi_i\right)\\
&+ 4r\epsilon_i + \frac{8d}{\chi_i}r\epsilon_i + \left(4 + \frac{8d}{\chi_i}\right)r\varphi_i < 0,
\end{aligned} \tag{4.4}$$

then, for each i and $0 \leq s < s(i)$, the estimate

$$\begin{aligned}
\rho_{i,s} &\leq c_{i-1}\left(1 + \frac{2d}{\chi_i}\right)^{-s}\\
&= \rho_{1,0}\left(1 + \frac{2d}{\chi_i}\right)^{-s}\prod_{k=1}^{i-1}\left(1 + \frac{2d}{\chi_k}\right)^{-s(k)+\frac{1}{2}}
\end{aligned} \tag{4.5}$$

holds true.

(ii) *If Assumption* 4.1 *(with* $l = 1$*) and relation* (3.20) *are valid, then there exists* i_0 *such that the estimate*

$$\begin{aligned}
&dist(u^{i,s}, U^* \cap S_{r^*})\\
&\leq \left(3 + \frac{4r\overline{\chi}}{d}\right)\epsilon_i + 2\left(1 + \frac{2r\overline{\chi}}{d} + \frac{2\mu(r)}{d}\right)\varphi_i + \frac{4r}{d}\sigma_i
\end{aligned} \tag{4.6}$$

holds for $i \geq i_0$ and $1 \leq s \leq s(i)$.

Proof. Because (4.4) implies (3.20), and also since (3.20), (3.21) imply (3.14), the conclusions of Lemmas 3.6 and 3.8 remain true (see Remark 3.5(ii)). For a fixed pair (i,s) with $1 \le s \le s(i)$, choose $v^{i,s} \in Q, y^{i,s} \in \mathcal{T} v^{i,s}$, and $q^{i,s-1} \in Q_i$ such that

$$\|v^{i,s} - \bar{u}^{i,s}\| \le \varphi_i, \qquad \|\mathcal{T}_i \bar{u}^{i,s} - y^{i,s}\|_{V'} \le \sigma_i$$

and

$$\|q^{i,s-1} - u^*(u^{i,s-1})\| \le \varphi_i.$$

This is possible due to Assumptions 3.4(d′) and (e′). Then we get

$$\begin{aligned}
&\langle \mathcal{T}_i \bar{u}^{i,s}, q^{i,s-1} - \bar{u}^{i,s}\rangle \\
&= \langle \mathcal{T}_i \bar{u}^{i,s} - y^{i,s}, q^{i,s-1} - \bar{u}^{i,s}\rangle + \langle y^{i,s}, q^{i,s-1} - u^*(u^{i,s-1})\rangle \\
&\ + \langle y^{i,s}, u^*(u^{i,s-1}) - v^{i,s}\rangle + \langle y^{i,s}, v^{i,s} - \bar{u}^{i,s}\rangle \\
&\le \langle y^{i,s}, u^*(u^{i,s-1}) - v^{i,s}\rangle + 2(r\sigma_i + \mu(r)\varphi_i).
\end{aligned}$$

However, with regard to Remark 3.5(ii), inequality (3.5) given with $H = V$, $\mathcal{P}$ the identity operator, $\mathcal{B} = 0, \mathcal{A} = \mathcal{T}_i, \chi = \chi_i, a^0 = u^{i,s-1}, a^1 = \bar{u}^{i,s}, v = q^{i,s-1}$ leads to

$$\begin{aligned}
&\|\bar{u}^{i,s} - q^{i,s-1}\|^2 - \|u^{i,s-1} - q^{i,s-1}\|^2 \\
&\le -\|\bar{u}^{i,s} - u^{i,s-1}\|^2 + \frac{2}{\chi_i}\langle \mathcal{T}_i \bar{u}^{i,s}, q^{i,s-1} - \bar{u}^{i,s}\rangle.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\|\bar{u}^{i,s} - q^{i,s-1}\|^2 - \|u^{i,s-1} - q^{i,s-1}\|^2 \\
&\le -\|\bar{u}^{i,s} - u^{i,s-1}\|^2 + \frac{2}{\chi_i}\langle y^{i,s}, u^*(u^{i,s-1}) - v^{i,s}\rangle + \frac{4}{\chi_i}(r\sigma_i + \mu(r)\varphi_i).
\end{aligned}$$

Using Assumption 4.1 and Lemma 4.2, this yields

$$\begin{aligned}
&\|\bar{u}^{i,s} - q^{i,s-1}\|^2 - \|u^{i,s-1} - q^{i,s-1}\|^2 \\
&\le -\|\bar{u}^{i,s} - u^{i,s-1}\|^2 - \frac{2d}{\chi_i}\|v^{i,s} - u^*(v^{i,s})\|^l + \frac{4}{\chi_i}(r\sigma_i + \mu(r)\varphi_i).
\end{aligned} \tag{4.7}$$

Now, we prove statement (i) and use the relations

$$\begin{aligned}
&\|\bar{u}^{i,s} - q^{i,s-1}\|^2 - \|u^{i,s-1} - q^{i,s-1}\|^2 \\
&= \big(u^{i,s} - u^*(u^{i,s-1}) - u^{i,s-1} + u^*(u^{i,s-1}) + \bar{u}^{i,s} - u^{i,s}, \\
&\ \ u^{i,s} + u^{i,s-1} - 2u^*(u^{i,s-1}) + \bar{u}^{i,s} - u^{i,s} + 2u^*(u^{i,s-1}) - 2q^{i,s-1}\big) \\
&\ge \|u^{i,s} - u^*(u^{i,s-1})\|^2 - \|u^{i,s-1} - u^*(u^{i,s-1})\|^2 - 4r(\epsilon_i + \varphi_i),
\end{aligned}$$

$$\|v^{i,s} - u^*(v^{i,s})\|^2 \ge \|u^{i,s} - u^*(v^{i,s})\|^2 - 4r(\varphi_i + \epsilon_i),$$

and

$$\|\bar{u}^{i,s} - u^{i,s-1}\|^2 \ge d_{i,s},$$

with

$$d_{i,s} = \begin{cases} (\delta_i - \epsilon_i)^2 & \text{if } 1 \le s < s(i), \\ 0 & \text{if } s = s(i). \end{cases}$$

Inserting these relations in (4.7), which is now valid with $l = 2$, one gets

$$\|u^{i,s} - u^*(u^{i,s-1})\|^2 + \frac{2d}{\chi_i}\|u^{i,s} - u^*(v^{i,s})\|^2 - \|u^{i,s-1} - u^*(u^{i,s-1})\|^2 \\ \le -d_{i,s} + \gamma_i. \tag{4.8}$$

By the use of the definition of $u^*(u)$ and $\rho_{i,s}$, inequality (4.8) leads to

$$\left(1 + \frac{2d}{\chi_i}\right)\rho_{i,s} - \rho_{i,s-1} \le -d_{i,s} + \gamma_i. \tag{4.9}$$

Obviously, condition (4.4) is stronger than the first condition (3.14) given with $\beta_2 = \beta_3 = 1$. Thus, the estimate

$$s(i) < \|u^{i,0} - v^i\| \left[\frac{1}{4r}\left((\delta_i - \epsilon_i)^2 - \frac{4}{\chi_i}(r\sigma_i + \mu(r)\varphi_i)\right) - \epsilon_i\right]^{-1} + 1 \tag{4.10}$$

remains true with some v^i satisfying $\|u^*(u^{i,0}) - v^i\| \le \varphi_i$ (see Remark 3.5(ii) and (3.19)).

Suppose now that for a given pair (i, s) with $0 \le s < s(i)$ the estimate

$$\rho_{i,s} \le c_{i-1}\left(1 + \frac{2d}{\chi_i}\right)^{-s} \tag{4.11}$$

is valid. Then (4.4), (4.10), and $\|u^*(u^{i,0}) - v^i\| \le \varphi_i$ ensure that

$$s(i) \le \bar{s}(i). \tag{4.12}$$

If $s < s(i) - 1$, then from (4.4), we obtain $-d_{i,s+1} + \gamma_i < 0$, and from (4.9) it follows that

$$\rho_{i,s+1} \le c_{i-1}\left(1 + \frac{2d}{\chi_i}\right)^{-s-1}. \tag{4.13}$$

But, if $s = s(i) - 1$, then $d_{i,s+1} = 0$, and (4.9) leads to

$$\left(1 + \frac{2d}{\chi_i}\right)\rho_{i,s(i)} - \rho_{i,s(i)-1} \le \gamma_i. \tag{4.14}$$

Combining (4.3), (4.12), (4.11), and (4.14), one can conclude that

$$\begin{aligned}\rho_{i+1,0} = \rho_{i,s(i)} \le & \left(1 + \frac{2d}{\chi_i}\right)^{-1}\left[c_{i-1}\left(1 + \frac{2d}{\chi_i}\right)^{-s(i)+1}\right. \\ & \left. + c_{i-1}\left(1 + \frac{2d}{\chi_i}\right)^{-s(i)+1}\left(\left(1 + \frac{2d}{\chi_i}\right)^{1/2} - 1\right)\right] \\ & = c_{i-1}\left(1 + \frac{2d}{\chi_i}\right)^{-s(i)+1/2} = c_i.\end{aligned}$$

To prove statement (ii), we start with the estimate

$$\|u^{i,s} - u^*(u^{i,s})\| \le \|u^{i,s} - u^*(v^{i,s})\| \le \|v^{i,s} - u^*(v^{i,s})\| + (\epsilon_i + \varphi_i). \tag{4.15}$$

Inserting (4.15) in inequality (4.7) (with $l = 1$), one gets

$$\begin{aligned}
&\frac{2d}{\chi_i}\|u^{i,s} - u^*(u^{i,s})\| \leq -\|\bar{u}^{i,s} - u^{i,s-1}\|^2 + \|u^{i,s-1} - q^{i,s-1}\|^2 \\
&\quad - \|\bar{u}^{i,s} - q^{i,s-1}\|^2 + \frac{4}{\chi_i}(r\sigma_i + \mu(r)\varphi_i) + \frac{2d}{\chi_i}(\epsilon_i + \varphi_i) \\
&\leq -\|\bar{u}^{i,s} - u^{i,s-1}\|^2 + \frac{4}{\chi_i}(r\sigma_i + \mu(r)\varphi_i) \ + \frac{2d}{\chi_i}(\epsilon_i + \varphi_i) \\
&\quad + \big(u^{i,s-1} - u^*(u^{i,s-1}) - u^{i,s} + u^*(u^{i,s-1}) + u^{i,s} - \bar{u}^{i,s}, \\
&\quad u^{i,s-1} - u^*(u^{i,s-1}) + u^{i,s} - u^*(u^{i,s-1}) \\
&\quad - u^{i,s} + \bar{u}^{i,s} + 2u^*(u^{i,s-1}) - 2q^{i,s-1}\big) \\
&\leq -\|\bar{u}^{i,s} - u^{i,s-1}\|^2 + \|u^{i,s-1} - u^*(u^{i,s-1})\|^2 - \|u^{i,s} - u^*(u^{i,s-1})\|^2 + \theta_i \\
&\leq -\|\bar{u}^{i,s} - u^{i,s-1}\|^2 + \|u^{i,s-1} - u^*(u^{i,s})\|^2 \\
&\quad - \|u^{i,s} - u^*(u^{i,s})\|^2 + \theta_i,
\end{aligned} \tag{4.16}$$

where $\theta_i = (4r + \frac{4\mu(r)+2d}{\chi_i})\varphi_i + (4r + \frac{2d}{\chi_i})\epsilon_i + \frac{4r}{\chi_i}\sigma_i$.

If $\|u^{i,s} - u^{i,s-1}\| \leq \epsilon_i$, then inequality (4.16) yields

$$\begin{aligned}
&\frac{2d}{\chi_i}\|u^{i,s} - u^*(u^{i,s})\| \\
&\leq \|u^{i,s-1} - u^{i,s}\| \left(\|u^{i,s-1} - u^*(u^{i,s})\| + \|u^{i,s} - u^*(u^{i,s})\|\right) + \theta_i \\
&\leq \|u^{i,s-1} - u^{i,s}\| \left(2\|u^{i,s} - u^*(u^{i,s})\| + \epsilon_i\right) + \theta_i.
\end{aligned} \tag{4.17}$$

But if $\|u^{i,s} - u^{i,s-1}\| > \epsilon_i$, then

$$\|\bar{u}^{i,s} - u^{i,s-1}\|^2 \geq \|u^{i,s} - u^{i,s-1}\|^2 + \epsilon_i^2 - 2\epsilon_i\|u^{i,s} - u^{i,s-1}\|,$$

and (4.16) leads to

$$\begin{aligned}
&\frac{2d}{\chi_i}\|u^{i,s} - u^*(u^{i,s})\| \\
&< -\|u^{i,s} - u^{i,s-1}\|^2 + 2\epsilon_i\|u^{i,s} - u^{i,s-1}\| + \|u^{i,s-1} - u^*(u^{i,s})\|^2 \\
&\quad - \|u^{i,s} - u^*(u^{i,s})\|^2 + \theta_i \\
&= 2\left(u^{i,s-1} - u^{i,s}, u^{i,s} - u^*(u^{i,s})\right) + 2\epsilon_i\|u^{i,s} - u^{i,s-1}\| + \theta_i \\
&\leq 2\|u^{i,s-1} - u^{i,s}\| \left(\|u^{i,s} - u^*(u^{i,s})\| + \epsilon_i\right) + \theta_i.
\end{aligned} \tag{4.18}$$

Due to the relation (2.5) and (3.34) (see also Remark 3.5(ii)), there exists i_0 such that

$$\|u^{i,s-1} - u^{i,s}\| \leq \frac{d}{2\bar{\chi}} \quad \text{for } i \geq i_0, 1 \leq s \leq s(i).$$

Together with (4.17), (4.18), and $\chi_i \leq \bar{\chi}$, this gives

$$\frac{d}{\chi_i}\|u^{i,s} - u^*(u^{i,s})\| \leq \frac{d\epsilon_i}{\bar{\chi}} + \theta_i,$$

proving relation (4.6). □

Remark 4.4. If problem (1.1) is uniquely solvable, estimate (4.5) shows that $\{u^{i,s}\}$ converges to u^* not slower than a geometrical progression with the factor $(1+\frac{2d}{\bar{\chi}})^{-1/4}$.

In the general case, analogous estimates can be proved for the distance of $u^{i,s}$ to U^*. If, moreover, Assumption 3.4(c$'$) is valid, then using the proof of Theorem 14.6 in [16], linear convergence of $\{u^{i,s}\}$ to an element $u^* \in U^*$ can be established, too.

Estimate (4.6), in particular, leads to the known result on “finite convergence” of the exact proximal point method (cf. [34], [27]).

It should be emphasized that statement (i) in Theorem 4.3 is mainly qualitative, because a very fast decrease of the parameters φ_i, σ_i, and ϵ_i could be necessary. The choice of the parameters according to statement (ii) can be performed in the same manner as described in Remark 3.12.

In [27], sections 2–3, and [34], section 3, the rate of convergence of the classical proximal point method applied to the problem

$$\text{find } \ u \in V: \quad 0 \in \tilde{\mathcal{T}} u \tag{4.19}$$

with $\tilde{\mathcal{T}} : V \to 2^V$ a maximal monotone operator, has been investigated. In these papers a data approximation is not included. If problem (1.1) is considered with $V' = V$, then it can be rewritten in the form (4.19) with

$$\tilde{\mathcal{T}} u = \begin{cases} \mathcal{T} u + N_K(u) & \text{if } u \in K, \\ \emptyset & \text{if } u \notin K, \end{cases} \tag{4.20}$$

where $N_K(u)$ is the normal cone to K at the point u. The operator $\tilde{\mathcal{T}}$ in (4.20) is maximal monotone, for instance, if $K \cap \text{int} D(\mathcal{T}) \neq \emptyset$ [33, Theorem 1].

To prove linear convergence, in [34] it is supposed that

(a) problem (4.19) is uniquely solvable and, for some $a > 0, \theta > 0$,

$$\|u - \bar{u}\| \leq a\|w\| \text{ whenever } u \in \tilde{\mathcal{T}}^{-1} w \text{ and } \|w\| \leq \theta$$

($\bar{u}$ is the solution of (4.19)).

In [27] this assumption is generalized to the case that problem (4.19) may have more than one solution. Here, denoting by $\bar{U}$ the solution set of (4.19), the corresponding assumption is

(b) for some $a > 0, \theta > 0$,

$$dist(u, \bar{U}) \leq a\|w\| \text{ whenever } u \in \tilde{\mathcal{T}}^{-1} w \text{ and } \|w\| \leq \theta.$$

The “finite convergence” of the exact method is established in [34] under the condition

(c) $0 \in \text{int} \tilde{\mathcal{T}} \bar{u}$,

which is generalized in [27] as follows:

(d) for some $\theta > 0$, the inclusion $u \in \bar{U}$ holds if $u \in \tilde{\mathcal{T}}^{-1} w$ and $\|w\| \leq \theta$.

The correlation between these conditions and Assumption 4.1 is not completely clear. There exist simple examples where Assumption 4.1 with $l = 1$ (resp., $l = 2$) is fulfilled, whereas conditions (a), (b) (resp., (c), (d)) are disturbed. Moreover, one can prove the following relations for the case $r^* = r$:

- If $\mathcal{T}$ is a potential operator, then (a) $\Rightarrow$ Assumption 4.1 ($l = 2$);
- (c) $\Rightarrow$ Assumption 4.1 ($l = 1$).

However, if $r^* = r$ and $K \subset S_r$, then Assumption 4.1 ($l = 2$) $\Rightarrow$ (b) and Assumption 4.1 ($l = 1$) $\Rightarrow$ (d).

Note that Assumption 4.1 as well as the conditions (b), (d) do not prevent from unboundedness or/and infinite-dimensionality of the solution set.

Acknowledgment. The authors are grateful to an anonymous referee for valuable and helpful comments and suggestions.

REFERENCES

[1] P. Alart and B. Lemaire, *Penalization in non-classical convex programming via variational convergence*, Math. Programming, 51 (1991), pp. 307–331.

[2] A.S. Antipin, *On a method for convex programs using a symmetrical modification of the Lagrange function*, Ekonomika i Mat. Metody, 12 (1976), pp. 1164–1173 (in Russian).

[3] A. Auslender, *Numerical methods for non-differentiable convex optimization*, Math. Progr. Study, 30 (1987), pp. 102–126.

[4] A. Auslender, J.P. Crouzeix, and P. Fedit, *Penalty-proximal methods in convex programming*, J. Optim. Theory Appl., 55 (1987), pp. 1–21.

[5] D.P. Bertsekas and P. Tseng, *Partial proximal minimization algorithms for convex programming*, SIAM J. Optim., 4 (1994), pp. 551–572.

[6] G. Chen and M. Teboulle, *A proximal-based decomposition method for convex minimization problems*, Math. Programming, 64 (1994), pp. 81–101.

[7] J. Eckstein, *Approximate iterations in Bregman-function-based proximal algorithms*, Math. Programming, 83 (1998), pp. 113–123.

[8] J. Eckstein and D.P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Programming, 55 (1992), pp. 293–318.

[9] M. Fukushima, *The primal Douglas-Rachford splitting algorithm for a class of monotone mappings with applications to the traffic equilibrium problem*, Math. Programming, 72 (1996), pp. 1–15.

[10] R. Glowinski, J.-L. Lions, and R. Tremolieres, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[11] Cu D. Ha, *A generalization of the proximal point algorithm*, SIAM J. Control Optim., 28 (1990), pp. 503–512.

[12] R. Hettich, A. Kaplan, and R. Tichatschke, *Regularized penalty methods for ill-posed optimal control problems with elliptic equations.* I. *Distributed control with bounded control set and state constraints*, Control Cybernet., 26 (1997), pp. 5–27.

[13] R. Hettich, A. Kaplan, and R. Tichatschke, *Regularized penalty methods for ill-posed optimal control problems with elliptic equations.* II. *Distributed and boundary control with unbounded control set and state constraints*, Control Cybernet., 26 (1997), pp. 29–43.

[14] S. Ibaraki, M. Fukushima, and T. Ibaraki, *Primal-dual proximal point algorithm for linearly constrained convex programming problems*, Comput. Optim. Appl., 1 (1992), pp. 207–226.

[15] A.A. Kaplan, *Algorithm for convex programming using smoothing of exact penalty functions*, Sibirsk. Mat. Zh., 23 (1982), pp. 53–64 (in Russian).

[16] A. Kaplan and R. Tichatschke, *Stable Methods for Ill-Posed Variational Problems. Prox-Regularization of Elliptic Variational Inequalities and Semi-Infinite Problems*, Akademie Verlag, Berlin, 1994.

[17] A. Kaplan and R. Tichatschke, *Path-following proximal approach for solving ill-posed convex semi-infinite programming problems*, J. Optim. Theory Appl., 90 (1996), pp. 113–137.

[18] A. Kaplan and R. Tichatschke, *Proximal Point Methods in Examples*, Forschungsbericht 96-20, Mathematik/Informatik, University of Trier, Germany, 1996.

[19] A. Kaplan and R. Tichatschke, *Prox-regularization and solution of ill-posed variational inequalities*, Appl. Math., 42 (1997), pp. 111–145.

[20] A. Kaplan and R. Tichatschke, *Regularized penalty method for non-coercive parabolic optimal control problems*, Control Cybernet., 27 (1998), pp. 5–27.

[21] A. Kaplan and R. Tichatschke, *Multi-step proximal method for variational inequalities with monotone operators*, in Recent Advances in Optimization, P. Gritzmann, R. Horst, E. Sachs, and R. Tichatschke, eds., Lecture Notes in Econom. Math. Systems 452, Springer, Berlin, 1997, pp. 138–153.

[22] K.C. Kiwiel, *Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities*, Math. Programming, 69 (1995), pp. 89–109.

[23] B. Lemaire, *Coupling optimization methods and variational convergence*, in Trends in Mathematical Optimization, Internat. Schriftenreihe Numer. Math 84, Birkhäuser, Basel, Boston, 1988, pp. 163–179.

[24] B. Lemaire, *The proximal algorithm*, in New Methods in Optimization and Their Industrial Use, Internat. Schriftenreihe Numer. Math 87, Birkhäuser, Basel, Boston, 1989, pp. 73–87.

[25] C. Lemaréchal and C.A. Sagastizábal, *An approach to variable metric bundle methods*, in Lecture Notes in Control and Inform. Sci. 197, Springer, Berlin, 1994, pp. 144–162.

[26] J.-L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[27] F.J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM J. Control Optim., 22 (1984), pp. 277–293.

[28] B. Martinet, *Régularisation d'inéquations variationelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.

[29] R. Mifflin, *A quasi-second-order proximal bundle algorithm*, Math. Programming, 73 (1996), pp. 51–72.

[30] U. Mosco, *Convergence of convex sets and of solutions of variational inequalities*, Adv. Math., 3 (1969), pp. 510–585.

[31] Z. Opial, *Weak convergence of the successive approximaions for nonexpansive mappings in Banach spaces*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.

[32] B.T. Polyak, *Introduction to Optimization*, Optimization Software, Inc., Publications Division, New York, 1987.

[33] R.T. Rockafellar, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.

[34] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[35] R.T. Rockafellar, *Augmented Lagrange multiplier functions and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.

[36] S. Rotin, *Konvergenz des Proximal-Punkt-Verfahrens für inkorrekt gestellte Optimalsteuerprobleme mit partiellen Differentialgleichungen*, Ph.D. thesis, University of Trier, Germany, 1999.

[37] H. Schmitt, *Normschwächere Prox-Regularisierungen*, Ph.D. thesis, University of Trier, Germany, 1996.

[38] M.V. Solodov and B.F. Svaiter, *A hybrid projection-proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.

[39] J.E. Spingarn, *Partial inverse of a monotone operator*, Appl. Math. Optim., 10 (1983), pp. 247–265.

[40] J.E. Spingarn, *Application of the method of partial inverses to convex programming: Decomposition*, Math. Programming, 32 (1985), pp. 199–223.

[41] F.P. Vasil'ev, *Methods for Solving Extremal Problems*, Nauka, Moscow, 1981 (in Russian).

[42] S.J. Wright, *Implementing proximal point methods for linear programming*, J. Optim. Theory Appl., 65 (1990), pp. 531–554.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1160–1181

STABILIZATION OF BERNOULLI–EULER BEAMS BY MEANS OF A POINTWISE FEEDBACK FORCE*

KAIS AMMARI[†] AND MARIUS TUCSNAK[‡]

Abstract. We study the energy decay of a Bernoulli–Euler beam which is subject to a pointwise feedback force. We show that both uniform and nonuniform energy decay may occur. The uniform or nonuniform decay depends on the boundary conditions. In the case of nonuniform decay in the energy space we give explicit polynomial decay estimates valid for regular initial data. Our method consists of deducing the decay estimates from observability inequalities for the associated undamped problem via sharp trace regularity results.

Key words. pointwise stabilization, observability inequality, unbounded feedback, exponential stability

AMS subject classifications. 35E15, 93D15, 93D20, 35B37, 35Q72

PII. S0363012998349315

1. Introduction.

The aim of this paper is to study the pointwise feedback stabilization of a Bernoulli–Euler beam. More precisely, we consider the following initial and boundary value problems:

$$\frac{\partial^2 u}{\partial t^2}(x,t) + \frac{\partial^4 u}{\partial x^4}(x,t) + \frac{\partial u}{\partial t}(\xi,t)\,\delta_\xi = 0, \quad 0 < x < \pi,\ t > 0, \tag{1.1}$$

$$u(0,t) = u(\pi,t) = \frac{\partial^2 u}{\partial x^2}(0,t) = \frac{\partial^2 u}{\partial x^2}(\pi,t) = 0, \quad t > 0, \tag{1.2}$$

$$u(x,0) = u^0(x), \frac{\partial u}{\partial t}(x,0) = u^1(x), \quad 0 < x < \pi, \tag{1.3}$$

and

$$\frac{\partial^2 u}{\partial t^2}(x,t) + \frac{\partial^4 u}{\partial x^4}(x,t) + \frac{\partial u}{\partial t}(\xi,t)\,\delta_\xi = 0, \quad 0 < x < \pi,\ t > 0, \tag{1.4}$$

$$u(0,t) = \frac{\partial u}{\partial x}(\pi,t) = \frac{\partial^2 u}{\partial x^2}(0,t) = \frac{\partial^3 u}{\partial x^3}(\pi,t) = 0, \quad t > 0, \tag{1.5}$$

$$u(x,0) = u^0(x), \frac{\partial u}{\partial t}(x,0) = u^1(x), \quad 0 < x < \pi. \tag{1.6}$$

Here u denotes the transverse displacement of the beam, δ_ξ is the Dirac mass concentrated in the point $\xi \in (0,\pi)$, and we suppose that the length of the beam is equal to π. The boundary condition (1.2) means that both ends of the beam are simply supported, whereas (1.5) signifies that the end $x = 0$ is simply supported and at $x = \pi$ there is a shear hinge end. Simple calculation shows that (1.1) is equivalent to the

*Received by the editors December 11, 1998; accepted for publication (in revised form) March 9, 2000; published electronically December 5, 2000.
http://www.siam.org/journals/sicon/39-4/34931.html

[†] Centre de Mathématiques Appliquées, UMR CNRS 7641, École Polytechnique, 91128 Palaiseau Cedex, France (kais@cmapx.polytechnique.fr).

[‡] Institut Elie Cartan, Université de Nancy I, 54506 Vandoeuvre les Nancy Cedex, France (tucsnak@iecn.u-nancy.fr).

equations modelling the vibrations of two Bernoulli–Euler beams with a dissipative joint (see [4] for further discussion of the model).

Pointwise stabilization of Bernoulli–Euler beams, or, equivalently, stabilization of serially connected beams with dissipative joints, has been widely studied in recent literature (see [4], [5], [6], [7], [17], [23]). In [4], [5], [6], [20], and [23] the authors give several examples showing that both uniform and nonuniform decay may occur. Their method is based on a classical result of Huang and Prüss (see [14] and [22]) combined with elaborate eigenvalues, calculations, or with concepts in system theory. In the cases when we have strong, but not exponential decay, as far as we know, no estimates were given in the literature.

In the present paper we give a simple proof of the fact that for any $\xi \in (0, \pi)$ solutions of (1.1)–(1.3) are not uniformly stable in the energy space. For the solutions of (1.4)–(1.6) we give a complete characterization of points ξ for which the solutions are uniformly stable in the energy space. The main novelty of the paper consists in the fact that, even in the cases when we have no uniform energy decay, we give explicit decay estimates for regular initial data. These estimates depend on the diophantine approximations properties of ξ. As far as we know, the results in previous literature concerning beam equations with pointwise feedbacks are essentially devoted to exponential stabilization. In the case of strong, but not exponential stability, no estimates were given. In the case of bounded feedback controls, similar estimates were given by Russell in [24]. Russell's method cannot be extended to unbounded feedbacks and, namely, to the case of pointwise stabilizers.

Even for the particular case of exponential decay our method is different from those previously used in pointwise stabilization problems. Our approach, avoiding frequency domain methods and spectrum calculations, is based on sharp trace regularity results combined with observability inequalities valid for solutions of appropriate conservative problems. As far as we know this is the first example in which observability estimates for the *undamped problem* are used to derive stability estimates in the presence of an *unbounded feedback*. Due to the appropriate choice of the associated undamped problem the basic observability estimates are simply obtained by applying Ingham's inequality. For bounded feedbacks a similar method was used in [13] in order to study uniform stabilization of second-order equations. Since in our case the feedback is unbounded we have to use new arguments, namely, some sharp trace regularity results.

The plan of the paper is as follows. In section 2 we give precise statements of the main results. Section 3 contains some new trace regularity results needed in the following sections. In section 4 we prove exact pointwise observability results for the associated undamped problem. The proof of the main result is given in section 5.

2. Statement of the main results.

If u is a solution of (1.1)–(1.3) or of (1.4)–(1.6), we define the energy of u at instant t by

$$E(u(t)) = \frac{1}{2}\int_0^\pi \left(\left|\frac{\partial u}{\partial t}(x,t)\right|^2 + \left|\frac{\partial^2 u}{\partial x^2}(x,t)\right|^2 \right) dx. \tag{2.1}$$

Simple formal calculations show that a sufficiently smooth solution of (1.1)–(1.3) or of (1.4)–(1.6) satisfies the energy estimate

$$E(u(0)) - E(u(t)) = \int_0^t \left|\frac{\partial u}{\partial t}(\xi, s)\right|^2 ds \qquad \forall\, t \ge 0. \tag{2.2}$$

In particular, (2.2) implies that

$$E(u(t)) \leq E(u(0)) \quad \forall t \geq 0.$$

The estimate above suggests that the natural wellposedness spaces for (1.1)–(1.3) (respectively, for (1.4)–(1.6)) are $V_1 \times L^2(0,\pi)$ (respectively, $V_2 \times L^2(0,\pi)$), where

$$V_1 = H^2(0,\pi) \cap H_0^1(0,\pi),\ V_2 = \left\{\phi \in H^2(0,\pi) | \phi(0) = \frac{d\phi}{dx}(\pi) = 0\right\}$$

are Hilbert spaces for the inner product

$$\left\langle \begin{pmatrix} u_1 \\ v_1 \end{pmatrix}, \begin{pmatrix} u_2 \\ v_2 \end{pmatrix} \right\rangle_{V_i \times L^2(0,\pi)} = \int_0^\pi \left[\frac{d^2 u_1}{dx^2}\frac{d^2 \bar{u}_2}{dx^2} + v_1 \bar{v}_2\right],\ i = 1, 2.$$

Denote

$$Y = \left[H^2(0,\pi) \cap H^4(0,\xi) \cap H^4(\xi,\pi)\right] \times H^2(0,\pi), \tag{2.3}$$

$$\begin{aligned}\mathcal{D}(A_1) =& \Big\{(u,v) \in Y,\ \ u(0) = v(0) = u(\pi) = v(\pi) = \tfrac{d^2u}{dx^2}(0) = \tfrac{d^2u}{dx^2}(\pi) = 0, \\ & \tfrac{d^2u}{dx^2}(\xi+) = \tfrac{d^2u}{dx^2}(\xi-),\ \tfrac{d^3u}{dx^3}(\xi+) - \tfrac{d^3u}{dx^3}(\xi-) = -v(\xi)\Big\},\end{aligned} \tag{2.4}$$

$$\begin{aligned}\mathcal{D}(A_2) =& \Big\{(u,v) \in Y,\ \ u(0) = v(0) = \tfrac{du}{dx}(\pi) = \tfrac{dv}{dx}(\pi) = \tfrac{d^2u}{dx^2}(0) = \tfrac{d^3u}{dx^3}(\pi) = 0, \\ & \tfrac{d^2u}{dx^2}(\xi+) = \tfrac{d^2u}{dx^2}(\xi-),\ \tfrac{d^3u}{dx^3}(\xi+) - \tfrac{d^3u}{dx^3}(\xi-) = -v(\xi)\Big\}.\end{aligned} \quad . \tag{2.5}$$

The corresponding operators A_1 and A_2 will be defined in section 5. If $(u^0, u^1) \in Y$, we denote

$$||(u^0, u^1)||_Y^2 = \|u^0\|_{H^4(0,\xi)}^2 + \|u^0\|_{H^4(\xi,\pi)}^2 + \|u^1\|_{H^2(0,\pi)}^2. \tag{2.6}$$

We first check that (1.1)–(1.3) (respectively, (1.4)–(1.6)) are well posed in the spaces above. Then we study the behavior of $E(u(t))$ when $t \to \infty$. The wellposedness and strong stability properties are summarized in the result below.

Proposition 2.1. *The following assertions hold true.*

1. *Suppose that* $(u^0, u^1) \in \mathcal{D}(A_1)$ *(respectively, that* $(u^0, u^1) \in \mathcal{D}(A_2)$*). Then the problem* (1.1)–(1.3) *(respectively,* (1.4)–(1.6)*) admits a unique solution*

$$\begin{pmatrix} u \\ \frac{\partial u}{\partial t} \end{pmatrix} \in C(0,T;\mathcal{D}(A_1)) \ \left(\textit{respectively, } \begin{pmatrix} u \\ \frac{\partial u}{\partial t} \end{pmatrix} \in C(0,T;\mathcal{D}(A_2))\right).$$

2. *If* $(u^0, u^1) \in V_1 \times L^2(0,\pi)$ *(respectively,* $(u^0, u^1) \in V_2 \times L^2(0,\pi)$*), then the problem* (1.1)–(1.3) *(respectively,* (1.4)–(1.6)*) admits a unique solution*

$$u \in C(0,T;V_1) \cap C^1(0,T;L^2(0,\pi)) \ (\textit{respectively, } u \in C(0,T;V_2)$$

$$\cap C^1(0,T;L^2(0,\pi))),$$

such that $u(\xi,\cdot) \in H^1(0,T)$ *and*

$$\|u(\xi,\cdot)\|_{H^1(0,T)}^2 \leq C(\|u^0\|_{H^2(0,\pi)}^2 + \|u^1\|_{L^2(0,\pi)}^2), \tag{2.7}$$

where the constant $C > 0$ *depends only on* ξ *and* T*. Moreover,* u *satisfies the energy estimate* (2.2).

3. *The estimate* $\lim_{t\to\infty} E(u(t)) = 0$ *holds true for any finite energy solution of* (1.1)–(1.3) *(respectively, of* (1.4)–(1.6)*) if and only if* $\frac{\xi}{\pi} \notin \mathbf{Q}$ *(respectively,* $\frac{\xi}{\pi} \neq \frac{2p}{2q-1}$ $\forall p, q \in \mathbb{N}$*).*

Remark 1. *The result above shows, in particular, that one cannot expect strong stabilization* $\forall \xi \in (0, \pi)$*.*

The main results in this paper concern the precise asymptotic behavior of the solutions of (1.1)–(1.3) and of (1.4)–(1.6). As we will see below, the systems (1.1)–(1.3) and (1.4)–(1.6) are generally not uniformly stable in the natural energy spaces. However, we prove that, in some cases of strong but not exponential stability, the energy decay is uniform for all initial data lying in more regular spaces.

Denote by $\mathbf{Q}$ the set of all rational numbers. Let us also denote by $\mathcal{S}$ the set of all numbers $\rho \in (0, \pi)$ such that $\frac{\rho}{\pi} \notin \mathbf{Q}$ and if $[0, a_1, \ldots, a_n, \ldots]$ is the expansion of $\frac{\rho}{\pi}$ as a continued fraction, then (a_n) is bounded. Let us notice that $\mathcal{S}$ is obviously uncountable and, by classical results on diophantine approximation (cf. [3, p. 120]), its Lebesgue measure is equal to zero. Roughly speaking, the set $\mathcal{S}$ contains the irrationals which are "badly" approximable by rational numbers. In particular, by the Euler–Lagrange theorem (cf. [18, p. 57]) $\mathcal{S}$ contains all $\xi \in (0, \pi)$ such that $\frac{\xi}{\pi}$ is an irrational quadratic number (i.e., satisfying a second degree equation with rational coefficients). According to a classical result (see, for instance, [26] and the references therein), if $\xi \in \mathcal{S}$, then there exists a constant $C_\xi > 0$ such that

$$|\sin(n\xi)| \geq \frac{C_\xi}{n} \qquad \forall\, n \geq 1. \tag{2.8}$$

Our main results can now be stated as follows.

Theorem 2.2.

1. *For any* $\xi \in (0, \pi)$*, the system described by* (1.1)–(1.3) *is not exponentially stable in* $V_1 \times L^2(0, \pi)$*.*
2. $\forall \xi \in \mathcal{S}$ *and* $\forall t \geq 0$ *we have*

$$E(u(t)) \leq \frac{C_\xi}{(t+1)^2} ||(u^0, u^1)||_Y^2 \qquad \forall\, (u^0, u^1) \in \mathcal{D}(A_1), \tag{2.9}$$

where $C_\xi > 0$ *is a constant depending only on* ξ*.*

3. $\forall \epsilon > 0$ *there exists a set* $B_\epsilon \subset [(0, \pi) \setminus \pi\mathbf{Q}]$*, the Lebesgue measure of* B_ϵ *being equal to* π*, such that* $\forall \xi \in B_\epsilon$ *and* $\forall t \geq 0$ *we have*

$$E(u(t)) \leq \frac{C_{\xi,\epsilon}}{(t+1)^{\frac{2}{1+\epsilon}}} ||(u^0, u^1)||_Y^2 \qquad \forall\, (u^0, u^1) \in \mathcal{D}(A_1), \tag{2.10}$$

where $C_{\xi,\epsilon} > 0$ *is a constant depending only on* ξ *and* ϵ*.*

Theorem 2.3.

1. *The system described by* (1.4)–(1.6) *is exponentially stable in* $V_2 \times L^2(0, \pi)$ *if and only if* $\frac{\xi}{\pi}$ *is a rational number with coprime factorization*

$$\frac{\xi}{\pi} = \frac{p}{q}, \quad \textit{where } p \textit{ is odd.} \tag{2.11}$$

2. $\forall \xi \in \mathcal{S}$ *and* $\forall t \geq 0$ *we have*

$$E(u(t)) \leq \frac{C_\xi}{(t+1)^2} ||(u^0, u^1)||_Y^2 \qquad \forall\, (u^0, u^1) \in \mathcal{D}(A_2), \tag{2.12}$$

where $C_\xi > 0$ *is a constant depending only on* ξ*.*

3. $\forall \epsilon > 0$ *there exists a set* $B_\epsilon \subset [(0,\pi) \setminus \pi\mathbf{Q}]$, *the Lebesgue measure of* B_ϵ *being equal to* π, *such that* $\forall \xi \in B_\epsilon$ *and* $\forall t \geq 0$ *we have*

$$E(u(t)) \leq \frac{C_{\xi,\epsilon}}{(t+1)^{\frac{2}{1+\epsilon}}} ||(u^0, u^1)||_Y^2 \qquad \forall \, (u^0, u^1) \in \mathcal{D}(A_2), \tag{2.13}$$

where $C_{\xi,\epsilon} > 0$ *is a constant depending only on* ξ *and* ϵ.

REMARK 2. *In the case of a string with pointwise stabilizer, the explicit eigenvalue calculation in* [27] *suggests that one cannot expect polynomial decay estimates like* (2.10) *for any* ξ *satisfying the assumption in the third assertion of Proposition* 2.1. *By analogy with the result in* [27] *we conjecture that* $\forall \epsilon > 0$ *there exists* ξ *satisfying the assumption in the third assertion of Proposition* 2.1 *and the sequences* (t_n) *(of real numbers), and* (u_n) *(of finite energy solutions), with* $t_n \to \infty$, *such that*

$$\lim_{n\to\infty} t_n^\epsilon \frac{E(u_n(t_n))}{\|(u_n(0), \frac{\partial u_n}{\partial t}(0))\|_Y^2} = \infty.$$

3. Some regularity results. Consider the initial and boundary value problems

$$\frac{\partial^2 v}{\partial t^2}(x,t) + \frac{\partial^4 v}{\partial x^4}(x,t) \; = k(t)\delta_\xi, \quad 0 < x < \pi, \; t > 0, \tag{3.1}$$

$$v(x,0) = 0, \frac{\partial v}{\partial t}(x,0) = 0, \quad 0 < x < \pi, \tag{3.2}$$

and either

$$v(0,t) = v(\pi,t) = \frac{\partial^2 v}{\partial x^2}(0,t) = \frac{\partial^2 v}{\partial x^2}(\pi,t) = 0, \quad t > 0, \tag{3.3}$$

or

$$v(0,t) = \frac{\partial v}{\partial x}(\pi,t) = \frac{\partial^2 v}{\partial x^2}(0,t) = \frac{\partial^3 v}{\partial x^3}(\pi,t) = 0, \quad t > 0. \tag{3.4}$$

The equations above are models for the vibrations of an undamped Bernoulli–Euler beam, in the presence of a pointwise force. The main result of this section gives regularity properties of the solutions of (3.1)–(3.3) and of (3.1),(3.2), and (3.4). These regularity results are sharp (according to Remark 4 below).

PROPOSITION 3.1. *Suppose that* $k \in L^2(0,T)$. *Then the problem* (3.1)–(3.3) *(respectively,* (3.1),(3.2),(3.4)*) admits a unique solution having the regularity*

$$v \in C(0,T;V_1) \cap C^1(0,T;L^2(0,\pi)) \quad (\textit{respectively}, \tag{3.5}$$

$$v \in C(0,T;V_2) \cap C^1(0,T;L^2(0,\pi))). \tag{3.6}$$

Moreover, $v(\xi,\cdot) \in H^1(0,T)$ *and there exists a constant* $C > 0$, *depending only on* T, *such that*

$$\|v(\xi,\cdot)\|_{H^1(0,T)} \leq C\|k\|_{L^2(0,T)} \qquad \forall \, k \in L^2(0,T). \tag{3.7}$$

REMARK 3. *We notice that the interior regularity* (3.5) *(respectively,* (3.6)*) does not follow from the Sobolev regularity of the right-hand side of* (3.1) *or from the results* *in* [25]*. Moreover, estimate* (3.7) *is not a consequence of the interior regularity* (3.5) *(respectively,* (3.6)*).*

In order to prove Proposition 3.1 we first study the case of free vibrations of an undamped beam, i.e., we consider the initial and boundary value problem

$$\frac{\partial^2\phi}{\partial t^2}(x,t)+\frac{\partial^4\phi}{\partial x^4}(x,t)=0, \quad 0<x<\pi,\ t>0, \tag{3.8}$$

$$\phi(0,t)=\phi(\pi,t)=\frac{\partial^2\phi}{\partial x^2}(0,t)=\frac{\partial^2\phi}{\partial x^2}(\pi,t)=0, \quad t>0, \tag{3.9}$$

$$\phi(x,0)=u^0(x),\frac{\partial\phi}{\partial t}(x,0)=u^1(x), \quad 0<x<\pi, \tag{3.10}$$

and the problem formed by (3.8), (3.10), and the boundary conditions

$$\phi(0,t)=\frac{\partial\phi}{\partial x}(\pi,t)=\frac{\partial^2\phi}{\partial x^2}(0,t)=\frac{\partial^3\phi}{\partial x^3}(\pi,t)=0, \quad t>0. \tag{3.11}$$

The following result, besides showing that the problems above are well posed in the natural energy spaces, gives a sharp inequality on the trace of ϕ at the point ξ.

LEMMA 3.2. *Suppose that* $(u^0,u^1)\in V_1\times L^2(0,\pi)$ *(respectively,* $(u^0,u^1)\in V_2\times L^2(0,\pi)$*). Then the initial and boundary value problem* (3.8)–(3.10) *(respectively,* (3.8),(3.10)*, and* (3.11)*) admits a unique solution*

$$\phi\in C(0,T;V_1)\cap C^1(0,T;L^2(0,\pi)), \tag{3.12}$$

respectively,

$$\phi\in C(0,T;V_2)\cap C^1(0,T;;L^2(0,\pi)), \tag{3.13}$$

satisfying

$$\phi(\xi,\cdot)\in H^1(0,T).$$

Moreover, there exists a constant $C>0$*, depending only on* T*, such that*

$$\|\phi(\xi,\cdot)\|^2_{H^1(0,T)}\le C(\|u^0\|^2_{H^2(0,\pi)}+\|u^1\|^2_{L^2(0,\pi)}). \tag{3.14}$$

Proof. We first notice that problem (3.8)–(3.10) can be written as

$$\frac{\partial}{\partial t}\begin{pmatrix}\phi\\ \dfrac{\partial\phi}{\partial t}\end{pmatrix}=A_0\begin{pmatrix}\phi\\ \dfrac{\partial\phi}{\partial t}\end{pmatrix},$$

where

$$\mathcal{D}(A_0)=\left\{\begin{pmatrix}u\\ v\end{pmatrix}\in\big(H^4(0,\pi)\cap H^1_0(0,\pi)\big)\times\big(H^2(0,\pi)\cap H^1_0(0,\pi)\big)\right.$$

$$\left.\left|\frac{d^2u}{dx^2}(0)=\frac{d^2u}{dx^2}(\pi)=0\right.\right\}, \tag{3.15}$$

and

$$A_0 : \mathcal{D}(A_0) \to V_1 \times L^2(0,\pi),\ A_0 \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ -\dfrac{d^4 u}{dx^4} \end{pmatrix}. \tag{3.16}$$

One can easily check that A_0 is skew-adjoint. So, by Stone's theorem, it generates a semigroup of isometries in $V_1 \times L^2(0,\pi)$. This implies that (3.8)–(3.10) admits a unique solution ϕ satisfying (3.12).

In order to prove (3.14) we put

$$u^0(x) = \sum_{n=1}^{\infty} a_n \sin(nx),\ u^1(x) = \sum_{n=1}^{\infty} n^2 b_n \sin(nx), \tag{3.17}$$

with $\sum_{n=1}^{\infty} n^4(a_n^2 + b_n^2) < \infty$. In this case the solution of (3.8)–(3.10) is given by

$$\phi(x,t) = \sum_{n\geq 1} [a_n \cos(n^2 t)\sin(nx) + b_n \sin(n^2 t)\sin(nx)], \tag{3.18}$$

which implies that

$$\phi(\xi,t) = \sum_{n=1}^{\infty} \left[a_n \cos(n^2 t)\sin(n\xi) + b_n \sin(n^2 t)\sin(n\xi)\right]. \tag{3.19}$$

If we consider the right-hand side of (3.19) as a Fourier series in t (see Theorem 4.1 from [10] for details) we obtain the existence of a constant C depending on T such that

$$\|\phi(\xi,\cdot)\|^2_{H^1(0,T)} \leq C \sum_{n=1}^{\infty} n^4(a_n^2 + b_n^2), \tag{3.20}$$

which obviously implies (3.14).

The problem (3.8), (3.10), (3.11) can be treated in a completely similar manner. It suffices to replace formulas (3.17) and (3.18) by the relations

$$u^0(x) = \sum_{n=0}^{\infty} a_n \sin\left(\frac{2n+1}{2}x\right),\ u^1(x) = \sum_{n=0}^{\infty} \frac{(2n+1)^2}{4} b_n \sin\left(\frac{2n+1}{2}x\right),$$

$$\begin{aligned} \phi(x,t) = \sum_{n=0}^{\infty} \Bigg[& a_n \cos\left(\frac{(2n+1)^2}{4}t\right)\sin\left(\frac{2n+1}{2}x\right) \\ & + b_n \sin\left(\frac{(2n+1)^2}{4}t\right)\sin\left(\frac{2n+1}{2}x\right)\Bigg]. \end{aligned} \tag{3.21}$$

The relations above clearly imply (3.14). □

In order to prove Proposition 3.1 we need the following technical result.

Lemma 3.3. *Let $\gamma > 0$, $\xi \in (0,\pi)$ be two fixed real numbers and $C_\gamma = \{w \in \mathbb{C} \mid Re(w)Im(w) = -\frac{\gamma}{2}\}$. Then the functions*

$$f_1(w) = \frac{i}{2w}\left\{-\frac{\sin(w\xi)\,\sin[w\,(\xi-\pi)]}{\sin(w\pi)} + \frac{sh(w\xi)\,sh[w\,(\xi-\pi)]}{sh(w\pi)}\right\} \tag{3.22}$$

and

$$f_2(w) = \frac{i}{2w}\left\{-\frac{\sin(w\,\xi)\,\cos[w\,(\xi-\pi)]}{\cos(w\pi)} + \frac{sh(w\,\xi)\,ch[w\,(\xi-\pi)]}{ch(w\pi)}\right\} \tag{3.23}$$

are bounded on C_γ, uniformly with respect to $\xi \in [0,\pi]$.

Proof. Let us suppose that f_1 is not bounded on C_γ. In this case there exists a sequence $(w_n) \subset C_\gamma$ such that

$$\lim_{n\to+\infty} |f_1(w_n)| = +\infty. \tag{3.24}$$

As f_1 is analytical in the open set $D = \{w \in \mathbb{C} \mid Re(w)Im(w) < 0\}$ and $C_\gamma \subset D$, relation (3.24) clearly implies that $|w_n| \to +\infty$. Due to the definition of C_γ this can happen in two situations:

$$|Re(w_n)| \to +\infty,\ |Im(w_n)| = \frac{\gamma}{2\,|Re(w_n)|} \to 0, \tag{3.25}$$

or

$$|Im(w_n)| \to +\infty,\ |Re(w_n)| = \frac{\gamma}{2\,|Im(w_n)|} \to 0. \tag{3.26}$$

Suppose that (3.25) holds true. In this case a simple calculation shows that

$$\lim_{n\to+\infty}\left|\frac{sh(w_n\,\xi)\,sh[w_n\,(\xi-\pi)]}{sh(w_n\pi)}\right| = \frac{1}{2}, \tag{3.27}$$

and

$$\limsup_{n\to+\infty} |\sin(w_n\,\xi)\,\sin[w_n\,(\xi-\pi)]| \leq 1. \tag{3.28}$$

Relations (3.24), (3.27), and (3.28) imply that

$$\lim_{n\to+\infty} |w_n\,\sin(w_n\,\pi)| = 0. \tag{3.29}$$

Since $\lim\limits_{n\to+\infty} |w_n| = +\infty$, relation (3.29) yields

$$\lim_{n\to+\infty} |\sin(w_n\,\pi)| = 0.$$

It is easily checked that the relation above implies the existence of a subsequence of (w_n), denoted also by (w_n), and of the sequences $(\alpha_n) \subset \mathbb{N}$, $(\beta_n) \subset [0,1[$, satisfying

$$|Re(w_n)| = \alpha_n + \beta_n,\ \lim_{n\to+\infty} \beta_n = 0. \tag{3.30}$$

We obviously have

$$|w_n \sin(w_n\pi)| = \left|\frac{w_n}{Re(w_n)}\right| \left|\frac{\sin\left(\beta_n\pi - i\pi\frac{\gamma}{2Re(w_n)}\right)}{\beta_n\pi - i\pi\frac{\gamma}{2Re(w_n)}}\right| \left|\beta_n\pi Re(w_n) - i\pi\frac{\gamma}{2}\right|. \tag{3.31}$$

On the other hand, (3.25) and (3.30) imply

$$\lim_{n\to+\infty}\left|\frac{\sin\left(\beta_n\pi - i\pi\frac{\gamma}{2Re(w_n)}\right)}{\beta_n\pi - i\pi\frac{\gamma}{2Re(w_n)}}\right| = 1,\ \lim_{n\to\infty}\left|\frac{w_n}{Re(w_n)}\right| = 1. \tag{3.32}$$

Moreover, we obviously have

$$\left|\beta_n \pi Re(w_n) - i\pi\frac{\gamma}{2}\right| \geq \frac{\pi\gamma}{2} \qquad \forall\, n \geq 1. \tag{3.33}$$

Relations (3.31)–(3.33) contradict (3.29). It follows that (3.24) and (3.25) cannot both be true. By a similar method we can show that (3.24) and (3.26) cannot both hold true. This means that assumption (3.24) is false, i.e., that f_1 is bounded on C_γ. The proof that f_2 is also bounded on C_γ can be done in a completely similar manner. The bounds are uniform with respect to ξ since $\sup_{w\in C_\gamma} |f_i(w)|$, $i = 1, 2$ depends continuously on $\xi \in [0, \pi]$. □

We can now give the proof of the main result of this section.

Proof of Proposition 3.1. We use the method of transposition. Let $\mathcal{D}(A_0)$ be the space defined in (3.15), and denote by $\mathcal{D}[(A_0)]'$ the dual space of $\mathcal{D}(A_0)$ with respect to the pivot space $V_1 \times L^2(0, \pi)$. It is well known that A_0 can be extended to a skew-adjoint operator (denoted also by A_0),

$$A_0 : V_1 \times L^2(0, \pi) \to [\mathcal{D}(A_0)]',$$

such that A_0 generates a group of isometries in $[\mathcal{D}(A_0)]'$, denoted by $S(t)$.

Moreover, we define the operator

$$B_0 : \mathbb{R} \to [\mathcal{D}(A_0)]',\ B_0 r = \begin{pmatrix} 0 \\ r\delta_\xi \end{pmatrix}\ \forall r \in \mathbb{R}. \tag{3.34}$$

With the notation above the problem (3.1)–(3.3) can be written as a Cauchy problem in $[\mathcal{D}(A_0)]'$ under the form

$$\frac{\partial}{\partial t}\begin{pmatrix} v(t) \\ \frac{\partial v}{\partial t}(t) \end{pmatrix} = A_0 \begin{pmatrix} v(t) \\ \frac{\partial v}{\partial t}(t) \end{pmatrix} + B_0 k(t) \quad \forall\, t > 0, \tag{3.35}$$

$$v(0) = \frac{\partial v}{\partial t}(0) = 0. \tag{3.36}$$

After a simple calculation we get that the operator $B_0^* : \mathcal{D}(A_0) \to \mathbb{R}$ is given by

$$B_0^* \begin{pmatrix} u \\ v \end{pmatrix} = v(\xi) \qquad \forall \begin{pmatrix} u \\ v \end{pmatrix} \in \mathcal{D}(A_0).$$

This implies that

$$B_0^* S^*(t) \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} = \frac{\partial \phi}{\partial t}(\xi, t) \qquad \forall \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} \in \mathcal{D}(A_0), \tag{3.37}$$

with ϕ satisfying (3.8)–(3.10). From (3.14) and (3.37) we deduce that there exists a constant $C > 0$ such that

$$\int_0^T \left| B_0^* S^*(t) \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} \right|^2 dt \leq C \left\| \begin{matrix} u^0 \\ u^1 \end{matrix} \right\|^2_{V_1 \times L^2(0,\pi)} \qquad \forall \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} \in \mathcal{D}(A_0). \tag{3.38}$$

According to Theorem 3.1 in [2, p. 173], inequality (3.38) implies that (3.35), (3.36) admit a unique solution

$$\begin{pmatrix} v \\ \frac{\partial v}{\partial t} \end{pmatrix} \in C(0, T; V_1 \times L^2(0, \pi)),$$

which obviously implies the conclusion (3.5). The proof that the interior regularity property (3.6) holds true for all solutions of (3.1), (3.2), and (3.4) can be obtained in a completely similar manner, so we skip it here.

We still have to prove the trace regularity property (3.7).

As (3.1) is time reversible, after extending k by zero for $t \in \mathbb{R} \setminus [0,T]$, we can solve (3.1)–(3.3) for $t \in \mathbb{R}$. In this way we obtain a function, denoted also by v, such that

$$v \in C(0,T;V_1) \cap C^1(0,T;L^2(0,\pi)),\ v(x,t)=0, \qquad \forall\, t \le 0, \tag{3.39}$$

and v satisfies (3.1)–(3.3) $\forall (x,t) \in [0,\pi] \times \mathbb{R}$.

Let $\widehat{v}(x,\lambda)$, where $\lambda = \gamma + i\eta,\ \gamma > 0$, and $\eta \in \mathbb{R}$, be the Laplace (with respect to time) transform of v. Since v satisfies (3.39), estimate (3.7) is equivalent to the fact that the function $t \to e^{-\gamma t} v(\xi,t)$ belongs to $H^1(\mathbb{R})$ and that there exists a constant $M_1 > 0$ such that

$$\|e^{-\gamma \cdot} v(\xi,\cdot)\|^2_{H^1(-\infty,\infty)} \le M_1 \|k\|^2_{L^2(-\infty,\infty)}.$$

Equivalently, by the Parseval identity (see, for instance, [9, p. 212]), it suffices to prove that the function

$$\eta \to (\gamma + i\eta)\widehat{v}(\xi, \gamma + i\eta)$$

belongs to $L^2(\mathbb{R}_\eta)$ for some $\gamma > 0$, and that there exists a constant $M_2 > 0$ such that

$$\|(\gamma + i\eta)\widehat{v}(\xi,\gamma + i\eta)\|^2_{L^2(\mathbb{R}_\eta)} \le M_2 \int_{-\infty}^{\infty} |k(\gamma + i\eta)|^2 d\eta. \tag{3.40}$$

It can be easily checked that $\widehat{v}$ satisfies

$$\lambda^2 \widehat{v}(x,\lambda) + \frac{\partial^4 \widehat{v}}{\partial x^4}(x,\lambda) = 0,\ x \in (0,\xi) \cup (\xi,1),\ Re\lambda > 0, \tag{3.41}$$

$$\widehat{v}(0,\lambda) = \widehat{v}(\pi,\lambda) = \frac{\partial^2 \widehat{v}}{\partial x^2}(0,\lambda) = \frac{\partial^2 \widehat{v}}{\partial x^2}(\pi,\lambda) = 0,\ Re\lambda > 0, \tag{3.42}$$

$$[\widehat{v}]_\xi = \left[\frac{\partial \widehat{v}}{\partial x}\right]_\xi = \left[\frac{\partial^2 \widehat{v}}{\partial x^2}\right]_\xi = 0, \tag{3.43}$$

$$\left[\frac{\partial^3 \widehat{v}}{\partial x^3}\right]_\xi = \widehat{k}(\lambda),\ Re\lambda > 0, \tag{3.44}$$

where we denote by $[f]_\xi$ the jump of the function f at the point ξ. As the equations above are linear, we deduce that, for every $\lambda \in \mathbb{C},\ Re\lambda > 0$, we can find $H_1(\lambda) \in \mathbb{C}$, such that

$$\lambda\, \widehat{v}(\xi,\lambda) = H_1(\lambda)\widehat{k}(\lambda) \quad \forall Re\,\lambda > 0. \tag{3.45}$$

In order to compute $H_1(\lambda)$ we notice that the solutions of (3.41) have the form

$$\widehat{v}(x,\lambda) = \begin{cases} Ae^{iwx} + Be^{-iwx} + Ce^{wx} + De^{-wx}, & x \in (0,\xi), \\ A_1 e^{iw(x-\pi)} + B_1 e^{-iw(x-\pi)} + C_1 e^{w(x-\pi)} + D_1 e^{-w(x-\pi)}, & x \in (\xi,\pi), \end{cases}$$

where $A, B, C, D, A_1, B_1, C_1$, and D_1 are constants and w is the unique complex number satisfying the conditions

$$\lambda = iw^2,\ w = re^{i\theta},\ \text{with } r > 0 \text{ and } \theta \in \left[-\frac{\pi}{2}, 0\right]. \tag{3.46}$$

Using (3.42), we obtain

$$\widehat{v}(x,\lambda) = \begin{cases} 2iA\sin(wx) + 2Csh(wx), & x \in (0,\xi), \\ 2iA_1\sin[w(x-\pi)] + 2C_1sh[w(x-\pi)], & x \in (\xi,\pi). \end{cases}$$

Consequently, the solutions of (3.41)–(3.43) have the following form:

$$\widehat{v}(x,\lambda) = \begin{cases} 2iA\sin(wx) - 2iA\dfrac{\sin(w\pi)sh[w(\xi-\pi)]}{sh(w\pi)\sin[w(\xi-\pi)]}sh(wx), & x \in (0,\xi), \\ 2iA\dfrac{\sin(w\xi)}{\sin[w(\xi-\pi)]}\sin[w(x-\pi)] & \\ -2iA\dfrac{\sin(w\pi)sh(w\xi)}{sh(w\pi)\sin[w(\xi-\pi)]}sh[w(x-\pi)], & x \in (\xi,\pi). \end{cases} \tag{3.47}$$

Then, using (3.44) and (3.45), we obtain

$$H_1(\lambda) = f_1(w), \tag{3.48}$$

where f_1 is defined by (3.22). By (3.46), the relation $Re\,\lambda = \gamma > 0$ implies that $w \in C_\gamma$, with C_γ defined in Lemma 3.3 . We can now apply Lemma 3.3 to obtain the existence of a constant $M_2 > 0$ such that (3.40) holds true. This ends the proof of the fact that (3.7) holds for all solutions of (3.1)–(3.3).

If v is the solution of (3.1), (3.2), and (3.4), similar calculations (see also [23]) imply that

$$\lambda\,\widehat{v}(\xi,\lambda) = H_2(\lambda)\widehat{k}(\lambda) \quad \forall Re\,\lambda > 0,$$

where $H_2(iw^2) = f_2(w)$ and f_2 is defined in (3.23). Again applying Lemma 3.3 and the method above, we can easily conclude that (3.7) holds for all solutions of (3.1), (3.2), and (3.4). □

REMARK 4. *It can be easily checked that* $\forall\varepsilon > 0$, $\lambda^\varepsilon H_i(\lambda), i = 1, 2$, *is not bounded on* C_γ. *This means that estimate* (3.7) *is no longer valid if we replace the* H^1 *norm by the* $H^{1+\varepsilon}$ *norm in the left-hand side of* (3.7). *This means that* (3.7) *is a sharp estimate.*

REMARK 5. *In* [23] *it is shown that the system* (3.1), (3.2), (3.4) *can be written as*

$$\dot{z} + Az = Bk$$

in an appropriate Hilbert space, with the input k *and the output* $y = B^*k$. *The results we proved in this section say that this system is well posed, in the sense used in* [23]. *According to classical results (see again* [23] *and the references therein), this fact is equivalent to the boundedness of* H_2 *on some half plane* $Re\lambda \geq \gamma > 0$. *This boundedness was proved in the appendix of* [23]. *Since we didn't use* H_2 *for the proof of the interior regularity, we a priori needed only the boundedness of* H_2 *on the line* $Re\,\lambda = \gamma > 0$. *Due to this fact, our approach can be easily adapted for other systems such as strings or Kirchhoff beams.*

4. Some observability inequalities. In this section we gather, for easy reference, some observability inequalities concerning the trace at the point $x = \xi$ of the solutions of (3.8)–(3.10) and of (3.8), (3.9), (3.11). The results in this section are similar to those obtained in [26] in a slightly different situation. Our first result concerns problem (3.8)–(3.10), and it can be stated as follows.

PROPOSITION 4.1. *Let $T > 0$ be fixed and $\mathcal{S} \subset [0, \pi]$ be the set introduced in section* 2. *Then we have the following.*

1. *$\forall \xi \in \mathcal{S}$ the solution ϕ of* (3.8)–(3.10) *satisfies*

$$\int_0^T \left|\frac{\partial \phi}{\partial t}(\xi, t)\right|^2 dt \geq C_\xi \left(\|u^0\|^2_{H^1(0,\pi)} + \|u^1\|^2_{H^{-1}(0,\pi)}\right)$$

$$\forall (u^0, u^1) \in V_1 \times L^2(0, \pi), \tag{4.1}$$

where $C_\xi > 0$ is a constant depending only on ξ.

2. *$\forall \epsilon > 0$ and for almost all $\xi \in (0, \pi)$ the solution ϕ of* (3.8)–(3.10) *satisfies*

$$\int_0^T \left|\frac{\partial \phi}{\partial t}(\xi, t)\right|^2 dt \geq C_{\xi,\epsilon} \left(\|u^0\|^2_{H^{1-\epsilon}(0,\pi)} + \|u^1\|^2_{H^{-1-\epsilon}(0,\pi)}\right)$$

$$\forall (u^0, u^1) \in V_1 \times L^2(0, \pi), \tag{4.2}$$

where $C_{\xi,\epsilon} > 0$ is a constant depending only on ξ and ϵ.

3. *The result in assertion* 1 *is sharp in the sense that, $\forall \xi \in (0, \pi)$, there exists a sequence $(u^0_m, u^1_m) \subset V_1 \times L^2(0, \pi)$ such that the corresponding sequence of solutions (ϕ_m) of* (3.8), (3.9) *with initial data (u^0_m, u^1_m) satisfies $\forall \epsilon > 0$*

$$\lim_{m \to \infty} \frac{\int_0^T \left|\frac{\partial \phi_m}{\partial t}(\xi, t)\right|^2 dt}{\|u^0_m\|^2_{H^{1+\epsilon}(0,\pi)} + \|u^1_m\|^2_{H^{-1+\epsilon}(0,\pi)}} = 0. \tag{4.3}$$

Proof. Notice first that, thanks to Lemma 3.2, the left-hand side of (4.1) is well defined and

$$\frac{\partial \phi}{\partial t}(\xi, t) = \sum_{n=1}^{\infty} \left[-n^2 a_n \sin(n^2 t) \sin(n\xi) + n^2 b_n \cos(n^2 t) \sin(n\xi)\right] \tag{4.4}$$

in $L^2(0, T)$, provided that u^0, u^1 are given by (3.17). Moreover, from (4.4) and the Ball–Slemrod generalization of Ingham's inequality (cf. [1], [11]) we obtain that, $\forall T > 0$, there exists a constant $C_T > 0$ such that

$$\int_0^T \left|\frac{\partial \phi}{\partial t}(\xi, t)\right|^2 dt \geq C_T \sum_{n=1}^{\infty} \left[n^4 a_n^2 \sin^2(n\xi) + n^4 b_n^2 \sin^2(n\xi)\right]. \tag{4.5}$$

Suppose now that ξ belongs to the set $\mathcal{S}$ defined in section 2. Then relations (4.5) and (2.8) imply the existence of a constant $K_{T,\xi} > 0$ such that

$$\int_0^T \left|\frac{\partial \phi}{\partial t}(\xi, t)\right|^2 dt \geq K_{T,\xi} \sum_{n=1}^{\infty} \left[n^2 a_n^2 + n^2 b_n^2\right] \qquad \forall\, \xi \in \mathcal{S},$$

which is exactly (4.1).

In order to prove (4.2) we use a result in [3, p. 120] (see also Proposition 2.4 in [26]) to get that $\forall \epsilon > 0$ there exists a set $B_\epsilon \subset (0,\pi)$ having the Lebesgue measure equal to π and a constant $C > 0$, such that for any $\rho \in B_\epsilon$

$$|\sin(n\rho)| \geq \frac{C}{n^{1+\epsilon}} \qquad \forall\, n \geq 1. \tag{4.6}$$

Let us notice that by Roth's theorem B_ϵ contains all numbers in $(0,\pi)$ having the property that $\frac{\xi}{\pi}$ is an algebraic irrational (see, for instance, [3, p. 104]). Inequalities (4.5) and (4.6) obviously imply (4.2).

We still have to show the existence of a sequence satisfying (4.3). By using continuous fractions (see again [26] and the references therein for details) we can construct a sequence $(q_m) \subset \mathbb{N}$ such that $q_m \to \infty$ and

$$|\sin(q_m\xi)| \leq \frac{\pi}{q_m} \qquad \forall\, m \geq 1. \tag{4.7}$$

Using (4.4) and (4.7), a simple calculation shows that the sequence $(\phi_m^0, \phi_m^1) = (\sin(q_m\pi x), 0)$ satisfies (4.3). □

The observability results for (3.8), (3.9), and (3.11) are given in the proposition below.

PROPOSITION 4.2. *Let $T > 0$ be fixed and $\mathcal{S}$ be the set introduced in section* 2. *Then the following assertions hold true.*

1. *The existence of a constant $C_\xi > 0$, such that the solutions ϕ of* (3.8), (3.10), *and* (3.11) *satisfy*

$$\int_0^T \left|\frac{\partial\phi}{\partial t}(\xi,t)\right|^2 dt \geq C_\xi \left(\|u^0\|^2_{H^2(0,\pi)} + \|u^1\|^2_{L^2(0,\pi)}\right)$$

$$\forall (u^0,u^1) \in V_1 \times L^2(0,\pi), \tag{4.8}$$

is equivalent to the fact that ξ satisfies (2.11).

2. *$\forall \xi \in \mathcal{S}$ the solution ϕ of* (3.8), (3.10), *and* (3.11) *satisfies* (4.1).
3. *$\forall \epsilon > 0$ and for almost all $\xi \in (0,\pi)$ the solution ϕ of* (3.8), (3.10), *and* (3.11) *satisfies* (4.2).

Proof. From (3.21) and the Ball–Slemrod generalization of Ingham's inequality we obtain the existence of a constant $C_T > 0$ such that the solution ϕ of (3.8), (3.10), and (3.11) satisfies

$$\int_0^T \left|\frac{\partial\phi}{\partial t}(\xi,t)\right|^2 dt \geq C_T \sum_{n\geq 0} \frac{(2n+1)^4}{16}(a_n^2+b_n^2)\left|\sin\left(\frac{2n+1}{2}\xi\right)\right|^2. \tag{4.9}$$

If ξ satisfies (2.11), then, by Lemma 2.9 in [23], there exists a constant $k_\xi > 0$ such that

$$\left|\sin\left[\frac{(2n+1)\xi}{2}\right]\right| \geq k_\xi \qquad \forall\, n \geq 0. \tag{4.10}$$

Inequalities (4.9) and (4.10) imply that (4.8) holds true $\forall \xi$ satisfying (2.11).

On the other hand, if ξ does not satisfy (2.11), we can again apply Lemma 2.9 from [23] to get the existence of a sequence $(p_m) \subset \mathbb{N}$, $\lim_{m\to\infty} p_m = \infty$ such that

$$\lim_{m\to\infty} \sin\left[\frac{(2p_m+1)\xi}{2}\right] = 0. \tag{4.11}$$

If we denote by ϕ_m the solution of (3.8), (3.11) with initial data

$$\phi_m(x,0) = \sin\left[\frac{(2p_m+1)x}{2}\right], \frac{\partial \phi_m}{\partial t}(x,0) = 0 \qquad \forall\, x \in (0,\pi),$$

a simple calculation using (4.11) implies that

$$\lim_{m\to\infty} \frac{\int_0^T \left|\frac{\partial \phi_m}{\partial t}(\xi,t)\right|^2 dt}{\|\phi_m(0)\|^2_{H^2(0,\pi)} + \|\frac{\partial \phi_m}{\partial t}(0)\|^2_{L^2(0,\pi)}} = 0,$$

so (4.8) is false for any ξ not satisfying (2.11). Assertions 2 and 3 of the proposition can be proved by simply adapting the proof of Proposition 4.1, so we skip the details. □

5. Proof of the main results.

Proof of Proposition 2.1. The existence and uniqueness of finite energy solutions of (1.1)–(1.3) (respectively, the problem (1.4)–(1.6)) can be obtained by standard semigroup methods. However, for the sake of completeness we sketch the proof here.

Consider the unbounded linear operator

$$A_1 : \mathcal{D}(A_1) \to V_1 \times L^2(0,\pi),\ A_1\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ -\frac{d^4u}{dx^4} - v(\xi)\delta_\xi \end{pmatrix},$$

where the derivatives with respect to x are calculated in $\mathcal{D}'(0,\pi)$, and $\mathcal{D}(A_1)$ is defined in (2.4). If $(u,v) \in \mathcal{D}(A_1)$, we denote by h_1 (respectively, by h_2) the function in $L^2(0,\xi)$ (respectively, in $L^2(\xi,\pi)$) defined by

$$h_1(x) = \frac{d^4u}{dx^4}, \text{ calculated in } \mathcal{D}'(0,\xi),$$

$$h_2(x) = \frac{d^4u}{dx^4}, \text{ calculated in } \mathcal{D}'(\xi,\pi).$$

Moreover, we define $\{\frac{d^4u}{dx^4}\} \in L^2(0,\pi)$ by

$$\left\{\frac{d^4u}{dx^4}\right\} = \begin{cases} h_1(x) & \text{if} \quad x \in (0,\xi), \\ h_2(x) & \text{if} \quad x \in (\xi,\pi). \end{cases}$$

A simple calculation shows that

$$A_1\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ -\left\{\frac{d^4u}{dx^4}\right\} \end{pmatrix} \qquad \forall \begin{pmatrix} u \\ v \end{pmatrix} \in \mathcal{D}(A_1). \tag{5.1}$$

We remark that $\mathcal{D}(A_1) \subset Y$ and that the graph norm in $\mathcal{D}(A_1)$ is equivalent to $\|\cdot\|_Y$. A simple calculation gives

$$\left\langle A_1\begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix}\right\rangle_{V_1\times L^2(0,\pi)} = -|v(\xi)|^2 \qquad \forall \begin{pmatrix} u \\ v \end{pmatrix} \in \mathcal{D}(A_1),$$

so A_1 is a dissipative operator. Moreover, it can be easily checked that A_1 is onto, so, according to Theorems 4.3 and 4.6 from [21, p. 14–15], we obtain that A_1 generates a continuous semigroup of linear contractions acting on $V_1 \times L^2(0,\pi)$.

This implies the existence and uniqueness of solutions u of (1.1)–(1.3) satisfying

$$\begin{pmatrix} u \\ \frac{\partial u}{\partial t} \end{pmatrix} \in C(0,T;\mathcal{D}(A_1)), \text{ if } (u^0,u^1) \in \mathcal{D}(A_1),$$

and

$$u \in C(0,T;V_1) \cap C^1(0,T;L^2(0,\pi)), \text{ if } (u^0,u^1) \in V_1 \times L^2(0,\pi).$$

In order to prove estimate (2.2) and the trace regularity property (2.7), it suffices to remark that, through simple integration by parts, they hold true for regular solutions (i.e., $\left(\begin{smallmatrix} u \\ \frac{\partial u}{\partial t} \end{smallmatrix}\right) \in C(0,T;\mathcal{D}(A_1))$. We can then use the density of $\mathcal{D}(A_1)$ in $V_1 \times L^2(0,\pi)$. The similar properties for problem (1.4)–(1.6) can be proved by simply replacing A_1 by the operator

$$A_2 : \mathcal{D}(A_2) \to V_2 \times L^2(0,\pi), \; A_2 \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ -\frac{d^4u}{dx^4} - v(\xi)\delta_\xi \end{pmatrix},$$

where $\mathcal{D}(A_2)$ is defined in (2.5).

The strong stability estimates at the end of Proposition 2.1 can be obtained by a simple application of LaSalle's invariance principle. However, for the sake of completeness we give here the proof. Let $S(t)$ be the semigroup of contractions generated by the operator A_1, already introduced. In order to prove the strong stability of the solutions of (1.1)–(1.3), it clearly suffices to show that

$$\lim_{t\to\infty} S(t) \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} = 0 \qquad \forall \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} \text{ in } \mathcal{D}(A_1).$$

We will show that this holds true provided that

$$\frac{\xi}{\pi} \notin \mathbf{Q}. \tag{5.2}$$

Since the imbedding $\mathcal{D}(A_1) \subset V_1 \times L^2(0,\pi)$ is compact, the set

$$\text{orb} \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} = \cup_{t\geq 0} S(t) \begin{pmatrix} u^0 \\ u^1 \end{pmatrix}$$

is precompact in $V_1 \times L^2(0,\pi)$ for any $\left(\begin{smallmatrix} u^0 \\ u^1 \end{smallmatrix}\right)$ in $\mathcal{D}(A_1)$. In this case the ω-limit set of $\left(\begin{smallmatrix} u^0 \\ u^1 \end{smallmatrix}\right)$ defined by

$$\omega \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} = \left\{ U \in V_1 \times L^2(0,\pi), \exists (t_n), \; t_n \to \infty, S(t_n) \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} \to U, \quad n \to \infty \right\}$$

is nonvoid for any $\left(\begin{smallmatrix} u^0 \\ u^1 \end{smallmatrix}\right)$ in $\mathcal{D}(A_1)$. On the other hand, by LaSalle's invariance principle (we refer to [8], [12, p. 18] for more details),

$$\text{if } \begin{pmatrix} \phi^0 \\ \phi^1 \end{pmatrix} \in \omega \begin{pmatrix} u^0 \\ u^1 \end{pmatrix},$$

then

$$\begin{aligned} \left\| \begin{pmatrix} \phi^0 \\ \phi^1 \end{pmatrix} \right\|_{V_1 \times L^2(0,\pi)} &= \lim_{t_n \to +\infty} \left\| S(t_n) \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} \right\|_{V_1 \times L^2(0,\pi)} \\ &= \lim_{n\to+\infty} \left\| S(t+t_n) \begin{pmatrix} u^0 \\ u^1 \end{pmatrix} \right\|_{V_1 \times L^2(0,\pi)} \\ &= \left\| S(t) \begin{pmatrix} \phi^0 \\ \phi^1 \end{pmatrix} \right\|_{V_1 \times L^2(0,\pi)}. \end{aligned}$$

Thus,

$$\left\| \begin{pmatrix} \phi(.,t) \\ \frac{\partial \phi}{\partial t}(\cdot,t) \end{pmatrix} \right\|_{V_1 \times L^2(0,\pi)} = \left\| \begin{pmatrix} \phi^0 \\ \phi^1 \end{pmatrix} \right\|_{V_1 \times L^2(0,\pi)} \text{ for any } t \geq 0.$$

The relation above and (2.2) imply that ϕ satisfies the system (1.1)–(1.2) together with

$$\frac{\partial \phi}{\partial t}(\xi, t) = 0 \quad \forall\, t \in (0,T). \tag{5.3}$$

In particular, this implies that ϕ is the solution of (3.8)–(3.9) with $\phi(x,0) = \phi^0(x)$, $\frac{\partial \phi}{\partial t}(x,0) = \phi^1(x)$. If we put

$$\phi^0(x) = \sum_{n \geq 1} c_n \sin(nx), \phi^1(x) = \sum_{n \geq 1} n^2 d_n \sin(nx),$$

with $(c_n),\ (d_n) \subset l^2(\mathbb{R})$, we have

$$\frac{\partial \phi}{\partial t}(\xi,t) = \sum_{n \geq 1} n^2 \Big\{ - c_n \sin(n^2 t) + d_n \cos(n^2 t) \Big\} \sin(n\xi).$$

The relation above, (5.2), and Ingham's inequality imply that $c_n = d_n = 0\ \forall n \in \mathbb{N}$ so $\phi^0 \equiv \phi^1 \equiv 0$. We can now conclude that condition (5.2) is sufficient for the strong stability of the solutions of (1.1)–(1.3).

If we suppose that ξ doesn't satisfy (5.2), i.e., that $\frac{\xi}{\pi} = \frac{p}{q}$, with $p, q \in \mathbb{Z}$, one can easily check that the solution u of (1.1)–(1.3) with initial data $u^0 = \sin(qx),\ u^1 = 0$ satisfies $E(u(t)) = E(u(0))\ \forall t \geq 0$, so (5.2) is also necessary for the strong stability of the solutions of (1.1)–(1.3).

In a completely similar manner we can tackle the strong stability for (1.4)–(1.6). □

Let $u \in C(0,T;V_1) \cap C^1(0,T;L^2(0,\pi))$ be the solution of (1.1)–(1.3). Then u can be written as

$$u = \phi + \psi, \tag{5.4}$$

where ϕ is the solution of (3.8)–(3.10) and ψ satisfies

$$\frac{\partial^2 \psi}{\partial t^2} + \frac{\partial^4 \psi}{\partial x^4} + \frac{\partial u}{\partial t}(\xi,t)\, \delta_\xi = 0 \quad \text{in } (0,\pi) \times (0,T), \tag{5.5}$$

$$\psi(0,t) = \psi(\pi,t) = \frac{\partial^2 \psi}{\partial x^2}(0,t) = \frac{\partial^2 \psi}{\partial x^2}(\pi,t) = 0,\ t \in (0,T), \tag{5.6}$$

$$\psi(x,0) = \frac{\partial \psi}{\partial t}(x,0) = 0,\ x \in (0,\pi). \tag{5.7}$$

In the same way the solution of (1.4)–(1.6) can be decomposed as in (5.4), where ϕ is the solution of (3.8), (3.10), (3.11), and ψ satisfies (5.5), (5.7) together with

$$\psi(0,t)=\frac{\partial^2\psi}{\partial x^2}(0,t)=\frac{\partial\psi}{\partial x}(\pi,t)=\frac{\partial^3\psi}{\partial x^3}(\pi,t)=0,\ t\in(0,T). \tag{5.8}$$

The main ingredient of the proofs of Theorems 2.2 and 2.3 is the following result.

LEMMA 5.1. *Suppose that* $(u^0,u^1)\in V_1\times L^2(0,\pi)$ *(respectively,* $(u^0,u^1)\in V_2\times L^2(0,\pi)$*). Then the solutions* u *of* (1.1)–(1.3) *(respectively, of* (1.4)–(1.6)*) and the solution* ϕ *of* (3.8)–(3.10) *(respectively, of* (3.8), (3.10), (3.11)*) satisfy*

$$C_1\int_0^T\left|\frac{\partial\phi}{\partial t}(\xi,t)\right|^2dt\le\int_0^T\left|\frac{\partial u}{\partial t}(\xi,t)\right|^2dt\le 4\int_0^T\left|\frac{\partial\phi}{\partial t}(\xi,t)\right|^2dt, \tag{5.9}$$

where $C_1>0$ *is a constant independent of* (u^0,u^1).

REMARK 6. *By Proposition* 2.1, $\frac{\partial u}{\partial t}(\xi,\cdot)\in L^2(0,T)$. *So,* (5.5) *makes sense. The result above shows that the* L^2 *norm of* $\frac{\partial u}{\partial t}(\xi,\cdot)$ *is equivalent to the* L^2 *norm of* $\frac{\partial\phi}{\partial t}(\xi,\cdot)$. *(Notice that* $\frac{\partial\phi}{\partial t}(\xi,\cdot)\in L^2(0,T)$ *by Lemma* 3.2.)

Proof of Lemma 5.1. We prove (5.9) only for u satisfying (1.1)–(1.3) and the ϕ solution of (3.8)–(3.10). As for u satisfying (1.4)–(1.6) and the ϕ solution of (3.8), (3.10), (3.11), the proof is a completely similar one.

Relation (5.4) implies that

$$\int_0^T\left|\frac{\partial\phi}{\partial t}(\xi,t)\right|^2dt\le 2\left\{\int_0^T\left|\frac{\partial u}{\partial t}(\xi,t)\right|^2dt+\int_0^T\left|\frac{\partial\psi}{\partial t}(\xi,t)\right|^2dt\right\}.$$

The estimate above combined with inequality (3.7) in Proposition 3.1 implies the existence of a constant $C_1>0$, independent of (u^0,u^1), such that

$$C_1\int_0^T\left|\frac{\partial\phi}{\partial t}(\xi,t)\right|^2dt\le\int_0^T\left|\frac{\partial u}{\partial t}(\xi,t)\right|^2dt. \tag{5.10}$$

On the other hand, according to Remark 6 and to relation (5.4), we have that $\frac{\partial\phi}{\partial t}(\xi,\cdot)\in L^2(0,T)$. This means that (5.5) can be rewritten as

$$\frac{\partial^2\psi}{\partial t^2}(x,t)+\frac{\partial^4\psi}{\partial x^4}(x,t)+\frac{\partial\psi}{\partial t}(\xi,t)\,\delta_\xi=-\frac{\partial\phi}{\partial t}(\xi,t)\,\delta_\xi\quad\text{in }(0,\pi)\times(0,T). \tag{5.11}$$

If we formally multiply (5.11) by $\frac{\partial\bar\psi}{\partial t}$ (this can be done rigorously by considering a regularizing sequence), we obtain

$$\int_0^T\left|\frac{\partial\psi}{\partial t}(\xi,t)\right|^2dt\le\left|\int_0^T\frac{\partial\phi}{\partial t}(\xi,t)\frac{\partial\bar\psi}{\partial t}(\xi,t)dt\right|,$$

which obviously yields

$$\left\|\frac{\partial\psi}{\partial t}(\xi,t)\right\|^2_{L^2(0,T)}\le\left\|\frac{\partial\phi}{\partial t}(\xi,t)\right\|^2_{L^2(0,T)}.$$

Relation (5.4) and the inequality above imply that

$$\left\|\frac{\partial u}{\partial t}(\xi,t)\right\|^2_{L^2(0,T)}\le 4\left\|\frac{\partial\phi}{\partial t}(\xi,t)\right\|^2_{L^2(0,T)}. \tag{5.12}$$

Inequalities (5.10) and (5.12) obviously yield the conclusion (5.9). □

Before giving the proof of the main results we need one more technical lemma. This lemma extends a result in [16].

LEMMA 5.2. *Let $(\mathcal{E}_k)$ be a sequence of positive real numbers satisfying*

$$\mathcal{E}_{k+1} \leq \mathcal{E}_k - C\mathcal{E}_{k+1}^{2+\alpha} \quad \forall k \geq 0, \tag{5.13}$$

where $C > 0$ and $\alpha > -1$ are constants. Then there exists a positive constant M (depending on α and C) such that

$$\mathcal{E}_k \leq \frac{M}{(k+1)^{\frac{1}{(1+\alpha)}}} \quad \forall k \geq 0. \tag{5.14}$$

Proof. Consider the sequence

$$\mathcal{F}_k = \frac{M}{(k+1)^{\frac{1}{1+\alpha}}},$$

where $M > 0$ is to be determined. After a simple calculation we obtain that

$$\frac{1}{M} \lim_{k\to\infty} \left[(\mathcal{F}_k - \mathcal{F}_{k+1})k(k+2)^{\frac{1}{1+\alpha}}\right] = \frac{1}{1+\alpha}, \tag{5.15}$$

so there exists $k_0 > 0$ such that

$$\mathcal{F}_k - \mathcal{F}_{k+1} \leq \frac{2M}{(1+\alpha)k(k+2)^{\frac{1}{1+\alpha}}} \quad \forall k \geq k_0.$$

The relation above implies that

$$\mathcal{F}_k - \mathcal{F}_{k+1} \leq \frac{4}{(1+\alpha)M^{1+\alpha}}\mathcal{F}_{k+1}^{2+\alpha} \quad \forall k \geq k_1 = \max\{k_0, 2\}. \tag{5.16}$$

If we suppose now that

$$\frac{4}{(1+\alpha)M^{1+\alpha}} < C \text{ and } \frac{M}{(k_1+1)^{\frac{1}{1+\alpha}}} \geq \mathcal{E}_{k_1}, \tag{5.17}$$

from (5.16) we get

$$\mathcal{F}_k - \mathcal{F}_{k+1} \leq C\mathcal{F}_{k+1}^{2+\alpha} \quad \forall k \geq k_1. \tag{5.18}$$

It obviously suffices to show that

$$\mathcal{E}_k \leq \mathcal{F}_k \quad \forall k \geq k_1. \tag{5.19}$$

We shall do that by induction over k.

For $k = k_1$, (5.19) follows directly from (5.17). If we suppose that (5.19) holds true for $k \leq m$, by combining (5.13) and (5.18) we obtain

$$\mathcal{E}_{m+1} + C\mathcal{E}_{m+1}^{2+\alpha} \leq \mathcal{F}_{m+1} + C\mathcal{F}_{m+1}^{2+\alpha},$$

which obviously implies that $\mathcal{E}_{m+1} \leq \mathcal{F}_{m+1}$. ☐

We can now prove the main results.

Proof of Theorem 2.2. 1. Suppose that there exists $\xi \in (0,\pi)$ such that solutions of (1.1)–(1.3) satisfy the estimate

$$E(u(t)) \leq Me^{-\omega t}E(u(0)) \qquad \forall\, t \geq 0, \tag{5.20}$$

where $M, \omega > 0$ are constants depending only on ξ. Relation (5.20) implies the existence of a time $T > 0$ and of a constant $C > 0$ (depending on T) such that

$$E(u(0)) - E(u(T)) \geq CE(u(0)) \qquad \forall\, (u^0, u^1) \in V_1 \times L^2(0,\pi).$$

The relation above combined with (2.2) yields

$$\int_0^T \left|\frac{\partial u}{\partial t}(\xi, s)\right|^2 ds \geq CE(u(0)) \qquad \forall\, (u^0, u^1) \in V_1 \times L^2(0,\pi),$$

which, by Lemma 5.1, implies that the solution ϕ of (3.8)–(3.10) satisfies

$$\int_0^T \left|\frac{\partial \phi}{\partial t}(\xi, s)\right|^2 ds \geq \frac{C}{4}E(u(0)) \qquad \forall\, (u^0, u^1) \in V_1 \times L^2(0,\pi).$$

The inequality above clearly contradicts assertion 3 in Proposition 4.1. So assumption (5.20) is false. We end in this way the proof of the first assertion of Theorem 2.2.

We pass now to the proof of the second assertion of this theorem. Let $\xi \in \mathcal{S}$. By Proposition 4.1 and Lemma 5.1, the solution u of (1.1)–(1.3) satisfies the inequality

$$\int_0^T \left|\frac{\partial u}{\partial t}(\xi, t)\right|^2 dt \geq K_1 \left(\|u^0\|^2_{H^1(0,\pi)} + \|u^1\|^2_{H^{-1}(0,\pi)}\right) \quad \forall (u^0, u^1) \in V_1 \times L^2(0,\pi),$$

where $K_1 > 0$ is a constant. The relation above and (2.2) imply that

$$\|\{u(T), u'(T)\}\|^2_{V_1 \times L^2(0,\pi)} \leq \|\{u^0, u^1\}\|^2_{V_1 \times L^2(0,\pi)}$$

$$-K_1 \|\{u^0, u^1\}\|^2_{H^1(0,\pi) \times H^{-1}(0,\pi)} \qquad \forall\, (u^0, u^1) \in \mathcal{D}(A_1). \tag{5.21}$$

By using a simple interpolation inequality (cf. [19, p. 49]), the fact that the function $t \rightarrow \|\{u(t), u'(t)\}\|^2_{V_1 \times L^2(0,\pi)}$ is nonincreasing, and relation (5.21), we obtain the existence of a constant $K_2 > 0$ such that

$$\|\{u(T), u'(T)\}\|^2_{V_1 \times L^2(0,\pi)} \leq \|\{u^0, u^1\}\|^2_{V_1 \times L^2(0,\pi)}$$

$$-K_2 \frac{\|\{u(T), u'(T)\}\|^3_{V_1 \times L^2(0,\pi)}}{\|\{u^0, u^1\}\|_Y}. \tag{5.22}$$

We follow now the method used in [24]. Estimate (5.22) remains valid in successive intervals $[kT, (k+1)T]$. So, $\forall k \geq 0$, we have

$$\|\{u((k+1)T), u'((k+1)T)\}\|^2_{V_1 \times L^2(0,\pi)}$$

$$\leq \|\{u(kT), u'(kT)\}\|^2_{V_1 \times L^2(0,\pi)} - K_2 \frac{\|\{u((k+1)T), u'((k+1)T)\}\|^3_{V_1 \times L^2(0,\pi)}}{\|\{u(kT), u'(kT)\}\|_Y}.$$

Since A_1 generates a semigroup of contractions in $\mathcal{D}(A_1)$ and the graph norm on $\mathcal{D}(A_1)$ is equivalent to $\|\cdot\|_Y$, the relation above implies the existence of a constant $K_3 > 0$ such that

$$\|\{u((k+1)T), u'((k+1)T)\}\|^2_{V_1\times L^2(0,\pi)} \leq \|\{u(kT), u'(kT)\}\|^2_{V_1\times L^2(0,\pi)}$$

$$(5.23)\quad -K_3\frac{\|\{u((k+1)T), u'((k+1)T)\}\|^3_{V_1\times L^2(0,\pi)}}{\|\{u^0, u^1\}\|_Y} \qquad \forall\ (u^0, u^1) \in \mathcal{D}(A_1).$$

If we adopt now the notation

$$(5.24)\qquad \mathcal{E}_k = \frac{\|\{u(kT), u'(kT)\}\|^2_{V_1\times L^2(0,\pi)}}{\|\{u^0, u^1\}\|^2_Y},$$

relation (5.23) gives

$$(5.25)\qquad \mathcal{E}_{k+1} \leq \mathcal{E}_k - K_3\mathcal{E}_{k+1}^{\frac{3}{2}} \quad \forall k \geq 0.$$

By applying Lemma 5.2 for $\alpha = -\frac{1}{2}$ and using relation (5.25), we obtain the existence of a constant $M > 0$ such that

$$\|\{u(kT), u'(kT)\}\|^2_{V_1\times L^2(0,\pi)} \leq \frac{M\|\{u^0, u^1\}\|^2_Y}{(k+1)^2} \quad \forall k \geq 0.$$

The conclusion (2.9) follows now by simply using the fact that the function

$$t \to \|\{u(t), u'(t)\}\|^2_{V_1\times L^2(0,\pi)}$$

is nonincreasing.

Let us now suppose that $\epsilon > 0$ and that ξ belongs to the set B_ϵ, introduced in section 4. From (2.2), (4.2), and Lemma 5.1, it follows that

$$\|\{u(T), u'(T)\}\|^2_{V_1\times L^2(0,\pi)} \leq \|\{u^0, u^1\}\|^2_{V_1\times L^2(0,\pi)}$$

$$-C\|\{u^0, u^1\}\|^2_{H^{1-\epsilon}(0,\pi)\times H^{-1-\epsilon}(0,\pi)}.$$

Using now the same method as above and the interpolation theorem from [19, p. 81], we obtain that the sequence $\mathcal{E}_k$, defined by (5.24), satisfies

$$\mathcal{E}_{k+1} \leq \mathcal{E}_k - K\mathcal{E}_{k+1}^{\frac{3+\epsilon}{2}} \quad \forall k \geq 1.$$

The relation above and Lemma 5.2 (with $\alpha = \frac{\epsilon-1}{2}$) give

$$\mathcal{E}_k \leq \frac{M}{(k+1)^{\frac{2}{1+\epsilon}}} \quad \forall k \geq 1,$$

which obviously implies (2.10). □

Proof of Theorem 2.3. As above, we use the fact that all finite energy solutions of (1.4)–(1.6) are exponentially stable in $V_2 \times L^2(0,\pi)$ if and only if there exist the positive constants T and K_T such that

$$(5.26)\qquad E(u(0)) - E(u(T)) \geq K_T E(u(0)) \qquad \forall\ (u^0, u^1) \in V_2 \times L^2(0,\pi).$$

Using now (2.2), (5.26), and Lemma 5.1, we obtain the existence of a constant $C_\xi > 0$ such that all solutions ϕ of (3.8), (3.9), and (3.11) satisfy (4.8). By Proposition 4.2, inequality (4.8) holds true if and only if ξ satisfies (2.11). Consequently, we obtain that the finite energy solutions of (1.4)–(1.6) are exponentially stable in $V_2 \times L^2(0,\pi)$ if and only if ξ satisfies (2.11).

The proof of estimates (2.12), (2.13) can be done by using obvious adaptations of the proof of estimates (2.9), (2.10), so it is omitted. □

REFERENCES

[1] J. M. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization of semilinear control systems*, Comm. Pure Appl. Math., 32 (1979), pp. 555–587.

[2] A. Bensoussan, G. Da Prato, M. Delfour, and S. Mitter, *Representation and Control of Infinite Dimensional Systems*, vol. 1, Birkhäuser Boston, Boston, MA, 1992.

[3] J. W. S. Cassels, *An Introduction to Diophantine Approximation*, Cambridge University Press, Cambridge, UK, 1966.

[4] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[5] G. Chen, S. G. Krantz, D. W. Ma, C. E. Wayne, and H. H. West, *The Euler-Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, Sung J.Lee, ed., Marcel Dekker, New York, 1988, pp. 67–96.

[6] G. Chen, S. G. Krantz, D. L. Russell, C. E. Wayne, H. H. West, and M. P. Coleman, *Analysis, designs and behavior of dissipative joints for coupled beams*, SIAM J. Appl. Math., 49 (1989), pp. 1665–1693.

[7] F. Conrad, *Stabilization of beams by pointwise feedback control*, SIAM J. Control Optim., 28 (1990), pp. 423–437.

[8] C. Dafermos and M. Slemrod, *Asymptotic behavior of solutions of nonlinear contraction semigroups*, J. Funct. Anal., 13 (1973), pp. 97–106.

[9] G. Doetsch, *Introduction to the Theory and Application of the Laplace Transformation*, Springer, Berlin, 1974.

[10] C. Fabre and J. P. Puel, *Pointwise controllabillity as limit of internal controllability for the wave equation in one space dimension*, Portugal. Math., 51 (1994), pp. 335–350.

[11] A. Haraux, *Quelques proprietes des séries lacunaires utiles dans l' étude des systèmes élastiques*, in Nonlinear Partial Differential Equations and Their Applications. Collège de France Seminar, Vol. XII (Paris, 1991–1993), Pitman Res. Notes Math. Ser. 302, Longman Sci. Tech., Harlow, UK, 1994, pp. 113–124.

[12] A. Haraux, *Systèmes dynamiques dissipatifs et applications*, Masson, Paris, 1991.

[13] A. Haraux, *Une remarque sur la stabilisation de certains systemes du deuxieme ordre en temps*, Portugal. Math., 46 (1989), pp. 245–258.

[14] F. Huang, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert space*, Ann. Differential Equations, 1 (1985), pp. 43–56.

[15] A. E. Ingham, *Some trigonometrical inequalities with applications in the theory of series*, Math. Z., 41 (1936), pp. 367–369.

[16] S. Jaffard, M. Tucsnak, and E. Zuazua, *Singular internal stabilization of the wave equation*, J. Differential Equations, 145 (1998), pp. 184–215.

[17] J.E. Lagnese, G. Leugering, and E. Schmidt, *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*, Birkhäuser, Basel, 1994.

[18] S. Lang, *Introduction to Diophantine Approximations*, Addison Wesley, New York, 1966.

[19] J. L. Lions and E. Magenes, *Problèmes aux limites non homogénes et applications*, Dunod, Paris, 1968.

[20] K.-S. Liu, F.-L. Huang, and G. Chen, *Exponential stability analysis of a long chain of coupled vibrating strings with dissipative linkage*, SIAM J. Appl. Math., 49 (1989), pp. 1694–1707.

[21] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[22] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 248 (1984), pp. 847–857.

[23] R. Rebarber, *Exponential stability of coupled beams with dissipative joints: A frequency domain approach*, SIAM J. Control. Optim., 33 (1995), pp. 1–28.

[24] D. L. Russell, *Decay rates for weakly damped systems in Hilbert space obtained with control theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.

[25] R. Triggiani, *Regularity with interior point control.* I. *Wave and Euler-Bernoulli equations*, in Boundary Control and Boundary Variation (Sophia-Antipolis, 1990), Lecture Notes in Control and Inform. Sci. 178, Springer, Berlin, 1992, pp. 321–355.

[26] M. Tucsnak, *Regularity and exact controllability for a beam with piezoelectric actuator*, SIAM J. Control Optim., 34 (1996), pp. 922–930.

[27] M. Tucsnak, *On the pointwise stabilization of a string*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math. 126, W. Desch, F Kappel, and K. Kunisch, eds., Birkhäuser, Basel, 1996, pp. 287–297.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1182–1203

PONTRYAGIN'S PRINCIPLE FOR LOCAL SOLUTIONS OF CONTROL PROBLEMS WITH MIXED CONTROL-STATE CONSTRAINTS*

E. CASAS†, J.-P. RAYMOND‡, AND H. ZIDANI§

Abstract. This paper deals with optimal control problems of semilinear parabolic equations with pointwise state constraints and coupled integral state-control constraints. We obtain necessary optimality conditions in the form of a Pontryagin's minimum principle for local solutions in the sense of L^p, $p \leq +\infty$.

Key words. optimal control, nonlinear boundary controls, semilinear parabolic equations, state constraints, Pontryagin's minimum principle, unbounded controls

AMS subject classifications. 49K20, 35K20

PII. S0363012998345627

1. Introduction. Let T be a positive number, Ω be a bounded open subset in $\mathbb{R}^N$ $(N \geq 2)$ with a Lipschitz boundary Γ, and q, σ, and $\bar{\sigma}$ be numbers satisfying

$$q > N/2 + 1 \quad \text{and} \quad \sigma > \bar{\sigma} > N + 1.$$

Consider the parabolic system

$$\frac{\partial y}{\partial t} + Ay + f(x,t,y) = 0 \text{ in } Q,\ \frac{\partial y}{\partial n_A} + g(s,t,y,v) = 0 \text{ on } \Sigma,\ y(0) = y_0 \text{ in } \Omega \tag{1.1}$$

(where $Q := \Omega\times]0,T[$, $\Sigma := \Gamma\times]0,T[$, $T > 0$, v is a boundary control, $y_0 \in C(\overline{\Omega})$, A is a second order elliptic operator) and the following control and state constraints:

$$v \in \widetilde{V}_{ad} := \{v \in L^\sigma(\Sigma) \mid v(s,t) \in V(s,t) \quad \text{for almost every (a.e.) } (s,t) \in \Sigma\},$$

$$\Phi(y) \in \mathcal{C}, \tag{1.2}$$

$$\begin{aligned} &\int_\Sigma \Psi_i(s,t,y(s,t),v(s,t))\,dsdt = 0, \quad 1 \leq i \leq m_0,\\ &\int_\Sigma \Psi_i(s,t,y(s,t),v(s,t))\,dsdt \leq 0, \quad m_0 + 1 \leq i \leq m. \end{aligned} \tag{1.3}$$

(V is a measurable set-valued mapping from Σ with closed and nonempty values in $\mathcal{P}(\mathbb{R}^k)$, the set of all subsets of $\mathbb{R}^k$, $\Psi = (\Psi_1, \ldots, \Psi_m)$, is a function with values in

*Received by the editors October 5, 1998; accepted for publication (in revised form) June 14, 2000; published electronically December 5, 2000.
http://www.siam.org/journals/sicon/39-4/34562.html

†Departamento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, 39005 Santander, Spain (casas@etsiso.macc.unican.es). The research of this author was partially supported by Dirección General de Enseñanza Superior e Investigación Científica (Madrid).

‡Université Paul Sabatier, UMR CNRS MIP, UFR MIG, 31062 Toulouse Cedex 4, France (raymond@mip.ups-tlse.fr).

§UMR CNRS 6628, Université d' Orléans, BP 6759, 45067 Orléans Cedex 2, France (hasna.zidani@labomath.univ-orleans.fr).

$\mathbb{R}^m$, Φ is a continuous mapping from $C(\overline{D})$ into $C(\overline{D})$, $\mathcal{C} \subset C(\overline{D})$, $\overline{D}$ is a nonempty compact subset of $\overline{Q}$.) Let us consider the following class of optimal control problems:

$$\text{(P)} \qquad \inf\{J(y,v) \mid y \in W(0,T) \cap C(\overline{Q}), v \in V_{ad}, (y,v) \text{ satisfies } (1.1), (1.2), (1.3)\},$$

where V_{ad} is a subset of $\widetilde{V}_{ad}$ (to be stated precisely later), and the cost functional is defined by

$$J(y,v) = \int_Q F(x,t,y(x,t))\,dx\,dt + \int_\Sigma G(s,t,y(s,t),v(s,t))\,ds\,dt + \int_\Omega L(x,y(x,T))dx.$$

We are mainly interested in optimality conditions for such problems, in the form of Pontryagin's principles. The existence of optimal solutions for (P) is a priori supposed.

In the case where $V_{ad} \equiv \widetilde{V}_{ad}$, and $\widetilde{V}_{ad}$ is a bounded subset in $L^\infty(\Sigma)$ (the case of bounded controls), Pontryagin's principles for (P) have been obtained in [3, 9, 16, 17, 11, 25, 26, 4]. In this case the Pontryagin's principle is of the form

$$H_\Sigma(\bar{y},\bar{v},\bar{p},\bar{\nu},\bar{\lambda}) = \min_{v \in \widetilde{V}_{ad}} H_\Sigma(\bar{y},v,\bar{p},\bar{\nu},\bar{\lambda}), \tag{1.4}$$

where

$$H_\Sigma(y,v,p,\nu,\lambda) = \int_\Sigma [\nu G(s,t,y,v) - pg(s,t,y,v) + \lambda \Psi(s,t,y,v)]\; dsdt,$$

$(\bar{y},\bar{v})$ is an optimal solution, $\bar{\lambda}$ is a multiplier associated with the mixed control-state constraints (1.3), $\bar{\nu}$ is a multiplier of the cost functional, $\bar{p}$ is the adjoint state (the multiplier associated with the state constraints (1.2) only intervenes in the adjoint equation satisfied by $\bar{p}$). Notice that (1.4) can also be replaced by a pointwise Pontryagin's principle.

Observe that in [9, 16, 17, 11, 4] there is no mixed control-state constraint. Results with mixed control-state constraint are obtained in [2].

As explained in [8, p. 595] and in [21], the case of unbounded controls, that is, when $V_{ad} \equiv \widetilde{V}_{ad}$ is not bounded in $L^\infty(\Sigma)$, leads to some difficulties. In this case Pontryagin's principles are more recent results [8, 10, 21].

Now consider a control set of the form

$$V_{ad} = \{v \in \widetilde{V}_{ad} \mid v \text{ satisfies } (1.6)\} \tag{1.5}$$

with

$$\begin{aligned} &\int_\Sigma h_i(s,t,v(s,t))\,dsdt = 0, \;\; 1 \le i \le \ell_0, \\ &\int_\Sigma h_i(s,t,v(s,t))\,dsdt \le 0, \;\; \ell_0 + 1 \le i \le \ell, \end{aligned} \tag{1.6}$$

where $h = (h_1, \ldots, h_\ell)$ is a function with values in $\mathbb{R}^\ell$. Obviously control constraints (1.6) can be considered as a particular case of mixed control-state constraints (1.3). The corresponding Pontryagin's principle for the problem (P), with the control set V_{ad} defined by (1.5), may be written in the form

$$H_\Sigma(\bar{y},\bar{v},\bar{p},\bar{\nu},\bar{\lambda},\hat{\lambda}) = \min_{v \in \widetilde{V}_{ad}} H_\Sigma(\bar{y},v,\bar{p},\bar{\nu},\bar{\lambda},\hat{\lambda}), \tag{1.7}$$

where $\hat{\lambda}$ is a multiplier for the control constraints (1.6).

The novelty of our paper is the following Pontryagin's principle for the problem (P) (Theorem 2.1):

$$H_\Sigma(\bar{y}, \bar{v}, \bar{p}, \bar{\nu}, \bar{\lambda}) = \min_{v \in V_{ad}} H_\Sigma(\bar{y}, v, \bar{p}, \bar{\nu}, \bar{\lambda}), \tag{1.8}$$

when the control set V_{ad} is defined by (1.5). Let us insist on the fact that the minimum in (1.7) is stated with controls in $\widetilde{V}_{ad}$, while in (1.8) it is stated with controls in V_{ad}. Since V_{ad} takes integral control constraints of isoperimetric type into account, the result is of a different nature. As an application, we are able to prove a Pontryagin's principle (Corollary 2.2) for local solutions of (P) (local in the $L^\sigma(\Sigma)$-sense). To our knowledge, this result is completely new. Control problems for semilinear elliptic equations, with integral control constraints, are considered in [5], but the Pontryagin's principle for local solutions was not obtained there. Also we can deduce from (1.8) the classical pointwise Pontryagin's principle for local solutions in $L^\sigma(\Sigma)$ of the previous control problems; see Corollaries 2.3 and 2.4.

Let us finally mention that we deal with parabolic equations of the form (1.1), where the coefficients of the operator A are not regular, and where the nonlinear terms $f(x,t,\cdot)$ and $g(s,t,\cdot)$ are neither Lipschitz nor monotone with respect to y. When $g(s,t,\cdot,v)$ is Lipschitz and monotone such an equation is studied in [4] for bounded controls. For unbounded controls, when $g(s,t,\cdot,v)$ is neither Lipschitz nor monotone, but when the coefficients of A are time independent and regular, (1.1) is studied in [20, 21] by means of estimates on analytic semigroups. Here we combine these different difficulties. Equation (1.1) and the adjoint state equation are studied in section 3.

Our main results are stated in section 2. Section 4 is devoted to the study of the metric space of the controls and to the existence of diffuse perturbations of controls. These perturbations are the key for the proof of Pontryagin's principle, which is done in section 5.

2. Main results.

We set $\overline{\Omega}_0 = \overline{\Omega} \times \{0\}$ and $\overline{\Omega}_T = \overline{\Omega} \times \{T\}$. For every $1 \le \tau \le \infty$, the usual norms of the spaces $L^\tau(\Omega)$, $L^\tau(\Gamma)$, $L^\tau(Q)$, $L^\tau(\Sigma)$ will be denoted by $\|\cdot\|_{\tau,\Omega}$, $\|\cdot\|_{\tau,\Gamma}$, $\|\cdot\|_{\tau,Q}$, $\|\cdot\|_{\tau,\Sigma}$. For every $t > 0$, we define the norm $\|.\|_{Q(t)}$ by $\|y\|^2_{Q(t)} := \|y\|^2_{L^2(0,t;H^1(\Omega))} + \|y\|^2_{L^\infty(0,t;L^2(\Omega))}$. The Hilbert space $W(0,T;H^1(\Omega),(H^1(\Omega))') = \{y \in L^2(0,T;H^1(\Omega)) \mid \frac{dy}{dt} \in L^2(0,T;(H^1(\Omega))')\}$, endowed with its usual norm, will be denoted by $W(0,T)$. We denote by V_{ad} the set of admissible controls

$$V_{ad} := \{v \in \widetilde{V}_{ad} \mid v \text{ satisfies } (1.6)\}.$$

2.1. Assumptions.

(A1) The operator A is defined by

$$Ay(x,t) = -\sum_{i=1}^N D_i\left(\sum_{j=1}^N (a_{ij}(x,t)D_j y(x,t)) + a_i(x,t)y(x,t)\right) + \sum_{i=1}^N (b_i(x,t)D_i y(x,t)),$$

the coefficients a_{ij} belong to $L^\infty(Q)$, a_i and b_i belong to $L^{2q}(Q)$, and

$$\Lambda|\xi|^2 \le \sum_{i,j=1}^N a_{ij}(x,t)\xi_j\xi_i \quad \text{for all } \xi \in \mathbb{R}^N \text{ and for a.e. } (x,t) \in Q \text{ with } \Lambda > 0. \tag{2.1}$$

We make the following assumptions on f, g, F, G, L, Φ, Ψ.

(A2) For every $y \in \mathbb{R}$, $f(\cdot, y)$ is measurable on Q. For almost every $(x,t) \in Q$, $f(x,t,\cdot)$ is of class C^1 on $\mathbb{R}$. The following estimates hold:

$$|f(x,t,0)| \leq M_1(x,t), \qquad C_0 \leq f'_y(x,t,y) \leq M_1(x,t)\eta(|y|),$$

where M_1 belongs to $L^q(Q)$, η is a nondecreasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$ and $C_0 \in \mathbb{R}$. (We have denoted by f'_y the partial derivative of f with respect to y, throughout what follows we adopt the same kind of notation for other functions.)

(A3) For every $(y,v) \in \mathbb{R}^2$, $g(\cdot, y, v)$ is measurable on Σ. For almost every $(s,t) \in \Sigma$ and every $v \in \mathbb{R}$, $g(s,t,\cdot,v)$ is of class C^1 on $\mathbb{R}$. For almost every $(s,t) \in \Sigma$, $g(s,t,\cdot)$ and $g'_y(s,t,\cdot)$ are continuous on $\mathbb{R} \times \mathbb{R}$. The following estimates hold:

$$|g(s,t,0,v)| \leq M_2(s,t) + \Lambda_1|v|, \qquad C_0 \leq g'_y(s,t,y,v) \leq (M_2(s,t) + \Lambda_1|v|)\eta(|y|),$$

where M_2 belongs to $L^\sigma(\Sigma)$, $\Lambda_1 > 0$, C_0 and η are as in (A2).

(A4) For every $y \in \mathbb{R}$, $L(\cdot, y)$ is measurable on Ω. For almost every $x \in \Omega$, $L(x,\cdot)$ is of class C^1 on $\mathbb{R}$. The following estimate holds:

$$|L(x,y)| + |L'_y(x,y)| \leq M_3(x)\eta(|y|),$$

where $M_3 \in L^1(\Omega)$, η is as in (A2).

(A5) For every $y \in \mathbb{R}$, $F(\cdot, y)$ is measurable on Q. For almost every $(x,t) \in Q$, $F(x,t,\cdot)$ is of class C^1 on $\mathbb{R}$. The following estimate holds:

$$|F(x,t,y)| + |F'_y(x,t,y)| \leq M_4(x,t)\eta(|y|),$$

where $M_4 \in L^1(Q)$, η is as in (A2).

(A6) For every $(y,v) \in \mathbb{R}^2$, $G(\cdot, y, v)$ is measurable on Σ. For almost every $(s,t) \in \Sigma$ and every $v \in \mathbb{R}$, $G(s,t,\cdot,v)$ is of class C^1 on $\mathbb{R}$. For almost every $(s,t) \in \Sigma$, $G(s,t,\cdot)$ and $G'_y(s,t,\cdot)$ are continuous on $\mathbb{R} \times \mathbb{R}$. The following estimates hold:

$$-M_5(s,t) - \Lambda_1|v|^{\bar{\sigma}} \leq G(s,t,0,v) \leq M_5(s,t) + \Lambda_1|v|^\sigma,$$

$$|G'_y(s,t,y,v)| \leq (M_5(s,t) + \Lambda_1|v|^{\bar{\sigma}})\eta(|y|),$$

where $M_5 \in L^1(\Sigma)$, Λ_1 and η are as in (A3).

(A7) The function $h = (h_1, \ldots, h_\ell)$ is a Carathéodory function from $\Sigma \times \mathbb{R}$ into $\mathbb{R}^\ell$ satisfying

$$|h_i(s,t,v)| \leq M_5(s,t) + \Lambda_1|v|^{\bar{\sigma}} \quad \text{for } i = 1, \ldots, \ell_0,$$

$$-M_5(s,t) - \Lambda_1|v|^{\bar{\sigma}} \leq h_i(s,t,v) \leq M_5(s,t) + \Lambda_1|v|^\sigma \quad \text{for } i = \ell_0 + 1, \ldots, \ell;$$

Λ_1 and M_5 are the same as above.

(A8) The function $\Psi = (\Psi_1, \ldots, \Psi_m)$ is a Carathéodory function from $\Sigma \times \mathbb{R}^2$ into $\mathbb{R}^m$. For almost every $(s,t) \in \Sigma$ and every $v \in \mathbb{R}$, $\Psi(s,t,\cdot,v)$ is of class C^1 on $\mathbb{R}$.

For almost every $(s,t)\in\Sigma$, $\Psi'_y(s,t,\cdot)$ is continuous on $\mathbb{R}\times\mathbb{R}$. The following estimates hold:

$$|\Psi_i(s,t,0,v)|\le M_5(s,t)+\Lambda_1|v|^{\bar\sigma}\quad\text{for } i=1,\dots,m_0,$$

$$-M_5(s,t)-\Lambda_1|v|^{\bar\sigma}\le\Psi_i(s,t,0,v)\le M_5(s,t)+\Lambda_1|v|^{\sigma}\quad\text{for } i=m_0+1,\dots,m,$$

$$|\Psi'_{iy}(s,t,y,v)|\le(M_5(s,t)+\Lambda_1|v|^{\bar\sigma})\eta(|y|)\quad\text{for } i=1,\dots,m,$$

where Λ_1, M_5, η are as before. We also suppose that the function $\Phi : C(\overline{D})\to C(\overline{D})$ is of class C^1, and that $\mathcal{C}\subset C(\overline{D})$ is a closed convex subset of finite codimension in $C(\overline{D})$, where $\overline{D}$ is a compact subset of $\overline{Q}$.

2.2. Statement of the main result. We define the boundary Hamiltonian function by

$$H_\Sigma(y,v,p,\nu,\lambda)=\int_\Sigma[\nu G(s,t,y,v)-pg(s,t,y,v)+\lambda\Psi(s,t,y,v)]\,dsdt$$

for every $(y,v,p,\nu,\lambda)\in C(\overline{Q})\times L^\sigma(\Sigma)\times L^{\sigma'}(\Sigma)\times\mathbb{R}^{1+m}$. (Here $\lambda=(\lambda^1,\dots,\lambda^m)$, $\lambda\Psi(s,t,y,v)=\sum_{i=1}^m\lambda^i\Psi_i(s,t,y,v)$. Throughout the paper we adopt the same kind of notation for scalar products in $\mathbb{R}^m$.)

THEOREM 2.1. *If* (A1)–(A8) *are fulfilled and if* $(\bar y,\bar v)$ *is a solution of* (P), *then there exist* $\bar p\in L^1(0,T;W^{1,1}(\Omega))$, $\bar\nu\in\mathbb{R}$, $\bar\lambda\in\mathbb{R}^m$, $\bar\mu\in\mathcal{M}(\overline{D})$ (*the space of Radon measures on* $\overline{D}$) *such that*

$$(\bar\nu,\bar\lambda,\bar\mu)\ne0,\ \ \bar\nu\ge0,\ \textit{for } m_0+1\le i\le m,\ \bar\lambda_i\ge0,\ \bar\lambda_i\int_\Sigma\Psi_i(s,t,\bar y,\bar v)\,dsdt=0, \tag{2.2}$$

$$\langle\bar\mu,z-\Phi(\bar y)\rangle_{\overline{D}}\le0\qquad\textit{for all } z\in\mathcal{C}, \tag{2.3}$$

$$\begin{cases}-\dfrac{\partial\bar p}{\partial t}+A^*\bar p+f'_y(x,t,\bar y)\bar p=\bar\nu F'_y(x,t,\bar y)+[\Phi'(\bar y)^*\bar\mu]|_Q & \textit{in } Q,\\[2ex] \dfrac{\partial\bar p}{\partial n_{A^*}}+g'_y(s,t,\bar y,\bar v)\bar p=\bar\nu G'_y(s,t,\bar y,\bar v)+\bar\lambda\Psi'_y(s,t,\bar y,\bar v)+[\Phi'(\bar y)^*\bar\mu]|_\Sigma & \textit{on } \Sigma,\\[2ex] \bar p(T)=\bar\nu L'_y(x,\bar y(T))+[\Phi'(\bar y)^*\bar\mu]|_{\overline{\Omega}_T} & \textit{on } \overline{\Omega},\end{cases} \tag{2.4}$$

$$\bar p\in L^{\delta'}(0,T;W^{1,d'}(\Omega))\qquad\textit{for every } (\delta,d)\ \textit{satisfying}\ \frac{N}{2d}+\frac{1}{\delta}<\frac{1}{2}, \tag{2.5}$$

$$H_\Sigma(\bar y,\bar v,\bar p,\bar\nu,\bar\lambda)=\min_{v\in V_{ad}}H_\Sigma(\bar y,v,\bar p,\bar\nu,\bar\lambda), \tag{2.6}$$

where $[\Phi'(\bar y)^*\bar\mu]|_Q$ *is the restriction of* $[\Phi'(\bar y)^*\bar\mu]$ *to* Q, $[\Phi'(\bar y)^*\bar\mu]|_\Sigma$ *is the restriction of* $[\Phi'(\bar y)^*\bar\mu]$ *to* Σ, *and* $[\Phi'(\bar y)^*\bar\mu]|_{\overline{\Omega}_T}$ *is the restriction of* $[\Phi'(\bar y)^*\bar\mu]$ *to* $\overline{\Omega}_T$, $[\Phi'(\bar y)^*\bar\mu]$ *is the Radon measure on* $\overline{D}$ *defined by* $z\longmapsto\langle\bar\mu,\Phi'(\bar y)z\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})}$ *for* $z\in C(\overline{D})$, $\langle\cdot,\cdot\rangle_{\overline{D}}$

denotes the duality pairing between $\mathcal{M}(\overline{D})$ *and* $C(\overline{D})$*,* A^* *is the formal adjoint of* A*, that is,*

$$A^*y(x,t) = -\sum_{i=1}^{N} D_i\left(\sum_{i=1}^{N}(a_{ji}(x,t)D_jy(x,t)) + b_i(x,t)y(x,t)\right) + \sum_{i=1}^{N} a_i(x,t)D_iy(x,t).$$

2.3. Pontryagin's principles for local solutions. By definition, a local solution $(\bar{y}, \bar{v})$ of (P) in $L^\sigma(\Sigma)$ is a solution of the problem

$$(\mathrm{P}_{\bar{v},\epsilon})\quad \inf\{J(y,v) \mid y \in C(\overline{Q}),\ v \in \widetilde{V}_{ad},\quad (y,v) \text{ satisfies (1.1)–(1.3)},\ \|\bar{v} - v\|_{\sigma,\Sigma} \leq \epsilon\}$$

for some $\epsilon > 0$. The following Pontryagin's principle for local solutions of (P) is a direct consequence of Theorem 2.1.

Corollary 2.2. *If* (A1)–(A8) *are fulfilled and if* $(\bar{y}, \bar{v})$ *is a solution of* $(\mathrm{P}_{\bar{v},\epsilon})$*, there then exist* $\bar{p} \in L^1(0,T; W^{1,1}(\Omega))$*,* $\bar{\nu} \in \mathbb{R}$*,* $\bar{\lambda} \in \mathbb{R}^m$*,* $\bar{\mu} \in \mathcal{M}(\overline{D})$ *satisfying* (2.2)–(2.5) *along with*

$$H_\Sigma(\bar{y}, \bar{v}, \bar{p}, \bar{\nu}, \bar{\lambda}) = \min_{v \in V_{ad}, \|\bar{v}-v\|_{\sigma,\Sigma} \leq \epsilon} H_\Sigma(\bar{y}, v, \bar{p}, \bar{\nu}, \bar{\lambda}).$$

As a consequence of this corollary we can get the classical pointwise Pontryagin principle for a local solution in $L^\sigma(\Sigma)$ of the control problem

$$(\tilde{\mathrm{P}})\qquad \inf\{J(y,v) \mid y \in W(0,T) \cap C(\overline{Q}), v \in \widetilde{V}_{ad}, (y,v) \text{ satisfies } (1.1), (1.2), (1.3)\}.$$

Corollary 2.3. *If* (A1)–(A8) *are fulfilled and if* $(\bar{y}, \bar{v})$ *is a local solution of* $(\tilde{\mathrm{P}})$ *in* $L^\sigma(\Sigma)$*, there then exist* $\bar{p} \in L^1(0,T; W^{1,1}(\Omega))$*,* $\bar{\nu} \in \mathbb{R}$*,* $\bar{\lambda} \in \mathbb{R}^m$*,* $\bar{\mu} \in \mathcal{M}(\overline{D})$ *satisfying* (2.2)–(2.5) *along with*

$$\mathcal{H}_\Sigma(s,t,\bar{y}(s,t),\bar{v}(s,t),\bar{p}(s,t),\bar{\nu},\bar{\lambda}) = \min_{\xi \in V(s,t)} \mathcal{H}_\Sigma(s,t,\bar{y}(s,t),\xi,\bar{p}(s,t),\bar{\nu},\bar{\lambda})$$

for almost all $(s,t) \in \Sigma$*, where*

$$\mathcal{H}_\Sigma(s,t,y,\xi,p,\nu,\lambda) = \nu G(s,t,y,\xi) - pg(s,t,y,\xi) + \lambda\Psi(s,t,y,\xi).$$

Proof. The pointwise Pontryagin's principle stated in the corollary may be derived from the integral Pontryagin's principle of Corollary 2.2 by using the same construction as in [21, proof of Theorem 2.1]. The idea in the proof of [21] is to construct a pointwise perturbation v_n of $\bar{v}$ such that $\lim_{(s,t)\to(s_0,t_0)} v_n(s,t) = \xi$, $\lim_n \mathcal{L}^N(\{(s,t) \in \Sigma \mid v_n(s,t) \neq \bar{v}(s,t)\}) = 0$, where $\xi \in V(s_0,t_0)$, $(s_0,t_0) \in \Sigma$, $\mathcal{L}^N$ denotes the N-dimensional Lebesgue measure. We obtain the pointwise Pontryagin's principle by replacing v by v_n in the integral Pontryagin's principle of Corollary 2.2, by dividing by $\mathcal{L}^N(\{(s,t) \in \Sigma \mid v_n(s,t) \neq \bar{v}(s,t)\}) \neq 0$, and by passing to the limit when n tends to infinity. The only difference with [21] is that v_n must satisfy $\|\bar{v} - v_n\|_{\sigma,\Sigma} \leq \epsilon$. Due to the condition $\lim_n \mathcal{L}^N(\{(s,t) \in \Sigma \mid v_n(s,t) \neq \bar{v}(s,t)\}) = 0$, it is clear that this condition will be realized for n big enough. □

Let us observe that integral control constraints can be studied in the framework of the problem $(\tilde{\mathrm{P}})$. Indeed, the mixed constraints (1.3) can include the integral control constraints. Then Corollary 2.3 provides a Pontryagin's principle for problems with integral constraints on the control and the state, even with mixed integral constraints, as well as pointwise constraints on the control and state too. The corresponding result is stated in the following corollary.

Corollary 2.4. *If* (A1)–(A8) *are fulfilled and if* $(\bar{y}, \bar{v})$ *is a local solution of* $(\tilde{\mathrm{P}})$ *in* $L^\sigma(\Sigma)$*, there then exist* $\bar{p} \in L^1(0, T; W^{1,1}(\Omega))$*,* $\bar{\nu} \in \mathbb{R}$*,* $\bar{\lambda} \in \mathbb{R}^m$*,* $\hat{\lambda} \in \mathbb{R}^\ell$*, and* $\bar{\mu} \in \mathcal{M}(\overline{D})$ *such that*

$$(\bar{\nu}, \bar{\lambda}, \hat{\lambda}, \bar{\mu}) \neq 0, \quad \bar{\nu} \geq 0, \quad \langle \bar{\mu}, z - \Phi(\bar{y}) \rangle_{\overline{D}} \leq 0 \quad \textit{for all } z \in \mathcal{C},$$

$$\bar{\lambda}_i \int_\Sigma \Psi_i(s, t, \bar{y}, \bar{v})\, dsdt = 0 \;\; \textit{for} \;\; 1 \leq i \leq m, \;\; \bar{\lambda}_i \geq 0 \;\; \textit{for} \;\; m_0 + 1 \leq i \leq m,$$

$$\hat{\lambda}_i \int_\Sigma h_i(s, t, \bar{v})\, dsdt = 0 \;\; \textit{for} \;\; 1 \leq i \leq \ell, \;\; \hat{\lambda}_i \geq 0 \;\; \textit{for} \;\; \ell_0 + 1 \leq i \leq \ell,$$

$$\begin{cases} -\dfrac{\partial \bar{p}}{\partial t} + A^* \bar{p} + \bar{f}'_y\, \bar{p} = \bar{\nu} \bar{F}'_y + [\Phi'(\bar{y})^* \bar{\mu}]|_Q & \textit{in } Q, \\[2ex] \dfrac{\partial \bar{p}}{\partial n_{A^*}} + \bar{g}'_y\, \bar{p} = \bar{\nu} \bar{G}'_y + \bar{\lambda} \bar{\Psi}'_y + \hat{\lambda} \bar{h} + [\Phi'(\bar{y})^* \bar{\mu}]|_\Sigma & \textit{on } \Sigma, \\[2ex] \bar{p}(T) = \bar{\nu} L'_y(x, \bar{y}(T)) + [\Phi'(\bar{y})^* \bar{\mu}]|_{\overline{\Omega}_T} & \textit{on } \overline{\Omega}, \end{cases}$$

where $\bar{f}'_y$ *stands for* $\bar{f}'_y(x, t, \bar{y})$*,* $\bar{G}'_y$ *for* $\bar{G}'_y(s, t, \bar{y}, \bar{v})$*, and the same convention is used for other functions. Also, the following pointwise Pontryagin's principle holds:*

$$\mathcal{H}_\Sigma(s, t, \bar{y}(s, t), \bar{v}(s, t), \bar{p}(s, t), \bar{\nu}, \bar{\lambda}, \hat{\lambda}) = \min_{\xi \in V(s,t)} \mathcal{H}_\Sigma(s, t, \bar{y}(s, t), \xi, \bar{p}(s, t), \bar{\nu}, \bar{\lambda}, \hat{\lambda})$$

for almost all $(s, t) \in \Sigma$*, where*

$$\mathcal{H}_\Sigma(s, t, y, \xi, p, \nu, \bar{\lambda}, \hat{\lambda}) = \nu G(s, t, y, \xi) - p g(s, t, y, \xi) + \bar{\lambda} \Psi(s, t, y, \xi) + \hat{\lambda} h(s, t, \xi).$$

3. State and adjoint equations.

3.1. State equation. Existence and regularity results for (1.1) and (2.4) rely on estimates in $C(\overline{Q})$ for solutions of linear equations of the form

$$(3.1)\quad \frac{\partial y}{\partial t} + Ay + ay = \phi - \mathrm{div}\xi \quad \text{in } Q, \quad \frac{\partial y}{\partial n_A} + by = \psi \quad \text{on } \Sigma, \quad y(0) = y_0 \quad \text{in } \Omega.$$

If assumption (A1) is satisfied, if $(a, \phi) \in L^q(Q) \times L^q(Q)$, $(b, \psi) \in L^{\bar{\sigma}}(\Sigma) \times L^{\bar{\sigma}}(\Sigma)$, the existence of a unique solution in $C([0, T]; L^2(\Omega)) \cap L^2([0, T]; H^1(\Omega))$ for (3.1) is proved in [12, Chapter 3, Theorem 5.1] when $\xi \equiv 0$. The result can be extended to (3.1) by the same method if the support of ξ is compact in Q and if ξ belongs to $L^\delta(0, T, (L^d(\Omega))^N)$ with $d > 1, \delta > 1, N/2d + 1/\delta < 1/2$. Recall that a weak solution in $L^2(0, T; H^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ of (3.1) is a function $y \in L^2(0, T; H^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ satisfying

$$\int_Q \left(-y \frac{\partial z}{\partial t} + \sum_{i,j} a_{ij} D_j y D_i z + \sum_i (a_i y D_i z + b_i D_i y z) + ayz \right) dxdt + \int_\Sigma byz\, dsdt$$
$$= \int_Q \left[\phi z + \sum_i \xi_i D_i z \right] dxdt + \int_\Sigma \psi z\, dsdt + \int_\Omega y(0) z(0)\, dx$$

for every $z \in C^1(\overline{Q})$ such that $z(\cdot, T) = 0$ on $\overline{\Omega}$. For linear equations with Dirichlet boundary conditions

$$\frac{\partial y}{\partial t} + Ay + ay = \phi - \text{div}\xi \quad \text{in } Q, \qquad y = \psi \quad \text{on } \Sigma, \qquad y(0) = y_0 \quad \text{in } \Omega,$$

estimates of the form

$$\|y\|_{L^\infty(Q)} \leq C \left(\|\phi\|_{q,Q} + \|\psi\|_{\infty,\Sigma} + \sum_i \|\xi_i\|_{d,\delta,\Omega} + \|y_0\|_{C(\overline{\Omega})} \right) \tag{3.2}$$

are obtained in [12, Chapter 3, Theorem 7.1] for $d > 1, \delta > 1, N/2d + 1/\delta < 1/2$. In this estimate the constant C depends on $T, \Omega, N, \Lambda, q, \bar{\sigma}, \delta, d, \sum_i \|a_i^2\|_{q,Q}, \sum_i \|b_i^2\|_{q,Q}$, but also on $\|a\|_{q,Q}$. The case of Robin boundary conditions is considered in [4] to study nonlinear equations of the form (1.1) when the function $g(s, t, \cdot, v)$, in the boundary condition, is monotone and Lipschitz, and when the boundary control v is bounded [4, Theorem 5.1]. The case when the function $g(s, t, \cdot, v)$ in (1.1) is neither Lipschitz nor monotone (g satisfies (A3)), and when the control v belongs to $L^{\bar{\sigma}}(\Sigma)$, but when the coefficients of the operator A are regular and independent of the time variable, is studied in [19]. Estimates in $C(\overline{Q})$ are obtained by semigroup techniques and comparison principles [19, Proposition 3.3 and Theorem 3.1]. Here we emphasize the fact that assumptions on the operator A are minimal (bounded leading coefficients, unbounded coefficients of order zero), that we deal with nonhomogeneous boundary conditions, and that source terms in the domain and in the boundary conditions are unbounded.

THEOREM 3.1. *Under assumptions* (A1)–(A3), *if* $v \in L^{\bar{\sigma}}(\Sigma)$, *then* (1.1) *admits a unique weak solution* y_v *in* $W(0, T) \cap C(\overline{Q})$. *This solution obeys*

$$\|y_v\|_{C(\overline{Q})} \leq C_1(\|v\|_{\bar{\sigma},\Sigma} + 1),$$

where $C_1 = C_1(T, \Omega, N, C_0, q, \bar{\sigma})$. *Moreover, the mapping* $v \longmapsto y_v$ *is continuous from* $L^{\bar{\sigma}}(\Sigma)$ *into* $C(\overline{Q})$.

Proof. The proof relies on Theorem 3.2 (see [19]). □

THEOREM 3.2. *Suppose that* (A1) *is satisfied,* $(a, \phi) \in L^q(Q) \times L^q(Q)$, $(b, \psi) \in L^{\bar{\sigma}}(\Sigma) \times L^{\bar{\sigma}}(\Sigma)$, *and* ξ *belongs to* $(\mathcal{D}(Q))^N$. *If in addition* $a \geq C_0$ *a.e. in* Q *and* $b \geq C_0$ *a.e. in* Σ *(for some* $C_0 \in \mathbb{R}$*), then the unique weak solution* y *of* (3.1) *belongs to* $C(\overline{Q})$ *and satisfies the following estimate:*

$$\|y\|_{C(\overline{Q})} \leq C_2 \left(\|\phi\|_{q,Q} + \|\psi\|_{\bar{\sigma},\Sigma} + \sum_i \|\xi_i\|_{L^\delta(0,T;L^d(\Omega))} + \|y_0\|_{C(\overline{\Omega})} \right),$$

where $d > 1$, $\delta > 1$ *satisfy* $N/2d + 1/\delta < 1/2$ *and the constant* C_2 *only depends on* T, Ω, N, C_0, Λ, q, $\bar{\sigma}$, δ, d, $\sum_i \|a_i^2\|_{q,Q}$, $\sum_i \|b_i^2\|_{q,Q}$.

Remark 3.3. Notice that the constant C_2 does not depend on $\|a\|_{q,Q}$ and $\|b\|_{\bar{\sigma},\Sigma}$. As in [12] (see the above estimate (3.2)), the assumption $a \geq C_0$ may be dropped out, and in this case the constant C_2 must be replaced by a constant also depending on $\|a\|_{q,Q}$. But the corresponding estimate cannot be used to treat nonlinear equations of the form (1.1).

Proof. To prove this theorem, we need only to establish the L^∞-estimate; the rest is classical. We prove the L^∞-estimate by using the so-called truncation method as in [12, Chapter 3, proof of Theorem 7.1]. If y is a weak solution of (3.1), then we have

$$\int_\Omega [y(x,t)z(x,t) - y(x,0)z(x,0)]\,dx$$

$$+\int_0^t\int_\Omega \left[-y\frac{\partial z}{\partial t} + \sum_{i,j} a_{ij}D_jyD_iz + \sum_i (a_iyD_iz + b_iD_iyz) + ayz\right] dxd\tau$$

$$+\int_0^t\int_\Gamma byz\,dsd\tau = \int_0^t\int_\Omega \left[\phi z + \sum_i \xi_iD_iz\right] dxd\tau + \int_0^t\int_\Gamma \psi z\,dsd\tau$$

for every $t \in [0,T]$ and every $z \in W_2^{1,1}(Q)$. We establish only the upper bound for y. (The lower bound can be obtained in the same way.) For $k \geq 0$ we set $y^k(x,t) = \max(y(x,t) - k, 0)$. By using Steklov averagings, as in [12, p. 183], we prove that

$$\frac{1}{2}\int_\Omega [y^k(x,t)^2 - y^k(x,0)^2]\,dx \tag{3.3}$$

$$+\int_0^t\int_\Omega \left[\sum_{i,j} a_{ij}D_jy^kD_iy^k + \sum_i (a_iyD_iy^k + b_iD_iy^ky^k) + ayy^k\right] dxd\tau$$

$$+\int_0^t\int_\Gamma byy^k\,dsd\tau = \int_0^t\int_\Omega \left[\phi y^k + \sum_i \xi_iD_iy^k\right] dxd\tau + \int_0^t\int_\Gamma \psi y^k\,dsd\tau$$

for every $t \in]0,T]$. Thus, it follows that

$$\frac{1}{2}\int_\Omega y^k(x,t)^2\,dx + \int_0^t\!\!\int_\Omega \left[\sum_{i,j} a_{ij}D_jy^kD_iy^k + (a - C_0 + \Lambda)yy^k\right] dxd\tau \tag{3.4}$$

$$+\int_0^t\!\!\int_\Gamma (b - C_0)yy^k\,dsd\tau = -\int_0^t\!\!\int_\Omega \left[\sum_i (a_iyD_iy^k + b_iD_iy^ky^k) + (C_0 - \Lambda)yy^k\right] dxd\tau$$

$$-\int_0^t\!\!\int_\Gamma C_0yy^k\,dsd\tau + \int_0^t\!\!\int_\Omega \left[\phi y^k + \sum_i \xi_iD_iy^k\right] dxd\tau + \int_0^t\int_\Gamma \psi y^k\,dsd\tau$$

for every $k > \tilde{k} := \|y_0\|_{C(\overline{\Omega})}$. Since $a - C_0 \geq 0$ a.e. in Q, $b - C_0 \geq 0$ a.e. on Σ, and $yy^k \geq (y^k)^2$ a.e. in Q, with (2.1) we obtain

$$\|y^k(t)\|_{2,\Omega}^2 + 2\Lambda\|y^k\|_{L^2(0,t;H^1(\Omega))}^2 \tag{3.5}$$

$$\leq -2\int_0^t\!\!\int_\Omega \left[\sum_i (a_iyD_iy^k + b_iD_iy^ky^k) + (C_0 - \Lambda)yy^k\right] dxd\tau$$

$$-2\int_0^t\!\!\int_\Gamma C_0yy^k\,dsd\tau + 2\int_0^t\!\!\int_\Omega \left[\phi y^k + \sum_i \xi_iD_iy^k\right] dxd\tau + 2\int_0^t\!\!\int_\Gamma \psi y^k\,dsd\tau$$

for every $k > \tilde{k}$. Set $A_k(t) = \{x \in \Omega \mid y(x,t) > k\}$, $B_k(t) = \{s \in \Gamma \mid y(s,t) > k\}$, $Q_k(t) = \{(x,\tau) \in \Omega\times]0,t[| \ y(x,\tau) > k\}$, $\Sigma_k(t) = \{(s,\tau) \in \Gamma\times]0,t[| \ y(s,\tau) > k\}$. We estimate the terms in the right-hand side of (3.5) by means of Young's inequality and we obtain

$$\|y^k(t)\|_{2,\Omega}^2 + \Lambda\|y^k\|_{L^2(0,t;H^1(\Omega))}^2$$

$$\leq \frac{3}{\Lambda}\int_0^t\int_{A_k(\tau)}\left[\sum_i((a_iy)^2 + (b_iy^k)^2) + (C_0 - \Lambda)^2y^2\right] dxd\tau + \frac{3K^2}{\Lambda}\int_0^t\int_{B_k(\tau)} C_0^2y^2\, dsd\tau$$

$$+2\int_0^t\int_{A_k(\tau)}\left[|\phi||y^k| + \sum_i|\xi_i||D_iy^k|\right] dxd\tau + 2\int_0^t\int_{B_k(\tau)}|\psi||y^k|\, dsd\tau,$$

where $K > 0$ satisfies $\|\varphi\|_{2,\Gamma} \leq K\|\varphi\|_{H^1(\Omega)}$ for all $\varphi \in H^1(\Omega)$. Since $y = y^k + k$ in $A_k(\tau)$ and $B_k(\tau)$ for a.e. τ, it follows that

$$\|y^k(t)\|_{2,\Omega}^2 + \Lambda\|y^k\|_{L^2(0,t;H^1(\Omega))}^2$$

$$\leq \frac{6}{\Lambda}\int_0^t\int_{A_k(\tau)}\left[\sum_i(a_i^2 + b_i^2) + (C_0 - \Lambda)^2\right]((y^k)^2 + k^2)\, dxd\tau$$

$$+\frac{6K^2}{\Lambda}\int_0^t\int_{B_k(\tau)} C_0^2((y^k)^2 + k^2)\, dsd\tau + 2\int_0^t\int_{A_k(\tau)}\left[|\phi||y^k| + \sum_i|\xi_i||D_iy^k|\right] dxd\tau$$

$$+2\int_0^t\int_{B_k(\tau)}|\psi||y^k|\, dsd\tau$$

for every $t \in [0,T]$ and every $k > \tilde{k}$. With Hölder's inequality we have

$$\|y^k(t)\|_{2,\Omega}^2 + \Lambda\|y^k\|_{L^2(0,t;H^1(\Omega))}^2 \tag{3.6}$$

$$\leq \Big(K_1(|Q_k(t)|^{\frac{1}{q'}} + |Q_k(t)|) + K_2|\Sigma_k(t)|\Big)\, k^2$$

$$+K_1(|Q_k(t)|^{\frac{2}{N+2}} + |Q_k(t)|^{\frac{1}{q'}-\frac{N}{N+2}})\|y^k\|_{\frac{2(N+2)}{N},\Omega\times]0,t[}^2$$

$$+2\|\phi\|_{q,Q}|Q_k(t)|^{\frac{1}{q'}-\frac{N}{2(N+2)}}\|y^k\|_{\frac{2(N+2)}{N},\Omega\times]0,t[}$$

$$+K_2|\Sigma_k(t)|^{\frac{1}{N+1}}\|y^k\|_{\frac{2(N+1)}{N},\Gamma\times]0,t[}^2 + 2\|\psi\|_{\bar{\sigma},\Sigma}|\Sigma_k(t)|^{\frac{1}{\bar{\sigma}'}-\frac{N}{2(N+1)}}\|y^k\|_{\frac{2(N+1)}{N},\Gamma\times]0,t[}$$

$$+\frac{\Lambda}{2}\|y^k\|_{L^2(0,t;H^1(\Omega))}^2 + \frac{2K_3}{\Lambda}\left(\int_0^t |A_k(\tau)|^{\frac{\delta(d-2)}{d(\delta-2)}}\, d\tau\right)^{\frac{\delta-2}{\delta}},$$

where

$$K_1 = \frac{6}{\Lambda}\left[\sum_i(\|a_i^2\|_{q,Q} + \|b_i^2\|_{q,Q}) + (C_0 - \Lambda)^2\right],$$

$$K_2 = \frac{6K^2}{\Lambda} C_0^2 \quad \text{and} \quad K_3 = \sum_i \|\xi_i\|^2_{L^\delta(0,T;L^d(\Omega))},$$

$|Q_k(t)|$ denotes the $(N+1)$-dimensional Lebesgue measure of $Q_k(t)$, $|\Sigma_k(t)|$ denotes the N-dimensional Lebesgue measure of $\Sigma_k(t)$, and $|A_k(\tau)|$ denotes the N-dimensional Lebesgue measure of $A_k(\tau)$. Notice that $\frac{N(d-2)}{2d} + \frac{\delta-2}{\delta} > \frac{N}{2}$. Then there exists $\tilde{r} > \frac{2\delta}{\delta-2} > 2$ such that $\frac{N(d-2)}{2d} + \frac{\delta-2}{\delta} > \frac{N}{2}\frac{\tilde{r}(\delta-2)}{2\delta} > \frac{N}{2}$. For such an $\tilde{r}$ we have $\frac{1}{\tilde{r}}(\frac{N}{2}\frac{\delta}{\delta-2}\frac{d-2}{d} + 1) > \frac{N}{4}$. We define $r > 2$ by $\frac{1}{r} = \frac{1}{\tilde{r}}\frac{\delta}{\delta-2}\frac{d-2}{d} < \frac{d-2}{2d} < \frac{1}{2}$ and we obtain $\frac{N}{2r} + \frac{1}{\tilde{r}} > \frac{N}{4}$. Thus the imbedding from $L^2(0,t;H^1(\Omega)) \cap C([0,t];L^2(\Omega))$ into $L^{\tilde{r}}(0,t;L^r(\Omega))$ is continuous; see [12, p. 75]. Observe that

$$|Q_k(t)|^{\frac{2}{N+2}} + |Q_k(t)|^{\frac{1}{q'}-\frac{N}{N+2}} \le (t|\Omega|)^{\frac{2}{N+2}} + (t|\Omega|)^{\frac{1}{q'}-\frac{N}{N+2}}, \qquad |\Sigma_k(t)|^{\frac{1}{N+1}} \le (t|\Gamma|)^{\frac{1}{N+1}}.$$

Let us choose $\bar{t} > 0$ small enough to have

$$(3.7)\quad K_1((\bar{t}|\Omega|)^{\frac{2}{N+2}} + (\bar{t}|\Omega|)^{\frac{1}{q'}-\frac{N}{N+2}})\|y\|^2_{\frac{2(N+2)}{N},\Omega\times]0,\bar{t}[} + K_2(\bar{t}|\Gamma|)^{\frac{1}{N+1}}\|y\|^2_{\frac{2(N+1)}{N},\Gamma\times]0,\bar{t}[}$$
$$\le \frac{1}{2}\min\left(1, \frac{\Lambda}{2}\right)\|y\|^2_{Q(\bar{t})}$$

for every $y \in L^2(0,\bar{t};H^1(\Omega)) \cap C([0,\bar{t}];L^2(\Omega))$. Then from (3.6) and imbedding theorems, it follows that

$$(3.8)\qquad \nu(\|y^k\|_{\frac{2(N+2)}{N},\Omega\times]0,\bar{t}[} + \|y^k\|_{\frac{2(N+1)}{N},\Gamma\times]0,\bar{t}[} + \|y^k\|_{L^{\tilde{r}}(0,\bar{t};L^r(\Omega))}) \le$$
$$\le \|y\|_{Q(\bar{t})} \le K_4\left(|Q_k(\bar{t})|^{\frac{1}{2q'}} + |Q_k(\bar{t})|^{\frac{1}{2}} + |\Sigma_k(\bar{t})|^{\frac{1}{2}}\right)k$$
$$+K_4\left(|Q_k(\bar{t})|^{\frac{1}{q'}-\frac{N}{2(N+2)}} + |\Sigma_k(\bar{t})|^{\frac{1}{\bar{\sigma}'}-\frac{N}{2(N+1)}}\right) + K_4\left(\int_0^{\bar{t}} |A_k(\tau)|^{\frac{\delta(d-2)}{d(\delta-2)}}\,d\tau\right)^{\frac{\delta-2}{2\delta}},$$

for $k > \tilde{k}$, where $\nu > 0$ depends on Λ, and where K_4 depends on K_1, K_2, K_3, $\|\phi\|_{q,Q}$, $\|\psi\|_{\bar{\sigma},\Sigma}$, and Λ. Now, we set $\theta(k) = |Q_k(\bar{t})|^{\frac{N}{2(N+2)}} + |\Sigma_k(\bar{t})|^{\frac{N}{2(N+1)}} + (\int_0^{\bar{t}} |A_k(\tau)|^{\frac{\tilde{r}}{r}}d\tau)^{\frac{1}{\tilde{r}}}$. Observe that, for every $\ell \ge k \ge 0$, we have $y^k \ge \ell - k$ a.e. in $Q_\ell(\bar{t})$, a.e. on $\Sigma_\ell(\bar{t})$, and a.e. in $A_\ell(\tau)$ for a.e. $\tau \in]0,\bar{t}[$; therefore

$$(3.9)\quad (\ell - k)\theta(\ell) \le \|y^k\|_{\frac{2(N+2)}{N},\Omega\times]0,\bar{t}[} + \|y^k\|_{\frac{2(N+1)}{N},\Gamma\times]0,\bar{t}[} + \|y^k\|_{L^{\tilde{r}}(0,\bar{t};L^r(\Omega))}.$$

Taking $k = 0$ in the above inequality, with the definition of the function θ we first obtain $\ell\theta(\ell) \le K_0$ for all $\ell \ge 0$, where $K_0 = \|y\|_{\frac{2(N+2)}{N},\Omega\times]0,\bar{t}[} + \|y\|_{\frac{2(N+1)}{N},\Gamma\times]0,\bar{t}[} + \|y\|_{L^{\tilde{r}}(0,\bar{t};L^r(\Omega))}$. In particular, for $\ell = K_0$, this implies $\theta(K_0) \le 1$. On the other hand, (3.8) and (3.9) give

$$(3.10)\qquad (\ell - k)\theta(\ell) \le \frac{K_4}{\nu}\left(|Q_k(\bar{t})|^{\frac{1}{2q'}} + |Q_k(\bar{t})|^{\frac{1}{2}} + |\Sigma_k(\bar{t})|^{\frac{1}{2}}\right.$$
$$\left.+|Q_k(\bar{t})|^{\frac{1}{q'}-\frac{N}{2(N+2)}} + |\Sigma_k(\bar{t})|^{\frac{1}{\bar{\sigma}'}-\frac{N}{2(N+1)}}\right)k + \frac{K_4}{\nu}\left(\int_0^{\bar{t}} |A_k(\tau)|^{\frac{\delta(d-2)}{d(\delta-2)}}d\tau\right)^{\frac{\delta-2}{2\delta}}$$

for all $\ell \geq k > \max(K_0, 1, \tilde{k})$. Set

$$\alpha_1 = \frac{N+2}{Nq'}, \quad \alpha_2 = \frac{N+1}{N\bar{\sigma}'}, \quad \alpha_3 = \tilde{r}\frac{\delta-2}{2\delta}, \quad \alpha = \min(\alpha_1, \alpha_2, \alpha_3),$$

and observe that $\alpha > 1$. Since $\theta(K_0) \leq 1$, and since θ is a nonincreasing function, we also have $|Q_k(\bar{t})| \leq 1$, $|\Sigma_k(\bar{t})| \leq 1$, and $\int_0^{\bar{t}} |A_k(\tau)|^{\frac{\tilde{r}}{r}} d\tau \leq 1$ for all $k \geq K_0$. Thus it follows that

$$|Q_k(\bar{t})|^{\frac{1}{2q'}} + |\Sigma_k(\bar{t})|^{\frac{1}{2}} + |Q_k(\bar{t})|^{\frac{1}{2}}$$

$$+|Q_k(\bar{t})|^{\frac{1}{q'} - \frac{N}{2(N+2)}} + |\Sigma_k(\bar{t})|^{\frac{1}{\bar{\sigma}'} - \frac{N}{2(N+1)}} + \left(\int_0^{\bar{t}} |A_k(\tau)|^{\frac{\tilde{r}}{r}} \, d\tau \right)^{\frac{\delta-2}{2\delta}} \leq 3\theta(k)^{\alpha}.$$

From (3.10), we deduce

$$(\ell - k)\theta(\ell) \leq K_5 \theta(k)^{\alpha} k \tag{3.11}$$

for every $\ell \geq k > \max(K_0, 1, \tilde{k})$. With the same arguments as in [12, Chapter 3, p. 186], still using (3.8), we finally obtain

$$\|y\|_{\infty,Q} \leq K_6, \tag{3.12}$$

where K_6 depends not only on $T, \Omega, N, C_0, \Lambda, q, \bar{\sigma}, \delta, d, \sum_i \|a_i^2\|_{q,Q}, \sum_i \|b_i^2\|_{q,Q}$, but also on K_0, $\|y_0\|_{C(\overline{\Omega})}$, $\|\phi\|_{q,Q}$, $\|\psi\|_{\bar{\sigma},\Sigma}$, and $\sum_i \|\xi_i\|^2_{L^{\delta}(0,T;L^d(\Omega))}$. The constant K_6 depends on $K_0 = \|y\|_{\frac{2(N+2)}{N},\Omega\times]0,\bar{t}[} + \|y\|_{\frac{2(N+1)}{N},\Gamma\times]0,\bar{t}[} + \|y\|_{L^{\tilde{r}}(0,\bar{t};L^r(\Omega))} \leq C\|y\|_{Q(\bar{t})}$. By using the same trick as in (3.4), we can obtain an estimate of $\|y\|_{Q(\bar{t})}$ depending on T, Ω, N, C_0, Λ, q, $\bar{\sigma}$, δ, d, $\sum_i \|a_i^2\|_{q,Q}$, $\sum_i \|b_i^2\|_{q,Q}$, $\|y_0\|_{C(\overline{\Omega})}$, $\sum_i \|\xi_i\|^2_{L^{\delta}(0,T;L^d(\Omega))}$, $\|\phi\|_{q,Q}$, and $\|\psi\|_{\bar{\sigma},\Sigma}$, but independent of $\|a\|_{q,Q}$ and $\|b\|_{\bar{\sigma},\Sigma}$. Since (3.1) is linear, the estimate given in Theorem 3.2 can be easily deduced from (3.12). □

3.2. Adjoint equation. Let (a, b) be in $L^q(Q) \times L^{\bar{\sigma}}(\Sigma)$ with $a \geq C_0$ and $b \geq C_0$. We consider the terminal boundary value problem

$$-\frac{\partial p}{\partial t} + A^* p + ap = \mu_Q \text{ in } Q, \quad \frac{\partial p}{\partial n_{A^*}} + bp = \mu_\Sigma \text{ on } \Sigma, \quad p(T) = \mu_{\overline{\Omega}_T} \text{ on } \overline{\Omega}, \tag{3.13}$$

where $\mu = \mu_Q + \mu_\Sigma + \mu_{\overline{\Omega}_T}$ is a bounded Radon measure on $\overline{Q} \backslash \overline{\Omega}_0$, μ_Q is the restriction of μ to Q, μ_Σ is the restriction of μ to Σ, and $\mu_{\overline{\Omega}_T}$ is the restriction of μ to $\overline{\Omega}_T$. A function $p \in L^1(0, T; W^{1,1}(\Omega))$ is a weak solution of (3.13) if

$$ap \in L^1(Q), \quad bp \in L^1(\Sigma), \quad a_i D_i p \in L^1(Q), \quad \text{and} \quad b_i p \in L^1(Q) \quad \text{for } i = 1, \ldots, N,$$

$$\int_Q \left(p\frac{\partial y}{\partial t} + \sum_{i,j} a_{ji} D_j p D_i y + \sum_i (a_i D_i p y + b_i p D_i y) + apy \right) dxdt + \int_\Sigma bpy \, dsdt$$

$$= \int_{\overline{Q} \backslash \overline{\Omega}_0} y d\mu(x,t) \quad \text{for every } y \in C^1(\overline{Q}) \text{ satisfying } y(x, 0) = 0 \text{ on } \overline{\Omega}.$$

As for elliptic equations [23], it is well known that (3.13) may admit more than one solution. However, uniqueness is guaranteed if we look for solutions of (3.13) satisfying some Green formula. (Such uniqueness results are proved in [1] for elliptic equations and in [4] for parabolic equations.)

THEOREM 3.4. *Let μ be in $\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)$ and let (a,b) be in $L^q(Q)\times L^{\bar\sigma}(\Sigma)$ satisfying $a\ge C_0$ a.e. in Q, $b\ge C_0$ a.e. Σ, for some $C_0\in\mathbb{R}$. Equation* (3.13) *admits a unique solution p in $L^1(0,T;W^{1,1}(\Omega))$ satisfying*

$$\int_Q p\left\{\frac{\partial y}{\partial t}+Ay+ay\right\}dxdt+\int_\Sigma p\left\{\frac{\partial y}{\partial n_A}+by\right\}dsdt=\langle y,\mu\rangle_{C_b(\overline{Q}\setminus\overline{\Omega}_0)\times\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)}$$

for every $y\in\{y\in W(0,T)\cap C(\overline{Q})\mid \frac{\partial y}{\partial t}+Ay\in L^q(Q),\ \frac{\partial y}{\partial n_A}\in L^{\bar\sigma}(\Sigma),\ y(x,0)=0$ on $\overline{\Omega}\}$. Moreover p belongs to $L^{\delta'}(0,T;W^{1,d'}(\Omega))$ for every $\delta>2$, $d>2$ satisfying $\frac{N}{2d}+\frac{1}{\delta}<\frac{1}{2}$ and we have

$$\|p\|_{L^{\delta'}(0,T;W^{1,d'}(\Omega))}\le C_4(\delta,d)\|\mu\|_{\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)},$$

where $C_4(\delta,d)=C_4(T,\Omega,N,C_0,q,\bar\sigma,\delta,d,\|a_i\|_{L^{2q}(Q)},\|b_i\|_{L^{2q}(Q)})$, but C_4 is independent of a and b.

Proof. Due to Theorem 3.2, the proof of Theorem 3.3 follows the lines of the proofs of Theorem 6.3 in [4] and of Theorem 4.2 in [18]. Since we improve the results given in [4, 18], we sketch the main points of the proof. Let $(h_n)_n$ be a sequence in $C_c(Q)$ (the space of continuous functions with compact support in Q), $(k_n)_n$ be a sequence in $C_c(\Sigma)$, and $(\ell_n)_n$ be a sequence in $C(\overline{\Omega})$ such that

$$\|h_n\|_{L^1(Q)}=\|\mu_Q\|_{\mathcal{M}_b(Q)},\quad \|k_n\|_{L^1(\Sigma)}=\|\mu_\Sigma\|_{\mathcal{M}_b(\Sigma)},\ \|\ell_n\|_{L^1(\Omega)}=\|\mu_{\overline{\Omega}_T}\|_{\mathcal{M}(\overline{\Omega}_T)},$$

$$\lim{}_n\int_Q h_n\phi dxdt=\langle\phi,\mu_Q\rangle_{C_b(Q)\times\mathcal{M}_b(Q)},$$

$$\lim{}_n\int_\Sigma k_n\phi dsdt=\langle\phi,\mu_\Sigma\rangle_{C_b(\Sigma)\times\mathcal{M}_b(\Sigma)},$$

$$\lim{}_n\int_\Omega \ell_n\phi dx=\langle\phi,\mu_{\bar\Omega_T}\rangle_{C(\overline{\Omega}_T)\times\mathcal{M}(\overline{\Omega}_T)}$$

for every $\phi\in C(\overline{Q})$. Let $(p_n)_n$ be the sequence in $W(0,T)$ defined by

$$-\frac{\partial p_n}{\partial t}+Ap_n+ap_n=h_n\quad\text{in } Q,\quad \frac{\partial p_n}{\partial n_A}+bp_n=k_n\quad\text{on }\Sigma,\quad p_n(T)=\ell_n\quad\text{in }\Omega.$$

Due to Theorem 3.2, and by using the same arguments as in [4, 18], we can prove that there exists a constant $C_5(\delta,d)=C_5(T,\Omega,N,C_0,q,\bar\sigma,\delta,d,\|a_i\|_{L^{2q}(Q)},\|b_i\|_{L^{2q}(Q)})$ such that

$$\|p_n\|_{L^{\delta'}(0,T;W^{1,d'}(\Omega))}\le C_5(\delta,d)\|\mu\|_{\mathcal{M}_b(\bar Q\setminus\bar\Omega_0)}$$

for every (δ,d) satisfying $\frac{N}{2d}+\frac{1}{\delta}<\frac{1}{2}$. Since $q>\frac{N}{2}+1$ and $\bar\sigma>N+1$, there exist (δ_1,d_1), (δ_2,d_2), (δ_3,d_3) satisfying $\frac{N}{2d_i}+\frac{1}{\delta_i}<\frac{1}{2}$ for $i=1,2,3$, such that $\delta_1'\ge q'$, $d_1'^{*}=\frac{Nd_1'}{N-d_1'}\ge q'$, $\delta_2'\ge\bar\sigma'$, $\frac{(N-1)d_2d_2'}{(N-1)d_2-d_2'}\ge\bar\sigma'$, $\delta_3'\ge(2q)'$, and $d_3'\ge(2q)'$. Therefore

$$\|p_n\|_{L^{q'}(Q)}\le C\|p_n\|_{L^{\delta'_1}(0,T;W^{1,d'_1}(\Omega))}\le CC_5(\delta_1,d_1)\|\mu\|_{\mathcal{M}_b(\bar Q\setminus\bar\Omega_0)},$$

$$\|p_n\|_{L^{\bar\sigma'}(\Sigma)} \le C\|p_n\|_{L^{\delta_2'}(0,T;W^{1,d_2'}(\Omega))} \le CC_5(\delta_2,d_2)\|\mu\|_{\mathcal{M}_b(\bar Q\setminus\bar\Omega_0)},$$

$$\|p_n\|_{L^{(2q)'}(0,T;W^{1,(2q)'}(\Omega))} \le C\|p_n\|_{L^{\delta_3'}(0,T;W^{1,d_3'}(\Omega))} \le CC_5(\delta_3,d_3)\|\mu\|_{\mathcal{M}_b(\bar Q\setminus\bar\Omega_0)}.$$

Then, there exist a subsequence, still indexed by n, and p such that $(p_n)_n$ converges to p for the weak-star topology of $L^{\delta'}(0,T;W^{1,d'}(\Omega))$ for every (δ,d) satisfying $\frac{N}{2d}+\frac{1}{\delta}<\frac{1}{2}$. By passing to the limit in the variational formulation satisfied by $(p_n)_n$, we prove that p is a solution of (3.13). The uniqueness can be proved as in [1, 4]. □

4. Technical results.

4.1. Metric space of controls. To apply the Ekeland variational principle, we have to define a metric space of controls in such a way that the mapping $v \longmapsto y_v$ be continuous from this metric space to $C(\overline{Q})$. Due to Theorem 3.1, this continuity condition will be realized if convergence in the metric space of controls implies convergence in $L^{\bar\sigma}(\Sigma)$. In the case where boundary controls are bounded, convergence in (V_{ad},d) (where d is the so-called Ekeland's distance) implies convergence in $L^{\bar\sigma}(\Sigma)$. This condition is no longer true for unbounded controls; see [10, p. 227]. To overcome this difficulty, we proceed as in [5] and we define a new metric space in the following way. Let $\tilde v$ be in V_{ad}. (In section 5, $\tilde v$ will be an optimal boundary control that we want to characterize.) For $0<M<\infty$, we define the set

$$V_{ad}(\tilde v,M)=\{v\in V_{ad}\mid \|v-\tilde v\|_{\sigma,\Sigma}\le M\}.$$

We endow the set $V_{ad}(\tilde v,M)$ with the Ekeland metric

$$d(v_1,v_2)=\mathcal{L}^N(\{(s,t)\in\Sigma\mid v_1(s,t)\ne v_2(s,t)\}).$$

PROPOSITION 4.1. *Let $\tilde v$ be in V_{ad}. Let $M>0$ and $\{(v_n)_n,v\}\subset V(\tilde v,M)$. If $(v_n)_n$ tends to v in $(V(\tilde v,M),d)$, then $(v_n)_n$ tends to v in $L^{\bar\sigma}(\Sigma)$.*

Proof. Since $1\le\bar\sigma<\sigma$, the proof is immediate if we notice that we have

$$\int_\Sigma |v-v_n|^{\bar\sigma}\,ds \le \|v-v_n\|^{\bar\sigma}_{\sigma,\Sigma}(d(v_n,v))^{\frac{\sigma-\bar\sigma}{\sigma}} \le (2M)^{\bar\sigma}(d(v_n,v))^{\frac{\sigma-\bar\sigma}{\sigma}}. \qquad \square$$

PROPOSITION 4.2. *For every $M>0$, we have that*

(i) *$(V_{ad}(\tilde v,M),d)$ is a complete metric space;*

(ii) *the mapping which associates y_v with v is continuous from $(V_{ad}(\tilde v,M),d)$ into $C(\overline{Q})$;*

(iii) *the mappings $v\to J(y_v,v)$ and $v\to\int_\Sigma \Psi_i(s,t,y_v,v)\,dsdt$ are continuous (respectively, lower semicontinuous) on $(V_{ad}(\tilde v,M),d)$ for $1\le i\le m_0$ (respectively, $m_0+1\le i\le m$).*

Proof. Claims (i) and (ii) are proved in [5], for control problems of elliptic equations; this proof can be repeated here with the obvious modifications. Contrary to [4], [21], the mapping $v\to J(y_v,v)$ is not necessarily continuous on the space of "truncated controls" endowed with the Ekeland metric. We can prove only a lower semicontinuity result. This result is stated in [5, Proposition 3.1] under the additional assumption that $G(s,t,y,\cdot)$ is convex. In fact we can prove the same result without this convexity assumption. Let $(v_n)_n$ be a sequence converging to v in $(V_{ad}(\tilde v,M),d)$. From Proposition 4.1 and Theorem 3.1 we know that $(v_n)_n$ converges to v in $L^{\bar\sigma}(\Sigma)$

and $(y_{v_n})_n$ converges to y_v uniformly on $\overline{Q}$. With assumption (A6), with Fatou's lemma, and with Lebesgue's dominated convergence theorem we have

$$\text{liminf}_n \int_\Sigma G(s,t,0,v_n)\,dsdt \geq \int_\Sigma G(s,t,0,v)\,dsdt,$$

$$\text{lim}_n \int_\Sigma \int_0^1 G'_y(s,t,\theta y_{v_n},v_n) y_{v_n} d\theta\,dsdt = \int_\Sigma \int_0^1 G'_y(s,t,\theta y_v,v) y_v d\theta\,dsdt.$$

Therefore we obtain

$$\text{liminf}_n \int_\Sigma G(s,t,y_{v_n},v_n)\,dsdt = \text{liminf}_n \int_\Sigma G(s,t,0,v_n)\,dsdt$$

$$+\text{lim}_n \int_\Sigma \int_0^1 G'_y(s,t,\theta y_{v_n},v_n) y_{v_n} d\theta\,dsdt$$

$$\geq \int_\Sigma G(s,t,0,v)\,dsdt + \int_\Sigma \int_0^1 G'_y(s,t,\theta y_v,v) y_v d\theta\,dsdt = \int_\Sigma G(s,t,y_v,v)\,dsdt.$$

Following the same ideas, we can prove the continuity (for $1 \leq i \leq m_0$) or the lower semicontinuity (for $m_0 + 1 \leq i \leq m$) of $v \to \int_\Sigma \Psi_i(s,t,y_v,v)\,dsdt$. □

4.2. Existence of diffuse perturbations. Let $\tilde{v}$ be an admissible control, and let v_1 and v_2 be in $V_{ad}(\tilde{v},M)$. A diffuse perturbation of v_1 by v_2 is a family of functions $(v_\rho)_{\rho>0}$ defined by

$$v_\rho(s,t) = \begin{cases} v_1(s,t) & \text{on } \Sigma \setminus E_\rho, \\ v_2(s,t) & \text{on } E_\rho, \end{cases}$$

where E_ρ is a measurable subset of Σ satisfying some conditions. Such perturbations are used to derive Pontryagin's principles from the Ekeland variational principle. In the case of bounded controls (when $V_{ad}(\tilde{v},M) \equiv V_{ad}$) the use of this kind of perturbations goes back to Yao [24, 25] and Li [13] (see also [17, 11, 26, 14]). Some variants have been developed in [4] for bounded controls, and in [21] for unbounded controls. In [5] we have investigated the case of unbounded controls with integral control constraints. Here we prove that the diffuse perturbations defined in [21] may be extended to derive a Pontryagin's principle for problems with integral coupled control-state constraints. Before proving the existence of such diffuse perturbations let us state an auxiliary lemma analogous to Lemma 3.2 of [5].

LEMMA 4.3. *Let ρ be such that $0 < \rho < 1$. For every $v_1, v_2, v_3 \in V_{ad}$, there exists a sequence of measurable sets $(E^n_\rho)_n$ in Σ such that*

$$\mathcal{L}^N(E^n_\rho) = \rho \mathcal{L}^N(\Sigma), \tag{4.1}$$

$$\int_{E^n_\rho} |v_i - v_3|^\sigma\,dsdt = \rho \int_\Sigma |v_i - v_3|^\sigma\,dsdt \quad \textit{for } i = 1,2, \tag{4.2}$$

$$\int_{E^n_\rho} h(s,t,v_i)\,dsdt = \rho \int_\Sigma h(s,t,v_i)\,dsdt \quad \textit{for } i = 1,2, \tag{4.3}$$

$$\frac{1}{\rho}\chi_{E^n_\rho} \rightharpoonup 1 \quad \textit{weakly-star in } L^\infty(\Sigma) \textit{ when } n \textit{ tends to infinity}, \tag{4.4}$$

where $\chi_{E^n_\rho}$ is the characteristic function of E^n_ρ.

Proof. We follow the ideas of [21, Lemma 4.1]. Let us take a sequence $(\varphi_n)_n$ dense in $L^1(\Sigma)$. For $n \geq 1$ we define $f^n \in (L^1(\Sigma))^{n+2\ell+3}$ by

$$f^n = (1, \varphi_1, \ldots, \varphi_n, |v_1 - v_3|^\sigma, |v_2 - v_3|^\sigma, h(\cdot, \cdot, v_1), h(\cdot, \cdot, v_2)).$$

Thanks to Lyapunov's convexity theorem, for every $n \geq 1$ and every $\rho \in (0, 1)$, there exists a measurable subset $E^n_\rho \subset \Sigma$ satisfying

$$\int_{E^n_\rho} f^n \, dsdt = \rho \int_\Sigma f^n \, dsdt.$$

As in [21], it is easy to prove that (4.1)–(4.4) hold for the sequence $(E^n_\rho)_n$. □

THEOREM 4.4. *Let ρ be such that $0 < \rho < 1$. For every $v_1, v_2, v_3 \in V_{ad}$, there exists a measurable subset $E_\rho \subset \Sigma$ such that*

$$\mathcal{L}^N(E_\rho) = \rho \mathcal{L}^N(\Sigma), \tag{4.5}$$

$$\begin{aligned}&\int_{\Sigma \setminus E_\rho} |v_1 - v_3|^\sigma \, dsdt + \int_{E_\rho} |v_2 - v_3|^\sigma \, dsdt \\ &= (1 - \rho) \int_\Sigma |v_1 - v_3|^\sigma \, dsdt + \rho \int_\Sigma |v_2 - v_3|^\sigma \, dsdt,\end{aligned} \tag{4.6}$$

$$\begin{aligned}&\int_{\Sigma \setminus E_\rho} h(s, t, v_1) \, dsdt + \int_{E_\rho} h(s, t, v_2) \, dsdt \\ &= (1 - \rho) \int_\Sigma h(s, t, v_1) \, dsdt + \rho \int_\Sigma h(s, t, v_2) \, dsdt,\end{aligned} \tag{4.7}$$

$$y_\rho = y_1 + \rho z + r_\rho \quad \textit{with} \quad \lim_{\rho \to 0} \frac{1}{\rho} ||r_\rho||_{C(\bar{Q})} = 0, \tag{4.8}$$

$$J(y_\rho, v_\rho) = J(y_1, v_1) + \rho[J'_y(y_1, v_1)z + J(y_1, v_2) - J(y_1, v_1)] + o(\rho), \tag{4.9}$$

$$\begin{aligned}&\int_\Sigma \Psi(s, t, y_\rho, v_\rho) \, dsdt \\ = &\int_\Sigma \Big(\Psi(s, t, y_1, v_1) + \rho[\Psi'_y(s, t, y_1, v_1)z + \Psi(s, t, y_1, v_2) - \Psi(s, t, y_1, v_1)] \Big) \, dsdt + o(\rho),\end{aligned} \tag{4.10}$$

where v_ρ is the control defined by

$$v_\rho(s, t) = \begin{cases} v_1(s, t) & \textit{on } \Sigma \setminus E_\rho, \\ v_2(s, t) & \textit{on } E_\rho, \end{cases} \tag{4.11}$$

y_ρ, y_1 are the solutions of (1.1) *corresponding, respectively, to v_ρ and to v_1, z is the weak solution of*

$$\begin{cases} \dfrac{\partial z}{\partial t} + Az + f'_y(x, t, y_1)z = 0 & \textit{in } Q, \\ \dfrac{\partial z}{\partial n_A} + g'_y(s, t, y_1, v_1)z = g(s, t, y_1, v_1) - g(s, t, y_1, v_2) & \textit{on } \Sigma, \\ z(0) = 0 & \textit{in } \Omega. \end{cases} \tag{4.12}$$

Proof. Using Lemma 4.3, the proof is similar to the one of Theorem 4.1 in [21] and the one of Theorem 3.4 in [4]. The relation (4.10), which does not appear in our previous papers, is deduced with the help of (4.4) and (4.8). □

5. Proof of Pontryagin's principle.

5.1. Penalized problem. Following [15, 16], since $C(\overline{D})$ is separable, there exists a norm $|\cdot|_{C(\overline{D})}$, which is equivalent to the usual norm $\|\cdot\|_{C(\overline{D})}$ such that $(C(\overline{D}), |\cdot|_{C(\overline{D})})$ is strictly convex, and $\mathcal{M}(\overline{D})$, endowed with the dual norm of $|\cdot|_{C(\overline{D})}$ (denoted by $|\cdot|_{\mathcal{M}(\overline{D})}$), is also strictly convex; see [7, Corollary 2, p. 148 or Corollary 2, p. 167]. We define the distance function to $\mathcal{C}$ (for the new norm $|\cdot|_{C(\overline{D})}$) by

$$d_{\mathcal{C}}(\varphi) = \inf_{z \in \mathcal{C}} |\varphi - z|_{C(\overline{D})}.$$

Since $\mathcal{C}$ is convex, then $d_{\mathcal{C}}$ is convex and Lipschitz of rank 1, and we have

$$\limsup_{\substack{\rho \searrow 0, \\ \varphi' \to \varphi}} \frac{d_{\mathcal{C}}(\varphi' + \rho z) - d_{\mathcal{C}}(\varphi')}{\rho} = \max\{\langle \xi, z \rangle_{\mathcal{M}(\overline{D}) \times C(\overline{D})} \mid \xi \in \partial d_{\mathcal{C}}(\varphi)\} \tag{5.1}$$

for every $\varphi, z \in C(\overline{D})$, where $\partial d_{\mathcal{C}}$ is the subdifferential in the sense of convex analysis (see [6]). Therefore, for a given $\varphi \in C(\overline{D})$ we have

(5.2)

$$\langle \xi, z - \varphi \rangle_{\mathcal{M}(\overline{D}) \times C(\overline{D})} + d_{\mathcal{C}}(\varphi) \le d_{\mathcal{C}}(z) \quad \text{for all } \xi \in \partial d_{\mathcal{C}}(\varphi) \text{ and for all } z \in C(\overline{D}),$$
$$|\xi|_{\mathcal{M}(\overline{D})} \le 1 \quad \text{for every } \xi \in \partial d_{\mathcal{C}}(\varphi).$$

Moreover it is proved in [16, Lemma 3.4] that, since $\mathcal{C}$ is a closed convex subset of $C(\overline{D})$, for every $\varphi \notin \mathcal{C}$, and every $\xi \in \partial d_{\mathcal{C}}(\varphi)$, then $|\xi|_{\mathcal{M}(\overline{D})} = 1$. Since $\partial d_{\mathcal{C}}(\varphi)$ is convex in $\mathcal{M}(\overline{D})$ and $(\mathcal{M}(\overline{D}), |\cdot|_{\mathcal{M}(\overline{D})})$ is strictly convex, if $\varphi \notin \mathcal{C}$, then $\partial d_{\mathcal{C}}(\varphi)$ is a singleton and $d_{\mathcal{C}}$ is Gâteaux-differentiable at φ.

Let $(\bar{y}, \bar{v})$ be an optimal solution of (P). Consider the penalized functional

$$J_k(y,v) = \left\{ \left[\left(J(y,v) - J(\bar{y},\bar{v}) + \frac{1}{k^2} \right)^+ \right]^2 + (d_{\mathcal{C}}(\Phi(y)))^2 \right.$$
$$\left. + \sum_{i=1}^{m_0} \left[\int_\Sigma \Psi_i(s,t,y,v)\, dsdt \right]^2 + \sum_{i=m_0+1}^{m} \left[\left(\int_\Sigma \Psi_i(s,t,y,v)\, dsdt \right)^+ \right]^2 \right\}^{\frac{1}{2}}.$$

We easily verify that $(\bar{y}, \bar{v})$ is a $\frac{1}{k^2}$-solution of the penalized problem

$$(\mathrm{P}_k^M) \qquad \inf\{J_k(y,v) \mid y \in W(0,T) \cap C(\overline{Q}), v \in V_{ad}(\bar{v}, M),\ (y,v) \text{ satisfies } (1.1)\}$$

for every $M > 0$ and every $k > 0$. For every $k > 0$, we set $M_k = k^{(\frac{1}{2\bar{\sigma}} - \frac{1}{2\sigma})}$ and we denote by (P^k) the penalized problem $(\mathrm{P}_k^{M_k})$.

5.2. Proof of Theorem 2.1. Step 1. For every $k \ge 1$, the metric space $(V_{ad}(\bar{v}, M_k), d)$ is complete; see Proposition 4.2. Let us prove that the functional $v \longmapsto J_k(y_v, v)$ is lower semicontinuous on this metric space. Since the mappings

$v \to J(y_v, v)$ and $v \to \int_\Sigma \Psi_i(s,t,y_v,v)\,dsdt$ ($m_0+1 \le i \le m$) are lower semicontinuous on $(V_{ad}(\tilde v, M_k), d)$, it is clear that $v \to \big(J(y_v,v) - J(\bar y, \bar v) + \frac{1}{k^2}\big)^+$ and $v \to \big(\int_\Sigma \Psi_i(s,t,y_v,v)\,dsdt\big)^+$ ($m_0+1 \le i \le m$) are also lower semicontinuous on $(V_{ad}(\tilde v, M_k), d)$ because $r \to r^+$ is a nondecreasing continuous mapping from $\mathbb{R}$ into $\mathbb{R}^+$. On the other hand, the mappings $v \to \int_\Sigma \Psi_i(s,t,y_v,v)\,dsdt$ ($1 \le i \le m_0$) are continuous on $(V_{ad}(\tilde v, M_k), d)$. Since the mappings $r \to r^2$ and $r \to r^{\frac12}$ are nondecreasing and continuous from $\mathbb{R}^+$ into $\mathbb{R}^+$, then $v \to J_k(y_v, v)$ is lower semicontinuous. Due to Ekeland's variational principle, for every $k \ge 1$, there exists $v_k \in V_{ad}(\bar v, M_k)$ such that

$$(5.3)\quad d(v_k, \bar v) \le \frac{1}{k} \text{ and } J_k(y_k, v_k) \le J_k(y_v, v) + \frac{1}{k} d(v_k, v) \text{ for every } v \in V_{ad}(\bar v, M_k).$$

(y_k and y_v are the solutions of (1.1) corresponding, respectively, to v_k and v.) Let v_0 be in V_{ad}. Let k_0 be large enough so that v_0 belong to $V_{ad}(\bar v, M_k)$ for every $k \ge k_0$. Observe that, for the above choice of M_k, $(v_k)_k$ tends to $\bar v$ in $L^{\bar\sigma}(\Sigma)$. Let us check this. Denote by Σ_k the set of points $(s,t) \in \Sigma$ where $v_k(s,t) \ne \bar v(s,t)$. From (5.3) we know that $\mathcal{L}^N(\Sigma_k) \le 1/k$. Then

$$(5.4)\qquad \begin{aligned} \int_\Sigma |\bar v - v_k|^{\bar\sigma} dsdt = \int_{\Sigma_k} |\bar v - v_k|^{\bar\sigma} dsdt &\le \|\bar v - v_k\|^{\bar\sigma}_{\sigma,\Sigma} \mathcal{L}^N(\Sigma_k)^{1-\frac{\bar\sigma}{\sigma}} \\ &\le M_k^{\bar\sigma} k^{\frac{\bar\sigma}{\sigma}-1} = k^{\frac12(\frac{\bar\sigma}{\sigma}-1)} \longrightarrow 0 \text{ when } k \to +\infty. \end{aligned}$$

Step 2. Theorem 3.1 gives the existence of measurable sets $E^k_\rho \subset \Sigma$, such that $\mathcal{L}^N(E^k_\rho) = \rho \mathcal{L}^N(\Sigma)$,

$$(5.5)\qquad \begin{aligned} &\int_{\Sigma \setminus E^k_\rho} |v_k - \bar v|^\sigma\, dsdt + \int_{E^k_\rho} |v_0 - \bar v|^\sigma\, dsdt \\ &\quad = (1-\rho)\int_\Sigma |v_k - \bar v|^\sigma\, dsdt + \rho \int_\Sigma |v_0 - \bar v|^\sigma\, dsdt, \end{aligned}$$

$$(5.6)\qquad \begin{aligned} &\int_{\Sigma \setminus E^k_\rho} h(s,t,v_k)\, dsdt + \int_{E^k_\rho} h(s,t,v_0)\, dsdt \\ &\quad = (1-\rho)\int_\Sigma h(s,t,v_k)\, dsdt + \rho \int_\Sigma h(s,t,v_0)\, dsdt, \end{aligned}$$

$$(5.7)\qquad \begin{aligned} &\int_\Sigma (\Psi(s,t,y^k_\rho, v^k_\rho) - \Psi(s,t,y_k,v_k))\, dsdt \\ &\quad = \rho \int_\Sigma (\Psi'_y(s,t,y_k,v_k) z_k + \Psi(s,t,y_k,v_0) - \Psi(s,t,y_k,v_k))\, dsdt + o(\rho), \end{aligned}$$

$$(5.8)\qquad y^k_\rho = y_k + \rho z_k + r^k_\rho, \quad \lim_{\rho \to 0} \frac{1}{\rho} \|r^k_\rho\|_{C(\overline{Q})} = 0,$$

$$(5.9)\qquad J(y^k_\rho, v^k_\rho) = J(y_k, v_k) + \rho \Delta J_k + o(\rho),$$

where v^k_ρ is defined by

$$(5.10)\qquad v^k_\rho(s,t) = \begin{cases} v_k(s,t) & \text{on } \Sigma \setminus E^k_\rho, \\ v_0(s,t) & \text{on } E^k_\rho, \end{cases}$$

y_ρ^k is the state corresponding to v_ρ^k, z_k is the weak solution of

$$\begin{cases} \dfrac{\partial z_k}{\partial t} + Az_k + f'_y(x,t,y_k)z_k = 0 & \text{in } Q, \\[2ex] \dfrac{\partial z_k}{\partial n_A} + g'_y(s,t,y_k,v_k)z_k = g(s,t,y_k,v_k) - g(s,t,y_k,v_0) & \text{on } \Sigma, \\[2ex] z_k(0) = 0 & \text{in } \Omega, \end{cases}$$

and

$$\Delta J_k = \int_Q F'_y(x,t,y_k(x,t))z_k(x,t)\,dxdt + \int_\Sigma G'_y(s,t,y_k(s,t),v_k(s,t))z_k(s,t)\,dsdt$$

$$+ \int_\Sigma [G(s,t,y_k(s,t),v_0(s,t)) - G(s,t,y_k(s,t),v_k(s,t))]\,dsdt + \int_\Omega L'_y(x,y_k(T))z_k(T)\,dx.$$

On the other hand, for every $k > k_0$ and every $0 < \rho < 1$, due to (5.5) and (5.6), v_ρ^k belongs to $V_{ad}(\bar{v}, M_k)$. If we set $v = v_\rho^k$ in (5.3), it follows that

$$\lim_{\rho\to 0} \frac{J_k(y_k,v_k) - J_k(y_\rho^k, v_\rho^k)}{\rho} \le \frac{1}{k}\mathcal{L}^N(\Sigma). \tag{5.11}$$

Taking (5.1), (5.7), (5.9), and the definition of J_k into account, we obtain

$$\begin{gathered} -\nu_k \Delta J_k \lambda_k \int_\Sigma \left[\Psi(\cdot,y_k,v_0) - \Psi(\cdot,y_k,v_k) + \Psi'_y(\cdot,y_k,v_k)z_k\right]\,dsdt \\ -\langle \mu_k, \Phi'(y_k)z_k\rangle_{\overline{D}} \le \frac{1}{k}\mathcal{L}^N(\Sigma), \end{gathered} \tag{5.12}$$

where

$$\lambda_k^i = \frac{\int_\Sigma \Psi_i(s,t,y_k,v_k)\,dsdt}{J_k(y_k,v_k)} \quad \text{for } 1 \le i \le m_0,$$

$$\lambda_k^i = \frac{\left(\int_\Sigma \Psi_i(s,t,y_k,v_k)\,dsdt\right)^+}{J_k(y_k,v_k)} \quad \text{for } m_0+1 \le i \le m,$$

$$\nu_k = \frac{(J(y_k,v_k) - J(\bar{y},\bar{v}) + \frac{1}{k^2})^+}{J_k(y_k,v_k)}, \quad \mu_k = \begin{cases} \dfrac{d_{\mathcal{C}}(\Phi(y_k))\nabla d_{\mathcal{C}}(\Phi(y_k))}{J_k(y_k,v_k)} & \text{if } \Phi(y_k) \notin \mathcal{C}, \\ 0 & \text{otherwise.} \end{cases}$$

For every $k > 0$, we consider the weak solution p_k of

$$\begin{cases} -\dfrac{\partial p_k}{\partial t} + A^* p_k + f'_y(x,t,y_k)p_k = \nu_k F'_y(x,t,y_k) + [\Phi'(y_k)^*\mu_k]|_Q, \\[2ex] \dfrac{\partial p_k}{\partial n_{A^*}} + g'_y(\cdot,y_k,v_k)p_k = \nu_k G'_y(\cdot,y_k,v_k) + \lambda_k \Psi'_y(\cdot,y_k,v_k) + [\Phi'(y_k)^*\mu_k]|_\Sigma, \\[2ex] p_k(T) = \nu_k L'_y(x,y_k(T)) + [\Phi'(y_k)^*\mu_k]|_{\overline{\Omega}_T}, \end{cases} \tag{5.13}$$

where $[\Phi'(y_k)^*\mu_k]|_Q$, $[\Phi'(y_k)^*\mu_k]|_\Sigma$, and $[\Phi'(y_k)^*\mu_k]|_{\bar\Omega_T}$ have the same meaning as in Theorem 2.1. By using the Green formula of Theorem 3.4, we obtain

$$\nu_k \int_Q F'_y(x,t,y_k)z_k\,dxdt + \nu_k \int_\Sigma G'_y(s,t,y_k,v_k)z_k\,dsdt + \nu_k \int_\Omega L'_y(x,y_k(T))z_k(T)\,dx$$

$$+\lambda_k \int_\Sigma \Psi'_y(s,t,y_k,v_k)z_k\,dsdt + \langle \mu_k, \Phi'(y_k)z_k\rangle_{\overline{D}}$$

$$= \int_Q p_k\left(\frac{\partial z_k}{\partial t} + Az_k + f'_y(x,t,y_k)z_k\right)dxdt + \int_\Sigma p_k\left(\frac{\partial z_k}{\partial n_A} + g'_y(s,t,y_k,v_k)z_k\right)dsdt$$

$$= \int_\Sigma p_k[g(s,t,y_k,v_k) - g(s,t,y_k,v_0)]\,dsdt.$$

With this equality, (5.12), and the definition of ΔJ_k, we have

$$(5.14)\qquad \int_\Sigma [\nu_k G(s,t,y_k,v_k) + \lambda_k\Psi(s,t,y_k,v_k) - p_k g(s,t,y_k,v_k)]\,dsdt$$

$$\leq \int_\Sigma [\nu_k G(s,t,y_k,v_0) + \lambda_k\Psi(s,t,y_k,v_0) - p_k g(s,t,y_k,v_0)]\,dsdt + \frac{1}{k}\mathcal{L}^N(\Sigma)$$

for every $k \geq k_0$.

Step 3. Notice that $\nu_k^2 + \sum_i(\lambda_k^i)^2 + |\mu_k|^2_{\mathcal{M}(\overline{D})} = 1$. Then there exist an element $(\bar\nu,\bar\lambda,\bar\mu)$ in $\mathbb{R}^{1+m}\times\mathcal{M}(\overline{D})$ with $\bar\nu \geq 0$ and $\bar\lambda_i \geq 0$ for $m_0+1 \leq i \leq m$, and a subsequence, still denoted by $(\nu_k,\lambda_k,\mu_k)_k$, such that

$$(\nu_k,\lambda_k) \longrightarrow (\bar\nu,\bar\lambda) \text{ in } \mathbb{R}^{1+m},\ \mu_k \rightharpoonup \bar\mu \text{ weak-star in } \mathcal{M}(\overline{D}).$$

From Theorem 3.4, we obtain the estimate

$$\|p_k\|_{L^{\delta'}(0,T;W^{1,d'}(\Omega))} \leq C_4(\delta,d)\Big\{\|F'_y(\cdot,y_k)\|_{1,Q} + \|G'_y(\cdot,y_k,v_k)\|_{1,\Sigma} +$$

$$\|L'_y(\cdot,y_k(T))\|_{1,\Omega} + |\lambda_k|\|\Psi'_y(.,y_k,v_k)\|_{1,\Sigma} + |\mu_k|_{\mathcal{M}(\overline{D})}\|\Phi'_y(y_k)\|_{\mathcal{L}(C(\overline{D});C(\overline{D}))}\Big\}$$

for every (δ,d) satisfying $\frac{N}{2d}+\frac{1}{\delta}<\frac{1}{2}$, where $\mathcal{L}(C(\overline{D});C(\overline{D}))$ denotes the space of linear continuous mappings from $C(\overline{D})$ to $C(\overline{D})$.

Since the sequences $(\nu_k)_k, (\lambda_k)_k, (\mu_k)_k, (y_k)_k$, and $(v_k)_k$ are bounded, respectively, in $\mathbb{R}$, $\mathbb{R}^m$, $\mathcal{M}(\overline{D})$, $C(\overline{Q})$, and in $L^{\bar\sigma}(\Sigma)$, the sequence $(p_k)_k$ is bounded in $L^{\delta'}(0,T;W^{1,d'}(\Omega))$. Then there exist $\bar p \in L^{\delta'}(0,T;W^{1,d'}(\Omega))$ and a subsequence, still denoted by $(p_k)_k$, such that $(p_k)_k$ weakly converges to $\bar p$ in $L^{\delta'}(0,T;W^{1,d'}(\Omega))$ for every (δ,d) satisfying $\frac{N}{2d}+\frac{1}{\delta}<\frac{1}{2}$. By using the same arguments as in [21], we can prove that $\bar p$ is the weak solution of (2.4).

Step 4. Recall that $(v_k)_k$ tends to $\bar v$ in $L^{\bar\sigma}(\Sigma)$ (see (5.4)).

By passing to the limit when k tends to infinity in (5.14), with Fatou's lemma (applied to the sequence of functions $(\nu_k G(\cdot,0,v_k(\cdot)), \lambda_k\Psi(\cdot,0,v_k(\cdot)))_k$ and the convergence results stated in Step 2, we obtain

$$(5.15)\qquad H_\Sigma(\bar y,\bar v,\bar p,\bar\nu,\bar\lambda) \leq H_\Sigma(\bar y,v_0,\bar p,\bar\nu,\bar\lambda),$$

for every $v_0 \in V_{ad}$. On the other hand, from definitions of μ_k and λ_k, and from (5.2), we deduce

$$\lambda_k^i \int_\Sigma \Psi_i(s,t,y_k,v_k)\,dsdt = 0,\ m_0+1 \le i \le m,$$

$$\langle \mu_k, z - \Phi(y_k)\rangle_{\mathcal{M}(\overline{D})\times C(\overline{D})} \le 0 \ \text{ for all } z \in \mathcal{C}.$$

We obtain (2.2) and (2.3) by passing to the limit in these expressions. Since $\mathcal{C}$ is of finite codimension, by using the same arguments as in [22], we prove that $(\bar\nu, \bar\lambda, \bar\mu)$ is nonzero. □

REFERENCES

[1] J. J. Alibert and J. P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim., 18 (1997), pp. 235–250.

[2] N. Basile and M. Mininni, *An extension of the maximum principle for a class of optimal control problems in infinite-dimensional spaces*, SIAM J. Control Optim., 28 (1990), pp. 1113–1135.

[3] J. F. Bonnans and E. Casas, *An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 274–298.

[4] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[5] E. Casas, J.-P. Raymond, and H. Zidani, *Optimal control problem governed by semilinear elliptic equations with integral control constraints and pointwise state constraints*, in International Conference on Control and Estimations of Distributed Parameter Systems, Vorau, Austria, 1996, W. Desch, F. Kappel, K. Kunisch, eds., Birkhaüser-Verlag, Basel, 1998, pp. 89–102.

[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, Toronto, 1983.

[7] J. Diestel, *Geometry of Banach Spaces: Selected Topics*, Lecture Notes in Math. 485, Springer-Verlag, Berlin, Heidelberg, New York, 1975.

[8] H. O. Fattorini, *Infinite Dimensional Optimization and Control Theory*, Cambridge University Press, Cambridge, UK, 1999.

[9] H. O. Fattorini and T. Murphy, *Optimal problems for nonlinear parabolic boundary control systems*, SIAM J. Control Optim., 32 (1994), pp. 1577–1596.

[10] H. O. Fattorini and S. Sritharan, *Necessary and sufficient conditions for optimal controls in viscous flow problems*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 211–251.

[11] B. Hu and J. Yong, *Pontryagin maximum principle for semilinear and quasilinear parabolic equations with pointwise state constraints*, SIAM J. Control Optim., 33 (1995), pp. 1857–1880.

[12] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Math. Monogr. 23, Amer. Math. Soc., Providence, RI, 1968.

[13] X. J. Li, *Vector measure and the necessary conditions for the optimal control problems of linear systems*, in Proceedings of the Third IFAC Symposium on the Control of Distributed Parameter Systems, Toulouse, France, Pergamon, Oxford, UK, 1982.

[14] X. J. Li and S. N. Chow, *Maximum principle of optimal control for functional differential systems*, J. Optim. Theory Appl., 54 (1987), pp. 335–360.

[15] X. J. Li and Y. Yao, *Maximum principle of distributed parameter systems with time lags*, in Proceedings on the Conference on Control Theory of Distributed Parameter Systems and Applications, F. Kappel and K. Kunish, eds., Springer-Verlag, New York, 1985, pp. 410–427.

[16] X. J. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.

[17] X. J. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, Basel, Berlin, 1995.

[18] J. P. Raymond, *Nonlinear boundary control of semilinear parabolic equations with pointwise state constraints*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 341–370.

[19] J. P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim., 39 (1999), pp. 143–177.

[20] J. P. Raymond and H. Zidani, *Optimal control problem governed by a semilinear parabolic equation*, in System Modelling and Optimization, J. Dolezal and J. Fidler, eds., Chapman and Hall, London, 1996, pp. 211–217.

[21] J. P. Raymond and H. Zidani, *Pontryagin's principles for state-constrained control problems governed by semilinear parabolic equations with unbounded controls*, SIAM J. Control Optim., 36 (1998), pp. 1853–1879.

[22] J. P. Raymond and H. Zidani, *Pontryagin's principle for time optimal problems*, J. Optim. Theory Appl., 101 (1999), pp. 375–402.

[23] J. Serrin, *Pathological solutions of elliptic differential equations*, Ann. Scuola Norm. Sup. Pisa, 18 (1964), pp. 385–387.

[24] Y. Yao, *Vector measure and maximum principle of distributed parameter systems*, Sci. Sinica Ser. A, 26 (1983), pp. 102–112.

[25] Y. Yao, *Maximum principle of semi-linear distributed systems*, in Proceedings of the Third IFAC Symposium on the Control of Distributed Parameter Systems, Toulouse, France, Pergamon, Oxford, UK, 1982.

[26] J. Yong, *Pontryagin maximum principle for semilinear second order elliptic partial differential equations and variational inequalities with state constraints*, Differential Integral Equations, 5 (1992), pp. 1307–1334.

SIAM J. Control Optim.
Vol. 39, No. 4, pp. 1204–1232

OPTIMIZABILITY AND ESTIMATABILITY FOR INFINITE-DIMENSIONAL LINEAR SYSTEMS[*]

GEORGE WEISS[†] AND RICHARD REBARBER[‡]

Abstract. An infinite-dimensional linear system described by $\dot{x}(t) = Ax(t) + Bu(t)$ $(t \geq 0)$ is said to be optimizable if for every initial state $x(0)$, an input $u \in L^2$ can be found such that $x \in L^2$. Here, A is the generator of a strongly continuous semigroup on a Hilbert space and B is an admissible control operator for this semigroup. In this paper we investigate optimizability (also known as the finite cost condition) and its dual, estimatability. We explore the connections with stabilizability and detectability. We give a very general theorem about the equivalence of input-output stability and exponential stability of well-posed linear systems: the two are equivalent if the system is optimizable and estimatable. We conclude that a well-posed system is exponentially stable if and only if it is dynamically stabilizable and input-output stable. We illustrate the theory by two examples based on PDEs in two or more space dimensions: the wave equation and a structural acoustics model.

Key words. strongly continuous semigroup, well-posed linear system, regular linear system, stabilizability, optimizability, detectability, input-output stability, dynamic stabilization

AMS subject classifications. 93B05, 93B52, 93D25, 93D22

PII. S036301299833519X

1. Introduction and outline of the results.

This paper is mainly about linear infinite-dimensional systems described either by

$$\left\{ \begin{array}{llll} \dot{x}(t) & = Ax(t) + Bu(t), & & x(0) = x_0, \\ y(t) & = x(t), & & t \geq 0, \end{array} \right. \tag{1.1}$$

or (in the dual situation) by

$$\left\{ \begin{array}{llll} \dot{x}(t) & = Ax(t) + u(t), & & x(-\infty) = 0, \\ y(t) & = C_\Lambda x(t), & & t \leq 0. \end{array} \right. \tag{1.2}$$

Here, A is the generator of a strongly continuous semigroup of operators $\mathbb{T}$ on the state space X, B is an admissible control operator for $\mathbb{T}$, defined on the input space U, and $C : \mathcal{D}(A) \rightarrow Y$ is an admissible observation operator for $\mathbb{T}$. U, X, and Y are Hilbert spaces and the operator C_Λ is the Λ-extension of C:

$$C_\Lambda z = \lim_{\lambda \rightarrow +\infty} C\lambda(\lambda I - A)^{-1} z \tag{1.3}$$

for all $z \in X$ for which the limit exists. In both types of systems, u is the input function, x is the state trajectory, and y is the output function. The first system is determined by the pair (A, B) and the second by the pair (A, C).

The pair (A, B) is called *optimizable* if for every $x_0 \in X$, a function u in L^2 can be found such that x is in L^2. If B is bounded, then optimizability is equivalent to

[*] Received by the editors March 9, 1998; accepted for publication (in revised form) August 14, 2000; published electronically December 7, 2000. This work was partially supported by National Science Foundation grant DMS-9623392.

http://www.siam.org/journals/sicon/39-4/33519.html

[†] Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, London SW7 2BT, UK (G.Weiss@ic.ac.uk).

[‡] Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588-0323 (rrebarbe@math.unl.edu).

stabilizability via a bounded state feedback operator: $u(t) = Fy(t)$ in (1.1); see [12]. For unbounded B, stabilizability implies optimizability, but we do not know if the converse holds. Here, by stabilizability we mean a rather general concept, which will be recalled in section 2 and which involves an unbounded state feedback operator. Obviously, exact controllability implies optimizability.

Estimatability is (by definition) the dual of optimizability: the pair (A, C) is *estimatable* if (A^*, C^*) is optimizable. If C is bounded, then estimatability is equivalent to detectability via a bounded output injection operator: $u(t) = Hy(t)$ in (1.2). For unbounded C, detectability implies estimatability, but we do not know if the converse holds. For the precise meaning of detectability which we have in mind, we refer to section 2. Obviously, exact observability implies estimatability.

In this paper we investigate the concepts of optimizability and estimatability. Optimizability (sometimes called the finite cost condition) has received some attention in recent years; see, for example, Flandoli, Lasiecka, and Triggiani [13], Jacob and Zwart [16], Rebarber and Zwart [23] and the references therein. As far as we know, estimatability has not yet been considered. Of course, every statement about optimizability can be translated into a dual statement about estimatability, but there is more to estimatability than just this. For example, it is interesting to derive the duality-free definition of estimatability, namely, that a final state estimator exists for the system (1.2). Also, certain estimates appear to be natural in the context of estimatability, but not in the dual context. It will be interesting to look at well-posed linear systems which are both optimizable and estimatable.

We now outline the contents of the sections and state some of the results. In section 2 we recall the necessary background on admissible control and observation operators, well-posed linear systems, stabilizability, and detectability.

In section 3 we study the concept of optimizability. We are concerned with systems described by (1.1). We review some LQ optimal control theory results for optimizable systems and derive some simple consequences. We derive a Hautus test for optimizability. We introduce the *open-loop L^2-stabilization problem,* which is to find a bounded operator $\mathbf{F}$ from X to $L^2([0, \infty), U)$ such that, taking in (1.1) $u = \mathbf{F}x_0$, the state trajectory x is in $L^2([0, \infty), X)$ and depends continuously on x_0. As is easy to guess, it turns out that this problem (and its optimal version, also defined in section 3) is solvable if and only if (A, B) is optimizable.

In section 4 we study the concept of estimatability. We show that estimatability of (A, C) is equivalent to the solvability of the associated *final state estimation problem.* This concerns systems described by (1.2). The function u is in $L^2((-\infty, 0], X)$, it has compact support, and $x(t) = 0$ for large negative t. The problem is to estimate $x(0)$, based on the knowledge of y, such that the estimation error depends continuously on u (with its L^2-norm). This is difficult because, unless $\mathbb{T}$ is exponentially stable, $x(0)$ and y do not depend continuously on u (with its L^2-norm). Indeed, if the system is unstable and the support of u is far from zero (i.e., it is in the "distant past"), then its influence on $x(0)$ and on the "recent past" of y can be very large.

A stronger version of this problem is the *optimal final state estimation problem*, which turns out to be the dual of the optimal open-loop L^2-stabilization problem introduced in section 3. Based on this duality, we derive that the estimatability of (A, C) implies the existence of a $K > 0$ such that

$$\int_0^T \|\mathbb{T}_t x_0\|^2 dt \le K^2 \left(\|x_0\|^2 + \int_0^T \|C\mathbb{T}_t x_0\|^2 dt \right)$$

for all $x_0 \in \mathcal{D}(A)$ and for all $T \geq 0$. We do not know if this estimate is sufficient for estimatability. Another consequence of estimatability is the following: there exist $\delta > 0$ and $m > 0$ such that for all $s \in \mathbb{C}$ with Re $s > -\delta$

$$\|(sI - A)z\| + \|Cz\| \geq m\|z\| \qquad \forall\, z \in \mathcal{D}(A).$$

Again, we do not know if the converse holds.

In section 5 we investigate the relation between two stability concepts. A well-posed linear system with input space U and output space Y is called *input-output stable* if it maps inputs in $L^2([0,\infty), U)$ into outputs in $L^2([0,\infty), Y)$. Here, the initial state is considered to be zero. Input-output stability is equivalent to the fact that the transfer function $\mathbf{G}$ of the system is an $\mathcal{L}(U,Y)$-valued H^∞ function (i.e., $\mathbf{G}$ is bounded on the open right half-plane). A well-posed linear system is called *exponentially stable* if its semigroup is exponentially stable.

It is well known that exponential stability implies input-output stability (see, e.g., [36]), but the converse is not true (not even for finite-dimensional systems). There has been much interest in recent years in finding additional conditions on the system which imply that these two kinds of stability are equivalent (we shall give a short account of this). The main result of section 7 is a generalization of all the results known to us in this direction. Our result is the following.

THEOREM 1.1. *A well-posed linear system is exponentially stable if and only if it is optimizable, estimatable, and input-output stable.*

In section 6 we show that optimizability and estimatability are preserved under output feedback. This is needed in the study of the dynamic stabilizability of well-posed linear systems. Such a system is called *dynamically stabilizable* if it can be embedded as a subsystem of a larger, exponentially stable well-posed linear system. The "other half" of this larger system is a so-called *stabilizing controller with internal loop* (see section 6 for the precise definition). Our definition of dynamic stabilization includes stabilization by static feedback, as well as stabilization by a controller in the usual sense, as discussed, for example, in Chapter 4 of Curtain and Zwart [12] (where the control and observation operators are assumed to be bounded, which simplifies the analysis). Sufficient conditions for dynamic stabilizability in our general sense (and a design procedure) were given in Weiss and Curtain [40], in the context of regular linear systems. (Regular linear systems are a large subclass of well-posed linear systems; see section 2.) The main result of section 6 is the following.

THEOREM 1.2. *A well-posed linear system is dynamically stabilized by a controller with internal loop if and only if* (1) *both subsystems are optimizable and estimatable,* (2) *the closed-loop system is input-output stable.*

This, combined with Theorem 1.1, implies the following useful corollary.

COROLLARY 1.3. *A well-posed linear system is exponentially stable if and only if it is dynamically stabilizable and input-output stable.*

We give examples of transfer functions which cannot be associated to dynamically stabilizable systems. For example, if $\mathbf{G}$ is a Blaschke product with zeros at $\frac{1}{n} \pm in$, then no system with transfer function $\mathbf{G}$ can be dynamically stabilizable.

In section 7 we give two examples based on partial differential equations (PDEs). One concerns the n-dimensional wave equation with mixed boundary control and boundary observation on a subset of the boundary, and the other is a structural acoustic system (the two-dimensional wave equation interacting with a beam equation). Both examples use the last corollary to conclude that the system is not dynamically stabilizable. Since the proof of Corollary 1.3 relies on a majority of the other results derived in this paper, these results about PDE systems rely on a complex chain of abstract systems-theoretical results.

2. Some background on infinite-dimensional systems. In this section we gather, for easy reference, some basic facts about admissible control and observation operators, about well-posed and regular linear systems, and about stabilizability and detectability. However, the material cannot be learned from this section, and for details we refer to the literature.

Throughout this section, X is a Hilbert space and $A : \mathcal{D}(A) \to X$ is the generator of a strongly continuous semigroup $\mathbb{T}$ on X. The Hilbert space X_1 is $\mathcal{D}(A)$ with the norm $\|z\|_1 = \|(\beta I - A)z\|$, where $\beta \in \rho(A)$ is fixed (this norm is equivalent to the graph norm). The Hilbert space X_{-1} is the completion of X with respect to the norm $\|z\|_{-1} = \|(\beta I - A)^{-1} z\|$. This space is isomorphic to $\mathcal{D}(A^*)^*$, and we have

$$X_1 \subset X \subset X_{-1}$$

densely and with continuous embeddings. $\mathbb{T}$ extends to a semigroup on X_{-1}, denoted by the same symbol. The generator of this extended semigroup is an extension of A, whose domain is X, so that $A : X \to X_{-1}$.

U is a Hilbert space and $B \in \mathcal{L}(U, X_{-1})$ is an *admissible control operator* for $\mathbb{T}$, defined as in Weiss [33]. This means that if x is the solution of

$$\dot{x}(t) = Ax(t) + Bu(t)\,, \tag{2.1}$$

with $x(0) = x_0 \in X$ and $u \in L^2([0,\infty), U)$, then $x(t) \in X$ for all $t \geq 0$. In this case, x is a continuous X-valued function of t. We have $x(t) = \mathbb{T}_t x_0 + \Phi_t u$, where $\Phi_t \in \mathcal{L}(L^2([0,\infty), U), X)$ is defined by

$$\Phi_t u = \int_0^t \mathbb{T}_{t-\sigma} B u(\sigma)\, d\sigma\,. \tag{2.2}$$

The above integration is done in X_{-1}, but the result is in X. The Laplace transform of x is

$$\hat{x}(s) = (sI - A)^{-1} \left[x_0 + B\hat{u}(s) \right] .$$

B is called *bounded* if $B \in \mathcal{L}(U, X)$ (and unbounded otherwise).

Y is another Hilbert space and $C \in \mathcal{L}(X_1, Y)$ is an *admissible observation operator* for $\mathbb{T}$, defined as in Weiss [35]. This means that for every $T > 0$ there exists a $K_T \geq 0$ such that

$$\int_0^T \|C\mathbb{T}_t x_0\|^2 dt \leq K_T^2 \|x_0\|^2 \qquad \forall\, x_0 \in \mathcal{D}(A)\,. \tag{2.3}$$

C is called *bounded* if it can be extended such that $C \in \mathcal{L}(X, Y)$.

We regard $L^2_{loc}([0,\infty), Y)$ as a Fréchet space with the seminorms being the L^2-norms on the intervals $[0,n]$, $n \in \mathbb{N}$. Then the admissibility of C means that there is a continuous operator $\Psi : X \to L^2_{loc}([0,\infty), Y)$ such that

$$(\Psi x_0)(t) = C\mathbb{T}_t x_0 \qquad \forall\, x_0 \in \mathcal{D}(A)\,. \tag{2.4}$$

The operator Ψ is completely determined by (2.4), because $\mathcal{D}(A)$ is dense in X. However, replacing C by its Λ-extension C_Λ, defined in (1.3), formula (2.4) becomes true for all $x_0 \in X$ and for almost every $t \geq 0$. If $y = \Psi x_0$, then its Laplace transform is

$$\hat{y}(s) = C(sI - A)^{-1} x_0 . \tag{2.5}$$

The operator Ψ is an *extended output map* for $\mathbb{T}$, which means that

$$\mathbf{S}_\tau^* \Psi = \Psi \mathbb{T}_\tau \qquad \forall\, \tau \geq 0 , \tag{2.6}$$

where $\mathbf{S}_\tau^*$ denotes the left shift by τ on $L^2_{loc}([0,\infty), Y)$. There is a representation theorem stating that every extended output map for $\mathbb{T}$ is of the form (2.4) for some admissible observation operator C.

By a *well-posed linear system* we mean a linear time-invariant system such that on any finite time interval, the operator from the initial state and the input function to the final state and the output function is bounded. The input, state, and output spaces are Hilbert spaces, and the input and output functions are of class L^2_{loc}. For the detailed definition, background, and examples we refer to Salamon [27], [28], Staffans [29], [30], Weiss [36], [37], Avalos and Weiss [5], and Weiss^2 [41].

We recall some necessary facts about well-posed linear systems. Let Σ be such a system, with input space U, state space X, and output space Y. We consider positive time, $t \geq 0$. The state trajectories of Σ satisfy (2.1) and the comments from the beginning of this section apply. $\mathbb{T}$ is called the *semigroup* of Σ and B is called the *control operator* of Σ. If u is the input function of Σ, x_0 is its initial state, and y is the corresponding output function, then

$$y = \Psi x_0 + \mathbb{F} u . \tag{2.7}$$

Here, Ψ is an extended output map for $\mathbb{T}$, so that it can be represented by (2.4), and C is called the *observation operator* of Σ.

The operator $\mathbb{F} : L^2_{loc}([0,\infty), U) \to L^2_{loc}([0,\infty), Y)$ satisfies the following functional equation:

$$\mathbf{S}_\tau^* \mathbb{F} = \Psi \Phi_\tau + \mathbb{F} \mathbf{S}_\tau^* \qquad \forall\, \tau \geq 0 , \tag{2.8}$$

where Φ_τ is the operator from (2.2) and $\mathbf{S}_\tau^*$ is as in (2.6). It follows from (2.6) and (2.8) that if $x(\tau) = \mathbb{T}_\tau x_0 + \Phi_\tau u$ and y is given by (2.7), then

$$\mathbf{S}_\tau^* y = \Psi x(\tau) + \mathbb{F} \mathbf{S}_\tau^* u . \tag{2.9}$$

$\mathbb{F}$ is easiest to represent using Laplace transforms. An operator-valued analytic function is called *well-posed* if its domain contains a right half-plane in $\mathbb{C}$ such that the function is uniformly bounded on this half-plane. We do not distinguish between two well-posed functions if one is a restriction of the other. There exists a unique $\mathcal{L}(U,Y)$-valued well-posed function $\mathbf{G}$, called the *transfer function* of Σ, which determines $\mathbb{F}$ as follows: if $u \in L^2([0,\infty), U)$ and $y = \mathbb{F} u$, then y has a Laplace transform $\hat{y}$ and, for Re s sufficiently large,

$$\hat{y}(s) = \mathbf{G}(s) \hat{u}(s) .$$

This determines $\mathbb{F}$, since L^2 is dense in L^2_{loc}. We have

$$\mathbf{G}(s) - \mathbf{G}(\beta) = C \left[(sI - A)^{-1} - (\beta I - A)^{-1} \right] B$$

for any s, β in the open right half-plane determined by the growth bound of $\mathbb{T}$. This shows that $\mathbf{G}$ is determined by A, B, and C up to an additive constant operator. If $\mathbb{T}$ is exponentially stable, then $\mathbf{G}$ is in H^∞, i.e., it is bounded on the half-plane where Re $s > 0$ (and even on a larger half-plane).

An operator $K \in \mathcal{L}(Y, U)$ is called an *admissible feedback operator* for Σ (or for $\mathbf{G}$) if $I - \mathbf{G}K$ has a well-posed inverse (equivalently, if $I - K\mathbf{G}$ has a well-posed inverse). If this is the case, then the system with output feedback shown in Figure 1 is well-posed (its input is v, its state and output are the same as for Σ). This new system is called the *closed-loop system* corresponding to Σ and K, and it is denoted by Σ^K. Its transfer function is $\mathbf{G}^K = \mathbf{G}(I - K\mathbf{G})^{-1} = (I - \mathbf{G}K)^{-1}\mathbf{G}$. We have that $-K$ is an admissible feedback operator for Σ^K and the corresponding closed-loop system is Σ. For more details on closed-loop systems we refer to [37]. Any interconnection of finitely many well-posed systems can be thought of as a closed-loop system in the above sense.

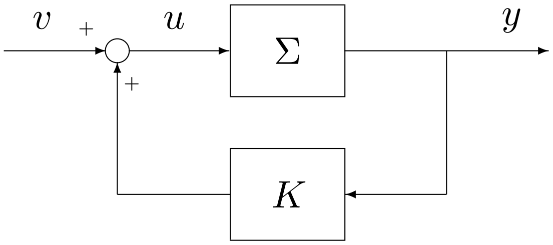


FIG. 1. *A well-posed linear system Σ with output feedback via K. If K is admissible, then this is a new well-posed linear system Σ^K, called the closed-loop system.*

The system Σ is called *regular* if the limit

$$\lim_{s \to +\infty} \mathbf{G}(s)\mathrm{v} = D\mathrm{v}$$

exists for every $\mathrm{v} \in U$, where s is real (see [36]). In this case, the operator $D \in \mathcal{L}(U, Y)$ is called the *feedthrough operator* of Σ. Regularity is equivalent to the fact that the product $C_\Lambda(sI - A)^{-1}B$ makes sense. In this case,

$$\mathbf{G}(s) = C_\Lambda(sI - A)^{-1}B + D, \tag{2.10}$$

as in finite dimensions. Moreover, the function y from (2.7) satisfies, for almost every $t \geq 0$,

$$y(t) = C_\Lambda x(t) + Du(t), \tag{2.11}$$

where x is the state trajectory of the system. The operators A, B, C, D are called the *generating operators* of Σ because they determine Σ via (2.1) and (2.11). Regular linear systems are usually considered on the time interval $[0, \infty)$, but other intervals are possible, for example, $(-\infty, 0]$. We refer to section 5 of [41] for a discussion of this latter case. The equations (2.1) and (2.11) are not affected by the time interval chosen. (A, B, C) is called a *regular triple* if $A, B, C, 0$ are the generating operators of a regular linear system. Equivalently, A generates a semigroup, B and C are admissible, the product $C_\Lambda(sI - A)^{-1}B$ exists and it is bounded on some right half-plane (see section 2 of [40]). In particular, if A is a generator, one of B and C is admissible and the other is bounded, then (A, B, C) is a regular triple. Thus, the systems in (1.1) and (1.2) are both regular and their generating operators are $A, B, I, 0$ for (1.1) and $A, I, C, 0$ for (1.2).

The generator of a semigroup is called *exponentially stable* if the corresponding semigroup is exponentially stable.

DEFINITION 2.1. *(A, B) is stabilizable if there exists $F \in \mathcal{L}(X_1, U)$ such that*

(i) *(A, B, F) is a regular triple;*

(ii) *I is an admissible feedback operator for $F_\Lambda(sI - A)^{-1}B$;*

(iii) *$A + BF_\Lambda$ (with its natural domain) is exponentially stable.*

In this case we say that F stabilizes (A, B).

For the concepts appearing in this definition and for further comments we refer to [22] and [40], but we summarize here the main facts in a few sentences. Since (A, B, F) is a regular triple, the transfer function $\mathbf{G}(s) = F_\Lambda(sI - A)^{-1}B$ is well defined. Recall that I being an admissible feedback operator for $\mathbf{G}$ means that $(I - \mathbf{G})^{-1}$ exists and is well-posed. The natural domain of $A^f = A + BF_\Lambda$ is

$$\mathcal{D}(A^f) = \left\{ z \in \mathcal{D}(F_\Lambda) \,\middle|\, Az + BF_\Lambda z \in X \right\} .$$

Conditions (i) and (ii) imply (see [37]) that A^f is the generator of a strongly continuous semigroup $\mathbb{T}^f$ on X and we have

$$\mathbb{T}^f_t x_0 = \mathbb{T}_t x_0 + \int_0^t \mathbb{T}_{t-\sigma} BF_\Lambda \mathbb{T}^f_\sigma x_0 \, d\sigma \,. \tag{2.12}$$

Moreover, (A^f, B, F_Λ) is a regular triple and

$$(I - \mathbf{G}(s))^{-1} = I + F_\Lambda(sI - A^f)^{-1}B \,.$$

It follows from the assumed exponential stability of $\mathbb{T}^f$ (point (iii) in the definition) that $(I - \mathbf{G})^{-1} \in H^\infty(\mathcal{L}(U))$, the space of bounded analytic $\mathcal{L}(U)$-valued functions on the half-plane where Re $s > 0$. Definition 2.1 is rather general, but we mention that Staffans [30] and Morris [19] have proposed a more general (and difficult) definition, with no regularity assumptions. It is not known if their definition is genuinely more general, i.e., if there exists a pair (A, B) which is stabilizable in the sense of [30] and [19], but not in the sense of Definition 2.1.

DEFINITION 2.2. *(A, C) is detectable if there exists $H \in \mathcal{L}(Y, X_{-1})$ such that*

(i) *(A, H, C) is a regular triple;*

(ii) *I is an admissible feedback operator for $C_\Lambda(sI - A)^{-1}H$;*

(iii) *$A + HC_\Lambda$ is exponentially stable.*

In this case we say that H detects (A, C).

The comments that can be made about the conditions above are similar to those at Definition 2.1, so we do not repeat them. This concept is almost the dual of stabilizability, but not quite: the trouble is that if (A, H, C) is a regular triple, it does not follow that (A^*, C^*, H^*) is a regular triple (it is weakly regular; see [41]). If Y is finite-dimensional, then the duality holds (i.e., (A, C) is detectable if and only if (A^*, C^*) is stabilizable); see section 3 of [40].

3. Optimizability.

In this section we are concerned with systems of the type (1.1). We use the notation X, X_{-1}, U, A, B, and $\mathbb{T}$ from sections 1 and 2.

DEFINITION 3.1. *The pair (A, B) is optimizable if for every $x_0 \in X$, there exists a $u \in L^2([0, \infty), U)$ such that $x \in L^2([0, \infty), X)$, where*

$$x(t) = \mathbb{T}_t x_0 + \int_0^t \mathbb{T}_{t-\tau} Bu(\tau) \, d\tau \,. \tag{3.1}$$

Note that x in (3.1) is the solution of (2.1) with $x(0) = x_0$.

REMARK 1. *It is easy to see that stabilizability implies optimizability. Indeed, assume that (A, B) is stabilized by F and let $\mathbb{T}^f$ be the semigroup generated by $A + BF_\Lambda$. For $x_0 \in X$, define the function u by $u(t) = F_\Lambda \mathbb{T}^f_t x_0$ for almost every $t \geq 0$. The exponential stability of $\mathbb{T}^f$ implies that $u \in L^2([0,\infty), U)$. From* (2.12) *we see that the function x from* (3.1) *is in fact $x(t) = \mathbb{T}^f_t x_0$, so that $x \in L^2$.*

We shall need the following cost functional associated with the pair (A, B): for every $x_0 \in X$ and every $u \in L^2([0,\infty), U)$, define

$$J(x_0, u) = \int_0^\infty (\|x(t)\|^2 + \|u(t)\|^2)\, dt, \tag{3.2}$$

where $x(t)$ is as in (3.1). Note that $J(x_0, u)$ might be infinite, and optimizability means that for every x_0, J can be made finite.

The following three results are known from linear quadratic optimal control theory, as developed in Flandoli, Lasiecka and Triggiani [13] (see also Staffans [31] or Zwart [42]). In [13], optimizability is called the finite cost condition. (In Weiss[2] [41], the particular case of a stable system is treated, and our notation follows this paper.)

PROPOSITION 3.2. *Suppose that (A, B) is optimizable. Then for every $x_0 \in X$ there is a unique function $u^{\rm opt} \in L^2([0,\infty), U)$ such that*

$$J(x_0, u^{\rm opt}) = \min_u J(x_0, u). \tag{3.3}$$

There is a positive operator $P \in \mathcal{L}(X)$ such that for all x_0 and $u^{\rm opt}$ as above,

$$J(x_0, u^{\rm opt}) = \langle P x_0, x_0 \rangle.$$

The following proposition tells us that the optimal state trajectories of the system (1.1), given by (3.1) with $u = u^{\rm opt}$, determine a semigroup.

PROPOSITION 3.3. *Suppose that (A, B) is optimizable. Then there is a strongly continuous semigroup $\mathbb{T}^{\rm opt}$ on X such that for every $x_0 \in X$, if $u^{\rm opt}$ is as in* (3.3), *then*

$$\mathbb{T}^{\rm opt}_t x_0 = \mathbb{T}_t x_0 + \int_0^t \mathbb{T}_{t-\sigma} B u^{\rm opt}(\sigma)\, d\sigma. \tag{3.4}$$

The semigroup $\mathbb{T}^{\rm opt}$ is exponentially stable.

The third result concerns an operator $F^{\rm opt}$, which in some weak sense can be thought of as an optimal feedback operator.

PROPOSITION 3.4. *With the notation of the previous two propositions, let $A^{\rm opt} : \mathcal{D}(A^{\rm opt}) \to X$ denote the generator of $\mathbb{T}^{\rm opt}$. Then $P : \mathcal{D}(A^{\rm opt}) \to \mathcal{D}(A^*)$, so that we can define $F^{\rm opt} : \mathcal{D}(A^{\rm opt}) \to U$ by*

$$F^{\rm opt} x_0 = -B^* P x_0 \qquad \forall\, x_0 \in \mathcal{D}(A^{\rm opt}). \tag{3.5}$$

$F^{\rm opt}$ is an admissible observation operator for $\mathbb{T}^{\rm opt}$, and for every $x_0 \in \mathcal{D}(A^{\rm opt})$,

$$u^{\rm opt}(t) = F^{\rm opt} \mathbb{T}^{\rm opt}_t x_0 \qquad \forall\, t \geq 0, \tag{3.6}$$

$$A^{\rm opt} x_0 = (A + BF^{\rm opt}) x_0. \tag{3.7}$$

The operator P satisfies a Riccati equation on $\mathcal{D}(A^{\rm opt})$ and possibly also on $\mathcal{D}(A)$, but this is not needed here, so we only refer to [13], [31], [41], [42].

It follows from (3.6) that, denoting by $F_\Lambda^{\rm opt}$ the Λ-extension of $F^{\rm opt}$, as in (1.3),

$$u^{\rm opt}(t) = F_\Lambda^{\rm opt}\mathbb{T}_t^{\rm opt}x_0\,, \tag{3.8}$$

for every $x_0 \in X$ and almost every $t \geq 0$. The Laplace transform of $u^{\rm opt}$ is

$$\hat{u}^{\rm opt}(s) = F^{\rm opt}(sI - A^{\rm opt})^{-1}x_0\,. \tag{3.9}$$

Combining (3.8) with (3.4), we get that

$$\mathbb{T}_t^{\rm opt}x_0 = \mathbb{T}_t x_0 + \int_0^t \mathbb{T}_{t-\sigma}BF_\Lambda^{\rm opt}\mathbb{T}_\sigma^{\rm opt}x_0\,d\sigma\,. \tag{3.10}$$

This formula resembles (2.12), but the context is different: for example, we do not know if B is an admissible control operator for $\mathbb{T}^{\rm opt}$. The formula (3.7) can be derived from (3.10) using the Laplace transformation.

The following proposition is an extension of the Hautus test for stabilizability to infinite-dimensional systems. It is a simple particular case of a result in section 2 of Rebarber and Zwart [23]. We denote by Ran $[sI - A\,|\,B]$ the subspace of X_{-1} which consists of all vectors of the form $(sI - A)z + B{\rm v}$, where $z \in X$ and ${\rm v} \in U$.

Proposition 3.5. *If (A, B) is optimizable, then there exists a $\delta > 0$ such that, for all $s \in \mathbb{C}$ with* Re $s > -\delta$,

$$\text{Ran }[sI - A\,|\,B] \supset X\,.$$

Proof. Let $A^{\rm opt}$ and $F^{\rm opt}$ be the operators from Proposition 3.4. Since $A^{\rm opt}$ is exponentially stable, there exists a $\delta > 0$ such that $sI - A^{\rm opt}$ is invertible for all s with Re $s > -\delta$. We have for every such s and for every $x_0 \in X$,

$$\begin{aligned} x_0 &= (sI - A^{\rm opt})(sI - A^{\rm opt})^{-1}x_0 \\ &= (sI - A)(sI - A^{\rm opt})^{-1}x_0 - BF^{\rm opt}(sI - A^{\rm opt})^{-1}x_0\,. \end{aligned}$$

Denoting $z = (sI - A^{\rm opt})^{-1}x_0$ and ${\rm v} = -F^{\rm opt}(sI - A^{\rm opt})^{-1}x_0$, we get the desired representation of x_0: $x_0 = (sI - A)z + B{\rm v}$. □

Definition 3.6. *The open-loop L^2-stabilization problem for (A, B) is to find a bounded linear operator*

$$\mathbf{F} : X \rightarrow L^2([0, \infty), U)$$

such that, taking in (3.1) $u = \mathbf{F}x_0$, *x should depend continuously on x_0, i.e.,*

$$\sup_{\|x_0\| \leq 1} \|x\|_{L^2([0,\infty),X)} < \infty\,. \tag{3.11}$$

It looks as if the solvability of this problem is a more restrictive condition than optimizability, but the optimal control theory results listed earlier imply that they are in fact equivalent (this will be needed in section 4).

Proposition 3.7. *The open-loop L^2-stabilization problem for (A, B) is solvable if and only if (A, B) is optimizable. If the latter condition holds, then one solution of the open-loop L^2-stabilization problem is the operator $\mathbf{F}^{\rm opt}$ defined by*

$$(\mathbf{F}^{\rm opt}x_0)(t) = F_\Lambda^{\rm opt}\mathbb{T}_t^{\rm opt}x_0 \tag{3.12}$$

for every $x_0 \in X$ *and almost every* $t \geq 0$*, where* F_Λ^{opt} *and* $\mathbb{T}^{\text{opt}}$ *are as in* (3.8)*.*

Proof. It is obvious that the solvability of the open-loop L^2-stabilization problem via an operator $\mathbf{F}$ implies optimizability by choosing $u = \mathbf{F}x_0$.

Conversely, suppose that (A, B) is optimizable. Since $\mathbb{T}^{\text{opt}}$ is exponentially stable and F^{opt} is admissible for $\mathbb{T}^{\text{opt}}$, the operator $\mathbf{F}^{\text{opt}}$ from (3.12) is bounded, as required in Definition 3.6. If we take in (3.1) $u = \mathbf{F}^{\text{opt}}x_0$, then $u = u^{\text{opt}}$ and by (3.4), $x(t) = \mathbb{T}_t^{\text{opt}}x_0$. Since $\mathbb{T}^{\text{opt}}$ is exponentially stable, the above operator from x_0 to the function x in $L^2([0,\infty), X)$ is bounded, i.e., (3.11) holds. □

We want to introduce the optimal version of the open-loop L^2-stabilization problem. For this, we introduce the operators

$$\Psi : X \to L^2_{loc}([0,\infty), X), \qquad \mathbb{F} : L^2([0,\infty), U) \to L^2_{loc}([0,\infty), X)$$

by

$$(\Psi x_0)(t) = \mathbb{T}_t x_0, \qquad (\mathbb{F}u)(t) = \int_0^t \mathbb{T}_{t-\sigma} Bu(\sigma)\, d\sigma. \tag{3.13}$$

(Both Ψx_0 and $\mathbb{F}u$ are in fact continuous X-valued functions of t.) Ψ and $\mathbb{F}$ are the operators from (2.7) for the regular linear system described by (1.1). If the semigroup $\mathbb{T}$ is exponentially stable, then the operators Ψ and $\mathbb{F}$ are bounded if we apply the L^2-norm to Ψx_0 and to $\mathbb{F}u$. Without exponential stability, Ψ and $\mathbb{F}$ are in general unbounded, and this is the case of interest.

The function x from (3.1) is $x = \Psi x_0 + \mathbb{F}u$, so that if $\mathbf{F}$ is an operator from X to $L^2([0,\infty), U)$, then taking $u = \mathbf{F}x_0$, x can be written as

$$x = (\mathbb{F}\mathbf{F} + \Psi)x_0. \tag{3.14}$$

This shows that the supremum appearing on the left of (3.11) is the norm of the operator $\mathbb{F}\mathbf{F} + \Psi$ (from X to $L^2([0,\infty), X)$). Thus, we obtain the following reformulation of the open-loop L^2-stabilization problem for (A, B): find a bounded linear operator $\mathbf{F} : X \to L^2([0,\infty), U)$ such that $\mathbb{F}\mathbf{F} + \Psi$ is bounded.

Definition 3.8. *The optimal open-loop L^2-stabilization problem for (A, B) is to find a bounded operator*

$$\mathbf{F}^{\text{opt}} : X \to L^2([0,\infty), U)$$

such that $\mathbb{F}\mathbf{F}^{\text{opt}} + \Psi$ *is bounded (from* X *to* $L^2([0,\infty), X)$*) and*

$$\left\| \begin{bmatrix} \mathbb{F}\mathbf{F}^{\text{opt}} + \Psi \\ \mathbf{F}^{\text{opt}} \end{bmatrix} \right\| = \min_{\mathbf{F}} \left\| \begin{bmatrix} \mathbb{F}\mathbf{F} + \Psi \\ \mathbf{F} \end{bmatrix} \right\|.$$

Proposition 3.9. *The optimal open-loop L^2-stabilization problem for (A, B) is solvable if and only if (A, B) is optimizable.*

If the latter condition holds, then one solution of the optimal open-loop L^2-stabilization problem is the operator $\mathbf{F}^{\text{opt}}$ *defined in* (3.12)*.*

Proof. It is clear that the solvability of the optimal open-loop L^2-stabilization problem via an operator $\mathbf{F}^{\text{opt}}$ implies optimizability, by choosing $u = \mathbf{F}^{\text{opt}}x_0$.

Conversely, suppose that (A, B) is optimizable, and let $\mathbf{F}^{\text{opt}}$ be defined by (3.12). We know from Proposition 3.7 that $\mathbf{F}^{\text{opt}}$ solves the open-loop L^2-stabilization problem, so that (using the reformulation of this problem given before Definition 3.8) $\mathbb{F}\mathbf{F}^{\text{opt}} + \Psi$ is bounded. From (3.14) it is easy to see that

$$J(x_0, \mathbf{F}x_0) = \left\| \begin{bmatrix} \mathbb{F}\mathbf{F} + \Psi \\ \mathbf{F} \end{bmatrix} x_0 \right\|^2,$$

where J is the cost functional from (3.2). For each $x_0 \in X$, the left-hand side above is minimized for $\mathbf{F} = \mathbf{F}^{\rm opt}$ (this follows from Proposition 3.4), so that the norm of the operator appearing on the right-hand side must be minimal for $\mathbf{F} = \mathbf{F}^{\rm opt}$. □

Recently, Jacob and Zwart [16] have obtained interesting necessary conditions for optimizability of systems with a finite-dimensional input space. These conditions concern the spectrum of A and the multiplicity of its eigenvalues.

4. Estimatability. In this section we are concerned with systems of the type (1.2). We use the notation $X, X_1, Y, A, C, C_\Lambda$, and $\mathbb{T}$ from sections 1 and 2. If we denote by X^*_{-1} the completion of X with respect to the norm $\|x\|^*_{-1} = \|(\beta I - A^*)^{-1}x\|$, then $X^*_{-1} = (X_1)^*$ and $C^* \in \mathcal{L}(Y, X^*_{-1})$. As is well known, C^* is an admissible control operator for $\mathbb{T}^*$. We have $A^* : \mathcal{D}(A^*) \to X$, which can be extended to $A^* : X \to X^*_{-1}$.

DEFINITION 4.1. *The pair* (A, C) *is estimatable if* (A^*, C^*) *is optimizable.*

Estimatability is equivalent to the solvability of the final state estimation problem, which we describe in what follows.

Consider the system Σ described by (1.2). Recall that the time is negative. We assume that $u \in L^2((-\infty, 0], X)$ and u has compact support. The system is at rest before u becomes active, i.e., $x(t) = 0$ if $u(\tau) = 0$ for all $\tau \le t$.

DEFINITION 4.2. *The final state estimation problem for* (A, C) *is to find a bounded linear operator*

$$\mathbf{E} : L^2((-\infty, 0], Y) \to X$$

such that for the system Σ *in* (1.2), *denoting* $e = \mathbf{E}y - x(0)$, e *should depend continuously on* u, *i.e,*

$$\sup_{\|u\| \le 1} \|e\| < \infty . \tag{4.1}$$

If (4.1) holds, then $\mathbf{E}y$ is a reasonable estimate (or guess) of $x(0)$, based only on the information y, and e is the estimation error; see Figure 2 (with $w = 0$). The operator $\mathbf{E}$ could be called a *final state estimator* for the system in (1.2).

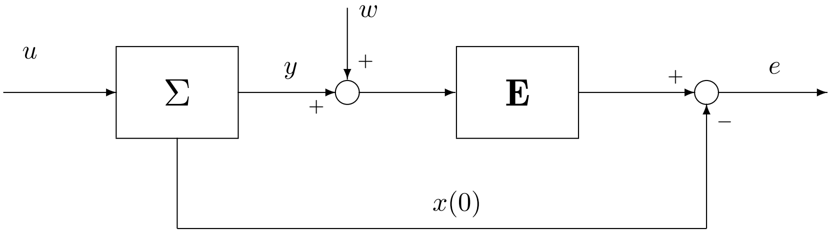


FIG. 2. *The final state estimation problem. The system* Σ *is described by* (1.2), *the time is negative, and* $w = 0$. *We are looking for a bounded operator* $\mathbf{E}$ *(the estimator) such that the operator from* u *to* e *should be bounded. Here,* e *is the estimation error. In the optimal version of this problem, the output noise* w *is in* L^2 *and we want to minimize the norm of the operator from the pair* (u, w) *to* e.

It follows from the description of Σ in (1.2) that

$$x(0) = \Phi u, \qquad y = \mathbb{L}u,$$

where

$$\Phi u = \int_0^\infty \mathbb{T}_\sigma u(-\sigma)\, d\sigma, \qquad (\mathbb{L}u)(t) = C_\Lambda \int_t^\infty \mathbb{T}_{t+\sigma} u(-\sigma)\, d\sigma . \tag{4.2}$$

If $\mathbb{T}$ is exponentially stable, then these operators are bounded if we apply the L^2-norm to u. If $\mathbb{T}$ is not exponentially stable, then the operators Φ and $\mathbb{L}$ are in general unbounded (this is the case of interest) and their domain contains the L^2-functions with compact support. With this notation, the formula for e becomes

$$e = (\mathbf{E}\mathbb{L} - \Phi)u. \tag{4.3}$$

The expression on the left of (4.1) is the norm of the operator $\mathbf{E}\mathbb{L} - \Phi$ (from $L^2((-\infty, 0], X)$ to X). Thus, we obtain the following reformulation of the final state estimation problem for (A, C): find a bounded linear operator $\mathbf{E}$ from $L^2((-\infty, 0], Y)$ to X such that $\mathbf{E}\mathbb{L} - \Phi$ is bounded.

The time-reflection operator $\mathbf{Я} : L^2((-\infty, 0], W) \to L^2([0, \infty), W)$ is defined by

$$(\mathbf{Я}w)(t) = w(-t) \qquad \forall t \geq 0,$$

where W is an arbitrary Hilbert space. Note that $(\mathbf{Я}^* v)(t) = v(-t)$ for all v in $L^2([0, \infty), W)$, so that $\mathbf{Я}\mathbf{Я}^* = I$ and $\mathbf{Я}^*\mathbf{Я} = I$.

THEOREM 4.3. *The final state estimation problem for* (A, C) *is solvable if and only if* (A, C) *is estimatable. If the latter condition holds, then one solution of the final state estimation problem is the operator* $\mathbf{E}^{\rm opt}$ *defined by*

$$\mathbf{E}^{\rm opt} = -\left(\mathbf{F}^{\rm opt}\right)^* \mathbf{Я}, \tag{4.4}$$

where $\mathbf{F}^{\rm opt}$ *is the operator from Proposition* 3.7*, but with* (A^*, C^*) *in place of* (A, B).

A system-theoretic interpretation of (4.4) will be given after the proof.

Proof. Suppose that (A, C) is estimatable. By Proposition 3.7, the open-loop L^2-stabilization problem is solvable for (A^*, C^*), and one solution is the operator $\mathbf{F}^{\rm opt}$ from (3.12). Thus, with the notation from (3.13) (with $\mathbb{T}^*$ and C^* in place of $\mathbb{T}$ and B) we have that $\mathbb{F}\mathbf{F}^{\rm opt} + \Psi$ is bounded (from X to $L^2([0, \infty), X)$).

We claim that $\mathbf{E}^{\rm opt}\mathbb{L} - \Phi$ is bounded and

$$\mathbf{E}^{\rm opt}\mathbb{L} - \Phi = -\left(\mathbb{F}\mathbf{F}^{\rm opt} + \Psi\right)^* \mathbf{Я}. \tag{4.5}$$

We have to be careful with the proof of (4.5), because the operators $\mathbb{L}$, Φ, $\mathbb{F}$, and Ψ are in general unbounded. Let $u \in L^2((-\infty, 0], X)$ have compact support and take $z \in X$ and $v \in L^2([0, \infty), Y)$. We need the identities

$$\langle \mathbf{Я}\mathbb{L}u, v\rangle = \langle \mathbf{Я}u, \mathbb{F}v\rangle, \qquad \langle \Phi u, z\rangle = \langle \mathbf{Я}u, \Psi z\rangle. \tag{4.6}$$

By $\langle \mathbf{Я}u, \mathbb{F}v\rangle$ we mean $\int_0^\infty \langle u(-t), (\mathbb{F}v)(t)\rangle dt$, which makes sense because u has compact support, and a similar explanation applies to $\langle \mathbf{Я}u, \Psi z\rangle$. The formulas (4.6) are verified by simple computations, using (3.13) and (4.2).

We have, using (4.4) and (4.6),

$$\begin{aligned}
\langle (\mathbf{E}^{\rm opt}\mathbb{L} - \Phi)u, z\rangle &= -\langle \mathbf{Я}\mathbb{L}u, \mathbf{F}^{\rm opt}z\rangle - \langle \Phi u, z\rangle \\
&= -\langle \mathbf{Я}u, \mathbb{F}\mathbf{F}^{\rm opt}z\rangle - \langle \mathbf{Я}u, \Psi z\rangle \\
&= -\langle \mathbf{Я}u, (\mathbb{F}\mathbf{F}^{\rm opt} + \Psi)z\rangle .
\end{aligned}$$

Since functions with compact support are dense in $L^2((-\infty,0],X)$, this shows that $\mathbf{E}^{\rm opt}\mathbb{L}-\Phi$ is bounded and (4.5) holds. Thus, $\mathbf{E}^{\rm opt}$ is a solution of the final state estimation problem for (A,C) (it is a final state estimator).

Conversely, suppose that the final state estimation problem for (A,C) is solvable and let $\mathbf{E}$ be a final state estimator for the system in (1.2), i.e., $\mathbf{E}\mathbb{L}-\Phi$ is bounded (from $L^2((-\infty,0],X)$ to X). Define $\mathbf{F}: X \rightarrow L^2([0,\infty),X)$ by

$$\mathbf{F} = -\boldsymbol{Я}\mathbf{E}^*,$$

so that $\mathbf{E}=-\mathbf{F}^*\boldsymbol{Я}$, like in (4.4). Then by a similar argument as in the first part of this proof, we obtain that

$$\langle(\mathbf{E}\mathbb{L}-\Phi)u, z\rangle = -\langle\boldsymbol{Я}u, (\mathbb{F}\mathbf{F}+\Psi)z\rangle, \tag{4.7}$$

which shows that $\mathbb{F}\mathbf{F}+\Psi$ is bounded. This means that the open-loop L^2-stabilization problem is solvable for (A^*,C^*). By Proposition 3.7, (A^*,C^*) is optimizable, i.e., (A,C) is estimatable. □

To understand the meaning of (4.4), we write it in a more explicit way. Let $\mathbb{T}^{\rm opt}$ denote the optimal semigroup from (3.4), corresponding to the optimal control problem (3.1), (3.2), (3.3), but with (A,B) replaced by (A^*,C^*) (in particular, $\mathbb{T}$ is replaced by its adjoint semigroup $\mathbb{T}^*$). We denote by $A^{\rm opt}$ the generator of $\mathbb{T}^{\rm opt}$ and by $F^{\rm opt}$ the operator from (3.5), so that $F^{\rm opt}\in\mathcal{L}(X_1^{\rm opt},Y)$, where $X_1^{\rm opt}$ is $\mathcal{D}(A^{\rm opt})$ with the graph norm. By Proposition 3.4, $F^{\rm opt}$ is an admissible observation operator for $\mathbb{T}^{\rm opt}$ and, according to (3.5), $F^{\rm opt}=-CP$.

Let $A_d^{\rm opt}$ and $H^{\rm opt}$ denote the adjoints of $A^{\rm opt}$ and $F^{\rm opt}$. Thus, $H^{\rm opt}\in\mathcal{L}(Y,X_{-1}^d)$, where X_{-1}^d is the dual of $X_1^{\rm opt}$ (the completion of X with respect to the norm $\|z\|_{-1}^d = \|(A_d^{\rm opt})^{-1}z\|$). $A_d^{\rm opt}$ is the generator of the semigroup $\mathbb{S}^{\rm opt}$, which is the adjoint of the semigroup $\mathbb{T}^{\rm opt}$. By duality we know that $H^{\rm opt}$ is an admissible control operator for $\mathbb{S}^{\rm opt}$. By an easy computation we can write

$$\left(\mathbf{F}^{\rm opt}\right)^* v = \int_0^\infty \mathbb{S}_t^{\rm opt}H^{\rm opt}v(t)\,dt. \tag{4.8}$$

Now we can rewrite the formula for $\mathbf{E}^{\rm opt}$:

$$\mathbf{E}^{\rm opt}y = -\int_0^\infty \mathbb{S}_t^{\rm opt}H^{\rm opt}y(-t)\,dt. \tag{4.9}$$

This corresponds to the following system:

$$\dot z(t) = A_d^{\rm opt}z(t) - H^{\rm opt}y(t) \tag{4.10}$$

with $t\le 0$ and $z(-\infty)=0$. The final state $z(0)$ of this system is $\mathbf{E}^{\rm opt}y$. For bounded C, see also the comments after Proposition 4.4.

Now we derive the dual counterparts of (3.10) and (3.7).

Proposition 4.4. *Assume that (A,C) is estimatable. Let $\mathbb{S}^{\rm opt}$ and $H^{\rm opt}$ be as in* (4.9) *and let $A_d^{\rm opt}$ be the generator of $\mathbb{S}^{\rm opt}$. Then for every $x_0\in X$,*

$$\mathbb{S}_t^{\rm opt}x_0 = \mathbb{T}_t x_0 + \int_0^t \mathbb{S}_{t-\sigma}^{\rm opt}H^{\rm opt}C_\Lambda\mathbb{T}_\sigma x_0\,d\sigma \tag{4.11}$$

and, for every $z_0\in\mathcal{D}(A)$,

$$Az_0 = \left(A_d^{\rm opt} - H^{\rm opt}C\right)z_0\,. \tag{4.12}$$

Proof. The formula (4.11) follows from (3.10) (with $\mathbb{T}_t^*$ and C^* in place of $\mathbb{T}_t$ and B) by taking adjoints and making some simple computations. Applying the Laplace transformation to (4.11), we obtain

$$(4.13)\quad (sI - A_d^{\rm opt})^{-1}x_0 - (sI-A)^{-1}x_0 \,=\, (sI - A_d^{\rm opt})^{-1}H^{\rm opt}C(sI-A)^{-1}x_0\,.$$

Applying $(sI - A_d^{\rm opt})$ to both sides and denoting $z_0 = (sI-A)^{-1}x_0$, we get

$$(sI-A)z_0 - (sI - A_d^{\rm opt})z_0 \,=\, H^{\rm opt}Cz_0$$

(this is an identity in X_{-1}^d). From here, (4.12) follows. □

Assuming for a moment that C (and hence also $H^{\rm opt}$) is bounded, using (4.12), we can rewrite (4.10) of the final state estimator in the form

$$\dot z(t) \,=\, Az(t) + H^{\rm opt}(Cz(t) - y(t))\,.$$

Here we recognize the equation of a Kalman estimator from finite-dimensional systems theory (this estimator has no access to the driving noise u).

Proposition 4.5. *If (A, C) is estimatable, then there is a $K > 0$ such that for every $x_0 \in D(A)$ and every $T \geq 0$*

$$\int_0^T \|\mathbb{T}_t x_0\|^2 dt \,\leq\, K^2\left(\|x_0\|^2 + \int_0^T \|C\mathbb{T}_t x_0\|^2 dt\right).$$

Proof. Since $H^{\rm opt}$ from (4.9) is admissible for $\mathbb{S}^{\rm opt}$ and $\mathbb{S}^{\rm opt}$ is exponentially stable, it follows that $(sI - A_d^{\rm opt})^{-1}H^{\rm opt}$ is in H^∞; see, for example, [38]. Denote

$$k \,=\, \sup_{{\rm Re}\, s>0} \|(sI - A_d^{\rm opt})^{-1}H^{\rm opt}\|_{\mathcal{L}(Y,X)}\,.$$

If $y \in L^2_{loc}([0,\infty), Y)$ and $w : [0,\infty)\to X$ is defined by

$$w(t) \,=\, \int_0^t \mathbb{S}_{t-\sigma}^{\rm opt} H^{\rm opt} y(\sigma)d\sigma,$$

then, since $\hat w(s) = (sI - A_d^{\rm opt})^{-1}H^{\rm opt}\hat y(s)$, we have that for all $T \geq 0$,

$$\int_0^T \|w(t)\|^2 dt \,\leq\, k^2 \int_0^T \|y(t)\|^2 dt.$$

Take in this inequality $y(t) = C_\Lambda \mathbb{T}_t x_0$. Then (4.11) implies that for all $T \geq 0$,

$$\left(\int_0^T \|\mathbb{T}_t x_0\|^2 dt\right)^{\frac12} \;\leq\; \left(\int_0^T \|\mathbb{S}_t^{\rm opt} x_0\|^2 dt\right)^{\frac12} + k\left(\int_0^T \|C_\Lambda \mathbb{T}_t x_0\|^2 dt\right)^{\frac12}.$$

Since the first term on the right-hand side is $\leq M\|x_0\|$, denoting $K^2 = M^2 + k^2$ we get the estimate in the proposition. □

Now consider the situation when there is an output noise (or uncertainty) w added to the output y of the plant from (1.2), as shown in Figure 2. The function w is in $L^2((-\infty, 0], Y)$. It is easy to see that in this situation, if $\mathbf{E}$ is a final state estimator for the system in (1.2), then the estimation error will be

$$(4.14)\qquad e \,=\, \begin{bmatrix} \mathbf{E}\mathbb{L} - \Phi & \mathbf{E} \end{bmatrix} \begin{bmatrix} u \\ w \end{bmatrix}.$$

Trying to minimize this error leads to the following problem.

DEFINITION 4.6. *The optimal final state estimation problem for* (A, C) *is to find a bounded linear operator*

$$\mathbf{E}^{\mathrm{opt}} : L^2((-\infty, 0], Y) \to X$$

such that $\mathbf{E}^{\mathrm{opt}}\mathbb{L} - \Phi$ *is bounded (from* $L^2((-\infty, 0], X)$ *to* X*) and*

$$\left[\begin{array}{cc} \mathbf{E}^{\mathrm{opt}}\mathbb{L} - \Phi & \mathbf{E}^{\mathrm{opt}} \end{array}\right] = \min_{\mathbf{E}} \left[\begin{array}{cc} \mathbf{E}\mathbb{L} - \Phi & \mathbf{E} \end{array}\right].$$

Note that this problem is the dual of the optimal open-loop L^2-stabilization problem introduced in Definition 3.8.

PROPOSITION 4.7. *The optimal final state estimation problem for* (A, C) *is solvable if and only if* (A, C) *is estimatable. If the latter condition holds, then one solution of the optimal final state estimation problem is the operator* $\mathbf{E}^{\mathrm{opt}}$ *defined in* (4.4).

The above result is the dual counterpart of Proposition 3.9 and it follows from it by duality, using formula (4.7).

PROPOSITION 4.8. *If* (A, C) *is estimatable, then there exist* $\delta > 0$ *and* $m > 0$ *such that, for all* $s \in \mathbb{C}$ *with* $\mathrm{Re}\ s > -\delta$,

$$\|(sI - A)z\| + \|Cz\| \geq m\|z\| \qquad \forall\, z \in \mathcal{D}(A). \tag{4.15}$$

This is approximately the dual version of Proposition 3.5 (the Hautus test). Actually, it is somewhat stronger than the dual of Proposition 3.5 because we prove a uniform lower bound instead of just a lower bound for each individual s.

Proof. According to (4.12), we have for every $z \in \mathcal{D}(A)$,

$$(sI - A_d^{\mathrm{opt}})^{-1}(sI - A - H^{\mathrm{opt}}C)z = z,$$

which means that

$$\left[(sI - A_d^{\mathrm{opt}})^{-1} \quad -(sI - A_d^{\mathrm{opt}})^{-1}H^{\mathrm{opt}}\right] \left[\begin{array}{c} sI - A \\ C \end{array}\right] z = z.$$

Since A_d^{opt} is exponentially stable, there is a $\delta > 0$ such that both $(sI - A_d^{\mathrm{opt}})^{-1}$ and $(sI - A_d^{\mathrm{opt}})^{-1}H^{\mathrm{opt}}$ are uniformly bounded on the half-plane where $\mathrm{Re}\ s > -\delta$; see, for example, [38]. Thus, there is an $M > 0$ such that, for all $s \in \mathbb{C}$ with $\mathrm{Re}\ s > -\delta$,

$$\left\|\left[(sI - A_d^{\mathrm{opt}})^{-1} \quad -(sI - A_d^{\mathrm{opt}})^{-1}H^{\mathrm{opt}}\right]\right\| \leq M$$

with the norm evaluated in $\mathcal{L}(X \times Y, X)$. Hence, for such s and every $z \in \mathcal{D}(A)$,

$$\begin{aligned} \|z\| &\leq M \left\| \left[\begin{array}{c} sI - A \\ C \end{array}\right] z \right\| \\ &\leq M\left(\|(sI - A)z\| + \|Cz\|\right). \end{aligned}$$

Denoting $m = \frac{1}{M}$, we get the desired estimate. □

It would be interesting to know whether the converse of Proposition 4.8 is true.

REMARK 2. *The estimate* (4.15) *can be replaced by the stronger looking*

$$\|(sI - A)z\| + \|Cz\| \geq m(1 + |\mathrm{Re}\, s|)\|z\| \qquad \forall\, z \in \mathcal{D}(A),$$

but this does not really tell more than (4.15)*. The last estimate holds because for large* Re s*, it holds even with* $C = 0$*, by semigroup theory. For this reason, numbers* s *with large* Re s *are not very interesting in this context: what is interesting is the area around* $\sigma(A)$ *in the half-plane where* Re $s > -\delta$*.*

For other infinite-dimensional extensions of the Hautus test we refer to Jacob and Zwart [16], Grabowski and Callier [14], and Russell and Weiss [26].

REMARK 3. *Using the operators* Φ *and* $\mathbb{L}$ *defined in* (4.2)*, it is possible to prove the following:* (A, C) *is estimatable if and only if the estimate*

$$\|\Phi u\|^2 \leq K^2 \left(\|u\|^2 + \|\mathbb{L}u\|^2\right)$$

holds for some $K > 0$ *and for all* $u \in L^2((-\infty, 0], X)$ *with compact support. The proof is a bit involved and we omit it, since this does not seem to be a practical way to check estimatability. The key step is that* $\mathbb{L}(I + \mathbb{L}^*\mathbb{L})^{-1}\Phi^*$ *is bounded.*

5. Equivalence of exponential and input-output stability. The purpose of this section is to prove Theorem 1.1. Before doing so, we say a few words about the history of this result. For finite-dimensional systems, it is well known that if the system is stabilizable, detectable and the transfer function is stable (i.e., all the poles are in the open left half-plane), then the system is stable (i.e., all the eigenvalues are in the open left half-plane). We cannot point to a specific source of this proposition. In the 1980s, this result was generalized to infinite-dimensional systems with bounded control and observation operators and finite-dimensional input and output spaces by S. A. Nefedov and F. A. Sholokhovich and by C. A. Jacobson and C. N. Nett. Related results were derived by H. Logemann. For precise bibliographic details on this period we refer to Curtain [8]. For a time-varying finite-dimensional version we refer to Ravi and Khargonekar [21].

The result was generalized to Pritchard–Salamon systems in [8] and then again to well-posed linear systems in Curtain [9]. In the latter reference, however, the definitions of stabilizability, of detectability, and of input-output stability are much more restrictive than the concepts used in this paper. In Rebarber [22] it was shown that a regular linear system is exponentially stable if and only if it is stabilizable, detectable, and input-output stable (see also section 3 of [40]). The papers [22] and [40] use exactly the same terminology which is used here. The result of [22] was generalized by Staffans [30] and (independently) by Morris [19]. Both of them have eliminated all the regularity assumptions: they consider well-posed linear systems, and their definitions of stabilizability and of detectability are also regularity-free, so that they are less restrictive (and more difficult) than the concepts used here. (Actually, in [30] input-output stability is replaced by a more restrictive condition, so that the result in [19] is stronger.) Their definitions of exponential stabilizability and detectability imply optimizability and estimatability, respectively, so that our Theorem 1.1 is a generalization of their result. We mention that [30] gives also related results concerning the strong stability of the semigroup.

To prove Theorem 1.1, we need some preliminary results. We use the notation $X, X_1, X_{-1}, U, Y, A, B, C$ from section 2.

PROPOSITION 5.1. *The following two statements are equivalent:*

(a) (A, B) *is optimizable and* $(sI - A)^{-1}B$ *is in* H^∞*,*

(b) A *is exponentially stable.*

Recall that by "$(sI - A)^{-1}B$ is in H^∞" we mean that $(sI - A)^{-1}B$, as an $\mathcal{L}(U, X)$-valued analytic function of s, is bounded on the half-plane where Re $s > 0$.

Proof. It is clear that (b) implies (a). Now assume that (a) holds. Applying the Laplace transformation to (3.4) and using (3.9), we obtain

$$(sI - A^{\rm opt})^{-1} - (sI - A)^{-1} \,=\, (sI - A)^{-1} B F^{\rm opt} (sI - A^{\rm opt})^{-1} \,. \tag{5.1}$$

Since $\mathbb{T}^{\rm opt}$ is exponentially stable and $F^{\rm opt}$ is admissible for $\mathbb{T}^{\rm opt}$, we have that $F^{\rm opt}(sI - A^{\rm opt})^{-1} \in H^\infty$. Since $(sI - A)^{-1} B \in H^\infty$, it follows that the right-hand side of (5.1) is in H^∞. Since $(sI - A^{\rm opt})^{-1} \in H^\infty$, it follows that $(sI - A)^{-1} \in H^\infty$. By a result of Prüss [20] (see also section 4 of [34]), $\mathbb{T}$ is exponentially stable. □

The following proposition is just the dual version of Proposition 5.1, so that no proof is necessary.

PROPOSITION 5.2. *The following two statements are equivalent:*

(a) *(A, C) is estimatable and $C(sI - A)^{-1}$ is in H^∞,*

(b) *A is exponentially stable.*

We now restate Theorem 1.1.

THEOREM 5.3. *Let Σ be a well-posed linear system with semigroup generator A, control operator B, observation operator C, and transfer function $\mathbf{G}$. Then Σ is exponentially stable if and only if*

(1) *(A, B) is optimizable,*

(2) *(A, C) is estimatable,*

(3) *$\mathbf{G} \in H^\infty$ (i.e, Σ is input-output stable).*

Proof. We denote by $\mathbb{T}$ the semigroup of Σ. It is clear that the exponential stability of $\mathbb{T}$ implies the properties (1), (2), and (3). To prove the converse, we assume that these properties hold. By Propositions 3.2 and 3.3, there is a semigroup $\mathbb{T}^{\rm opt}$ on X such that (3.4) holds. Let Ψ and $\mathbb{F}$ be the operators from (2.7) and $\mathbf{F}^{\rm opt}$ the operator from (3.12). Using the operator Φ_τ from (2.2), we rewrite (3.4):

$$\mathbb{T}^{\rm opt}_\tau x_0 \,=\, \mathbb{T}_\tau x_0 + \Phi_\tau u^{\rm opt} \,. \tag{5.2}$$

We define $\Psi^{\rm opt} = \Psi + \mathbb{F}\mathbf{F}^{\rm opt}$, so that

$$\Psi^{\rm opt} x_0 \,=\, \Psi x_0 + \mathbb{F} u^{\rm opt} \tag{5.3}$$

($\Psi^{\rm opt} x_0$ is the output function of the optimally controlled system).

We claim that $\Psi^{\rm opt}$ is an extended output map for $\mathbb{T}^{\rm opt}$, i.e.,

$$\mathbf{S}^*_\tau \Psi^{\rm opt} \,=\, \Psi^{\rm opt} \mathbb{T}^{\rm opt}_\tau \qquad \forall\, \tau \geq 0 \,, \tag{5.4}$$

as explained in section 2. Indeed, from (5.3), (2.8), and (5.2) we see that

$$\begin{aligned} \mathbf{S}^*_\tau \Psi^{\rm opt} x_0 \;&=\; \mathbf{S}^*_\tau \Psi x_0 + \mathbf{S}^*_\tau \mathbb{F} u^{\rm opt} \\ &=\; \Psi\, \mathbb{T}_\tau x_0 + \Psi \Phi_\tau u^{\rm opt} + \mathbb{F}\, \mathbf{S}^*_\tau u^{\rm opt} \\ &=\; \Psi\, \mathbb{T}^{\rm opt}_\tau x_0 + \mathbb{F}\, \mathbf{S}^*_\tau \mathbf{F}^{\rm opt} x_0 \,. \end{aligned}$$

We know that $\mathbf{F}^{\rm opt}$ is an extended output map for $\mathbb{T}^{\rm opt}$, i.e.,

$$\mathbf{S}^*_\tau \mathbf{F}^{\rm opt} \,=\, \mathbf{F}^{\rm opt} \mathbb{T}^{\rm opt}_\tau \qquad \forall\, \tau \geq 0 \,.$$

Substituting this into the previous formula, we get

$$\mathbf{S}^*_\tau \Psi^{\rm opt} x_0 \,=\, \left(\Psi + \mathbb{F}\mathbf{F}^{\rm opt} \right) \mathbb{T}^{\rm opt}_\tau x_0 \,.$$

Now by the definition of Ψ^{opt} we obtain that (5.4) holds.

By the representation theorem for extended output maps (explained in section 2), there is an operator $C^{\mathrm{opt}}:\mathcal{D}(A^{\mathrm{opt}})\to Y$ with

$$(\Psi^{\mathrm{opt}}x_0)(t) = C^{\mathrm{opt}}\mathbb{T}_t^{\mathrm{opt}}x_0 \qquad \forall\, x_0\in\mathcal{D}(A^{\mathrm{opt}})\,.$$

Applying the Laplace transformation to (5.3) and using (2.5) and (3.9), we get

$$C^{\mathrm{opt}}(sI-A^{\mathrm{opt}})^{-1} = C(sI-A)^{-1} + \mathbf{G}(s)F^{\mathrm{opt}}(sI-A^{\mathrm{opt}})^{-1}\,. \tag{5.5}$$

According to Proposition 3.4, F^{opt} (like C^{opt}) is an admissible observation operator for $\mathbb{T}^{\mathrm{opt}}$. Moreover, according to Proposition 3.3, $\mathbb{T}^{\mathrm{opt}}$ is exponentially stable. This implies (see, e.g., Weiss [38]) that

$$F^{\mathrm{opt}}(sI-A^{\mathrm{opt}})^{-1}\in H^\infty\,, \qquad C^{\mathrm{opt}}(sI-A^{\mathrm{opt}})^{-1}\in H^\infty\,.$$

Using this and property (3) in the theorem, we conclude from (5.5) that $C(sI-A)^{-1}\in H^\infty$. By Proposition 5.2, A is exponentially stable. □

We make a remark about the particular case when Σ is regular, with feedthrough operator D. Then from (5.1) and (2.10) we see that $\mathcal{D}(A^{\mathrm{opt}})\subset\mathcal{D}(C_\Lambda)$ and

$$C_\Lambda(sI-A^{\mathrm{opt}})^{-1} - C(sI-A)^{-1} = [\mathbf{G}(s)-D]\,F^{\mathrm{opt}}(sI-A^{\mathrm{opt}})^{-1}\,.$$

Comparing this with (5.5), we conclude that

$$C^{\mathrm{opt}}z = (C_\Lambda + DF^{\mathrm{opt}})z \qquad \forall\, z\in\mathcal{D}(A^{\mathrm{opt}})\,.$$

We introduce two more concepts. We say that (A,C) is *exactly observable* if (A^*,C^*) is exactly controllable. It is clear that if (A,C) is exactly observable, then it is also estimatable. We say that a transfer function $\mathbf{G}$ is in H^∞_α (with $\alpha\in\mathbb{R}$) if $\mathbf{G}$ is bounded (and analytic) on the half-plane where Re $s>\alpha$.

COROLLARY 5.4. *Let Σ, A, B, C, and $\mathbf{G}$ be as in Theorem 5.3 and let $\mathbb{T}$ be the semigroup of Σ. If (A,B) is exactly controllable, (A,C) is exactly observable and if $\mathbf{G}\in H^\infty_\alpha$, then there exist $\beta<\alpha$ and $M\geq 1$ such that*

$$\|\mathbb{T}_t\| \leq M\,e^{\beta t} \qquad \forall\, t\geq 0\,. \tag{5.6}$$

Proof. Introduce the "shifted" well-posed linear system Σ^α with semigroup generator $A-\alpha I$, control operator B, observation operator C, and transfer function $\mathbf{G}^\alpha(s)=\mathbf{G}(s+\alpha)$. This system is again exactly controllable and exactly observable, so that it is optimizable and estimatable. Its transfer function is in H^∞, so that by Theorem 5.3, $A-\alpha I$ is exponentially stable. This is equivalent to (5.6). □

Finally, we give a proposition which is related to Proposition 5.2. For this, we need the following concept: C is called *infinite-time admissible* for $\mathbb{T}$ if it is possible to take $T=\infty$ in (2.3). For a detailed discussion of this concept (e.g., its relation to Lyapunov equations) we refer to Grabowski and Callier [14] and to Hansen and Weiss [15]. Recall that, according to the notation used in this section, C is an admissible observation operator for $\mathbb{T}$, the semigroup generated by A.

PROPOSITION 5.5. *The following two statements are equivalent:*

(a) *(A,C) is estimatable and C is infinite-time admissible for $\mathbb{T}$,*

(b) *A is exponentially stable.*

Proof. It is clear that (b) imples (a). Conversely, if C is infinite-time admissible, then for every $x_0\in X$, $C(sI-A)^{-1}x_0$ is in the Hardy space H^2 (since it is the Laplace transform of a function in L^2). Since (A, C) is estimatable, Proposition 4.4 applies and, in particular, the formula (4.13) from its proof holds. Since A_d^{opt} is exponentially stable, we have that $(sI - A_d^{\text{opt}})^{-1} H^{\text{opt}} \in H^\infty$. This implies that the right-hand side of (4.13) is in H^2. On the left-hand side, $(sI - A_d^{\text{opt}})^{-1} x_0$ is in H^2, so that $(sI - A)^{-1} x_0$ must also be in H^2. Thus, the trajectories of $\mathbb{T}$ are in L^2. By a well-known proposition of Datko, $\mathbb{T}$ is exponentially stable. □

We leave it to the reader to formulate the dual version of this proposition.

6. Dynamic stabilization. In the first part of this section we prove that optimizability and estimatability are invariant under feedback. In what follows, U, X, Y, and R are Hilbert spaces.

LEMMA 6.1. *Let A be the generator of a strongly continuous semigroup $\mathbb{T}$ on X and let $B \in \mathcal{L}(U, X_{-1})$ be an admissible control operator for $\mathbb{T}$. Assume that $x_0 \in X$ and $u \in L^2([0,\infty), U)$ are such that $x \in L^2([0,\infty), X)$, where x is the state trajectory defined by* (3.1). *Then for every $\tau > 0$, we have that*

$$\sum_{n=1}^{\infty} \|x(n\tau)\|^2 < \infty .$$

Proof. We define the sequences $u_n \in L^2([0,\tau], U)$ and $x_n \in L^2([0,\tau], X)$ by $u_n(t) = u((n-1)\tau + t)$ and $x_n(t) = x((n-1)\tau + t)$, so that the sequences $\|u_n\|$ and $\|x_n\|$ are square summable. Using the notation from (2.2) and (2.6), we have that for every $n \in \mathbb{N}$ and every $t \in [0,\tau]$,

$$x(n\tau) = \mathbb{T}_t x_n(\tau - t) + \Phi_t \mathbf{S}^*_{\tau - t} u_n .$$

Denoting $m_1 = \sup_{t\in[0,\tau]} \|\mathbb{T}_t\|$ and $m_2 = \|\Phi_\tau\|$ (so that $\|\Phi_t\| \le m_2$ for all $t \in [0,\tau]$), we have that

$$\|x(n\tau)\| \le m_1 \|x_n(\tau - t)\| + m_2 \|u_n\| ,$$

so that, denoting $m = m_1^2 + m_2^2$, $\|x(n\tau)\|^2 \le m \left(\|x_n(\tau - t)\|^2 + \|u_n\|^2\right)$. Integrating this inequality with respect to $t \in [0,\tau]$, we obtain

$$\tau \|x(n\tau)\|^2 \le m \|x_n\|^2 + m\tau \|u_n\|^2 .$$

This shows that the sequence $\|x(n\tau)\|$ is square summable. □

LEMMA 6.2. *Let Σ be a well-posed linear system with input space U, state space X, and output space Y. We denote by $\mathbb{T}$ the semigroup of Σ and by B its control operator. Assume that $x_0 \in X$ and $u \in L^2([0,\infty), U)$ are such that $x \in L^2([0,\infty), X)$, where x is the state trajectory defined by* (3.1). *Let y be the corresponding output function, given by* (2.7). *Then $y \in L^2([0,\infty), Y)$.*

Proof. Take $\tau > 0$. We define the sequences $u_n \in L^2([0,\tau], U)$ and $y_n \in L^2([0,\tau], Y)$ by $u_n(t) = u((n-1)\tau + t)$ and $y_n(t) = y((n-1)\tau + t)$. We denote by $\mathbf{P}_\tau$ the projection from $L^2_{loc}([0,\infty), Y)$ to $L^2([0,\tau], Y)$ (by truncation). It follows from (2.9) that (with the notation from (2.7))

$$y_n = \mathbf{P}_\tau \Psi x((n-1)\tau) + \mathbf{P}_\tau \mathbb{F} u_n .$$

Since $\mathbf{P}_\tau \Psi$ and $\mathbf{P}_\tau \mathbb{F}$ are bounded operators, there is a $k > 0$ such that

$$\|y_n\| \le k \left(\|x((n-1)\tau)\| + \|u_n\|\right) .$$

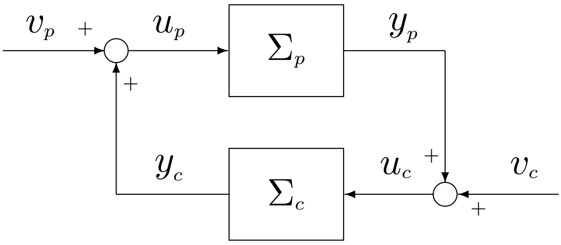


FIG. 3. *A plant Σ_p with a stabilizing controller Σ_c. The closed-loop system is well-posed and exponentially stable.*

By the assumption and by Lemma 6.1, both sequences $\|u_n\|$ and $\|x((n-1)\tau)\|$ are square summable, so that $\|y_n\|$ is square summable. □

THEOREM 6.3. *Let Σ be a well-posed linear system, let K be an admissible feedback operator for Σ, and let Σ^K be the corresponding closed-loop system (see Figure* 1*). We denote by A the semigroup generator of Σ, by B the control operator of Σ, and by C the observation operator of Σ. We denote by A^K, B^K, and C^K the corresponding operators for Σ^K. Then the following holds:*

(a) *(A, B) is optimizable if and only if (A^K, B^K) is optimizable.*

(b) *(A, C) is estimatable if and only if (A^K, C^K) is estimatable.*

Proof. We denote by U, X, and Y the input, state, and output space of Σ (and also of Σ^K). First we prove statement (a). Suppose that (A^K, B^K) is optimizable, let $x_0 \in X$ be an initial state for Σ^K, and let $v \in L^2([0,\infty), U)$ be an input function which causes the state trajectory of Σ^K to be in $L^2([0,\infty), X)$. Let y be the corresponding output function of Σ^K. We know from Lemma 6.2 that $y \in L^2([0,\infty), Y)$. If u is the corresponding input function of Σ (which causes the same state trajectory and the same output function in Σ) then $u = v + Ky$ (see Figure 1). Thus, $u \in L^2([0,\infty), U)$, so that (A, B) is optimizable. To prove the converse direction, assuming that (A, B) is optimizable, we follow the same argument but we regard Σ as a closed-loop system obtained from Σ^K via the output feedback through $-K$.

Now we prove statement (b). We introduce the dual systems Σ^d and Σ^{Kd} (see section 6 of [41] or [39]). Then Σ^{Kd} is the closed-loop system obtained from Σ^d via the output feedback K^*. The semigroup generator of Σ^d is A^*, its control operator is C^*, and its observation operator is B^*. A similar characterization applies to Σ^{Kd}. Now statement (b) is equivalent to statement (a) applied to these dual systems. □

We mention that, with the notation of the last theorem, for every $x_0 \in \mathcal{D}(A^K)$ and for every $z_0 \in \mathcal{D}(A)$,

$$A^K x_0 = \left(A + BKC^K\right) x_0\,, \qquad A z_0 = \left(A^K - B^K KC\right) z_0\,.$$

For the proof of these formulas and for further details we refer to [37].

We say that a well-posed linear system is *optimizable* if the corresponding pair (A, B) is optimizable. The meaning of a system being *estimatable* is similar.

Let Σ_p be a well-posed linear system. A *stabilizing controller* for Σ_p is another well-posed linear system Σ_c such that the interconnection shown in Figure 3 is well-posed and the closed-loop system is exponentially stable. The input signals of the closed-loop system are v_p and v_c and the output signals are y_p and y_c.

This is the framework for dynamic stabilization in Chapters 4, 5, and 9 of Curtain and Zwart [12] and many earlier references, where the systems are assumed to have bounded B and C operators. A rather different concept of dynamic control can be found in Russell [25], to model “indirect damping.” The original system (more precisely, its semigroup generator) is embedded in (the semigroup generator of) a larger, coupled system. The input and output signals are not restricted to be in L^2_{loc}, like they are in our framework. This approach is applied in [25] to modelling thermal effects in vibrating systems.

The concept of a stabilizing controller as in Figure 3, while much used and intuitively appealing, is too narrow: it does not admit all the observer-based controllers for Σ_p, and the Youla parametrization of such stabilizing controllers is not clean, requiring extra invertibility and well-posedness conditions. This is the case even for finite-dimensional plants Σ_p if they are not strictly proper. To overcome these difficulties, the following generalization was introduced in Weiss and Curtain [40] (and was further investigated in Curtain, Weiss, and Weiss [10] and [11]).

A *stabilizing controller with internal loop* for Σ_p is a well-posed linear system Σ_k with two inputs and two outputs such that the interconnection shown in Figure 4 is a well-posed linear system $\Sigma_{p,k}$ and this system is exponentially stable. The inputs of the *closed-loop system* $\Sigma_{p,k}$ are the three external signals going to the summation points, and the outputs are the outputs of the two subsystems.

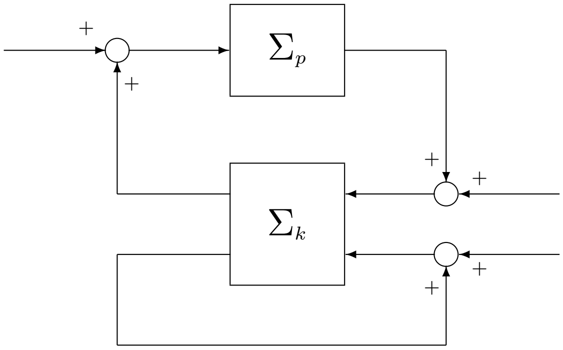


FIG. 4. *A plant Σ_p with a stabilizing controller with internal loop Σ_k. Again, the closed-loop system is well-posed and exponentially stable. Closing only one of the two loops may lead to a non-well-posed system.*

Each of the signals in Figure 4 may be Hilbert space-valued, of course. The lower loop in Figure 4 is referred to as the *internal loop* of the controller. Closing only one of the two loops may lead to a non-well-posed system. Thus, in particular, it may be impossible to close the internal loop in the absence of the plant Σ_p, in which case the controller with internal loop cannot be reduced to a usual stabilizing controller (as in Figure 3). Examples for this are given in section 6 of [40]. However, every stabilizing controller can be thought of as a stabilizing controller with internal loop (with no signal in the internal loop). We say that Σ_p is *dynamically stabilizable* if there exists a stabilizing controller with internal loop for Σ_p. Broadly speaking, Σ_p is dynamically stabilizable if it can be made into a subsystem of a stable system.

In Proposition 5.3 of [40], a list of three conditions were given which imply dynamic stabilizability for a regular linear system (a design procedure for Σ_k was also provided). The first two of these conditions are stabilizability and detectability. We do not know if these two conditions alone imply dynamic stabilizability, and we also do not know if dynamic stabilizability implies stabilizability and detectability. Some of the results from [40] were generalized to well-posed linear systems by Staffans [30].

For systems with bounded B and C, dynamic stabilizability is equivalent to stabilizability by a bounded F and detectability by a bounded H (see Exercise 6.13 in

Curtain and Zwart [12]) and no internal loop is needed in this case. For systems with bounded control and observation operators, stabilizability and detectability in the general sense which we are using are equivalent to stabilizability and detectability in the more restrictive sense of [12], i.e., with bounded F and H. This follows from the results in section 3 (in particular, formula (3.5)) and duality.

Now we look at dynamic stabilization from the point of view of transfer functions. We denote by $\mathbf{P}$ the transfer function of the plant Σ_p and by $\mathbf{K}$ the transfer function of the controller Σ_k. Thus, if the input and output spaces of Σ_p are U and Y, then $\mathbf{P}(s) \in \mathcal{L}(U, Y)$. We denote by R the Hilbert space where the signal of the internal loop takes values, so that $\mathbf{K} \in \mathcal{L}(Y \times R, U \times R)$. $\mathbf{K}$ is naturally partitioned into $\mathbf{K}_{11}$, $\mathbf{K}_{12}$, $\mathbf{K}_{21}$, and $\mathbf{K}_{22}$. We introduce Σ_l, the *parallel connection* of Σ_p and Σ_k, which is a well-posed linear system with two components, Σ_p and Σ_k, operating independently (see also section 4 of [40]). The transfer function of Σ_l is

$$\mathbf{L} = \begin{bmatrix} 0 & \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{P} & 0 & 0 \\ 0 & \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}. \tag{6.1}$$

The closed-loop system $\Sigma_{p,k}$ from Figure 4 is obtained from Σ_l via the output feedback operator I (acting on $U \times Y \times R$). Thus, its transfer function is $\mathbf{L}(I - \mathbf{L})^{-1} = (I - \mathbf{L})^{-1} - I$. The following result, a restatement of Theorem 1.2, is a strengthened version of Proposition 4.11 from [40]. (By this we mean that the result in [40] is an immediate consequence of the theorem below, but not conversely.)

THEOREM 6.4. *Let Σ_p and Σ_k be well-posed linear systems, with transfer functions $\mathbf{P}$ and $\mathbf{K}$, such that $\mathbf{P}(s) \in \mathcal{L}(U, Y)$ and $\mathbf{K}(s) \in \mathcal{L}(Y \times R, U \times R)$. After a proper partitioning of $\mathbf{K}$, we define $\mathbf{L}$ by* (6.1). *Then Σ_k is a stabilizing controller with internal loop for Σ_p if and only if the following two conditions hold:*

(1) *Both Σ_p and Σ_k are optimizable and estimatable.*

(2) *$(I - \mathbf{L})^{-1} \in H^\infty$ (i.e., the closed-loop system is input-output stable).*

Proof. First assume that Σ_k is a stabilizing controller with internal loop for Σ_p, i.e., the closed-loop system $\Sigma_{p,k}$ from Figure 4 is exponentially stable. As explained earlier, its transfer function is $(I - \mathbf{L})^{-1} - I$. Since exponential stability implies input-output stability, we have that $(I - \mathbf{L})^{-1} \in H^\infty$. The system $\Sigma_{p,k}$ is obtained from the parallel connection Σ_l (introduced before (6.1)) via the output feedback operator I. Since $\Sigma_{p,k}$ is obviously optimizable and estimatable, by Theorem 6.3 these properties are shared also by Σ_l. Because of the nature of Σ_l, this implies that Σ_p and Σ_k are optimizable and estimatable.

Conversely, assume that (1) and (2) hold. Then (1) implies that Σ_l is optimizable and estimatable. By Theorem 6.3, $\Sigma_{p,k}$ is also optimizable and estimatable. By (2), $\Sigma_{p,k}$ is input-output stable. By Theorem 5.3, $\Sigma_{p,k}$ is exponentially stable. □

REMARK 4. *It follows from Theorems* 5.3 *and* 6.4 *that a well-posed linear system is exponentially stable if and only if it is dynamically stabilizable and input-output stable. This is stated in Corollary* 1.3 *and it will be useful in the examples in section* 7. *Hence, if an unstable system (i.e., one that is not exponentially stable) has its transfer function in H^∞, then it cannot be dynamically stabilizable.*

REMARK 5. *If for some $\alpha \geq 0$ we have that $\mathbf{G} \in H^\infty_\alpha$ (defined in section* 5*), but for all $\varepsilon > 0$, $\mathbf{G}$ cannot be extended to a function in $H^\infty_{\alpha-\varepsilon}$, then any system with transfer function $\mathbf{G}$ is not dynamically stabilizable. For $\alpha = 0$, the proof goes as follows: If a system Σ with transfer function $\mathbf{G}$ were dynamically stabilizable, then by Theorem* 6.4 *it would be optimizable and estimatable. Now Theorem* 5.3 *would imply* *that Σ is exponentially stable, and this would imply (see* [36]*) that $\mathbf{G} \in H^\infty_{-\varepsilon}$ for some $\varepsilon > 0$. To extend this argument to $\alpha > 0$, we have to use shifted systems, as in the proof of Corollary* 5.4*. For example, a system whose transfer function is $e^{-\frac{1}{s}}$, or a Blaschke product with zeros arbitrarily close to the imaginary axis (as described at the end of section* 1*), cannot be dynamically stabilizable.*

7. Examples.

Example 1. In this example we illustrate Corollary 1.3 (Remark 4) with two results about a model for elastic structure/acoustics interaction. These results are improvements on results in Avalos, Lasiecka, and Rebarber [2].

Let Ω be either a rectangular region in $\mathbb{R}^2$ or a region in $\mathbb{R}^2$ with Lipschitz boundary Γ. Let Γ_0 be a smooth (C^2) segment of Γ with endpoints a and b, called the active boundary. Let $z = z(t,\zeta)$ for $t \in [0,\infty)$ and $\zeta \in \Omega$, let $v = v(t,\xi)$ for $t \in [0,\infty)$ and $\xi \in \Gamma_0$, and let $\partial/\partial\nu$ denote the outward normal derivative to Γ. Let $U = \mathbb{R}^r$ and $\mathcal{B} \in \mathcal{L}(U, H^{-\alpha}(\Gamma_0))$, where $\alpha = 7/4$ when Ω is rectangular, and $\alpha = 5/3$ when Ω has a smooth boundary. In applications Ω is a cross section of an acoustic cavity, z is the acoustic velocity potential, and v is the normal displacement of the active wall. The following model has been studied extensively in recent years—see Banks and Smith [6] for a discussion of the modelling:

$$\begin{aligned} z_{tt} &= \Delta z \quad \text{on } [0,\infty)\times\Omega\,,\\ \frac{\partial z}{\partial\nu} &= v_t \quad \text{on } [0,\infty)\times\Gamma_0\,,\\ \frac{\partial z}{\partial\nu} &= 0 \quad \text{on } [0,\infty)\times\Gamma\setminus\Gamma_0\,,\\ v_{tt} &= -\Delta^2 v - \Delta^2 v_t - z_t + \mathcal{B}u \quad \text{on } [0,\infty)\times\Gamma_0\,,\\ v(a,t) = v(b,t) &= v_x(a,t) = v_x(b,t) = 0 \quad \text{for } t\in[0,\infty)\,. \end{aligned} \tag{7.1}$$

We consider three natural quantities that can be observed for (7.1): the active boundary displacement v, the active boundary velocity v_t, and the acoustic velocity z_t, which is proportional to the acoustic pressure. The observation of these quantities is typically taken at points on Γ_0 for v and v_t, and at points in $\overline{\Omega}$ for z_t. The output signal may be a vector of such measurements.

Define the state space

$$X := H^1(\Omega)/\mathbb{R} \times L^2(\Omega)/\mathbb{R} \times H^2_0(\Gamma_0) \times L^2(\Gamma_0),$$

where $H^r(\Omega)/\mathbb{R} = \{f \in H^r(\Omega) \mid \int_\Omega f\,dx = 0\}$ for $r > 0$. Letting

$$x(t) = [z(t,\cdot), z_t(t,\cdot), v(t,\cdot), v_t(t,\cdot)]^T,$$

it is shown in Avalos and Lasiecka [1] that (7.1) can be put in the form

$$\dot{x}(t) = Ax(t) + Bu(t).$$

Here, A generates a strongly stable but not exponentially stable semigroup $\mathbb{T}$ on X, and B is an admissible control operator for $\mathbb{T}$. If $Y = \mathbb{R}^{k+j}$ and

$$y(t) = C_1 x(t) := [v(\alpha_1),\ldots,v(\alpha_j), v_t(\beta_1),\ldots,v_t(\beta_k)]^T \tag{7.2}$$

with $\alpha_i, \beta_i \in \Gamma_0$, then it is shown in [2] that (7.1), (7.2) is a regular linear system with feedthrough operator 0. In Theorem 5.9 of [2] it is shown that if there exists a stabilizing controller for (7.1), (7.2), then the stabilization is not robust with respect to delays in the feedback loop. In this context, when we say that stabilization is "not robust with respect to delays," we mean that there exist sequences $\{\varepsilon_n\}$ and $\{p_n\}$, with $\varepsilon_n > 0$, $\varepsilon_n \to 0$, and $p_n \in \mathbb{C}_0$ such that if a delay of length ε_n is introduced into the feedback loop, then the closed-loop transfer function has a pole at p_n; for details see [2] or Logemann, Rebarber, and Weiss [18]. This result in [2] uses Remark 4 from this paper, and "stabilizing controller" can be easily replaced with the more general "stabilizing controller with internal loop."

In Remark 5.11 of [2] it is mentioned that it might not even be possible to dynamically stabilize (7.1), (7.2). Using a recent result from Avalos, Lasiecka, and Rebarber [4], we prove that this lack of stabilization is indeed the case.

THEOREM 7.1. *Let* $Y = H_0^2(\Gamma_0) \times H_0^2(\Gamma_0)$ *and define the observation*

$$y(t) = [\,v(t),\, v_t(t)\,]^T . \tag{7.3}$$

Then (7.1)*,* (7.3) *cannot be dynamically stabilized.*

Proof. It is shown in [4] that if the initial state for (7.1) is zero, then there exists an $M > 0$ such that

$$\int_0^\infty \left(\|v(t,\cdot)\|^2_{H_0^2(\Gamma_0)} + \|v_t(t,\cdot)\|^2_{H_0^2(\Gamma_0)} \right) dt \le M \int_0^\infty \|u(t)\|_U^2 \, dt.$$

Hence (7.1), (7.3) is input-output stable. Since $\mathbb{T}$ is not exponentially stable, by Remark 4 this system is not dynamically stabilizable. □

COROLLARY 7.2. *The system* (7.1)*,* (7.2) *is not dynamically stabilizable.*

Proof. Clearly, the output signal in (7.2) is obtained by applying a (static) bounded operator to y from (7.3). Using Theorem 7.1, we see that (7.1), (7.2) is not dynamically stabilizable. □

Since it is impossible to stabilize (7.1) with observations taken only along the beam component, we turn our attention to output signals which include point observations of the acoustic pressure. Let $Y = \mathbb{R}^{j+k+l}$ and

$$C_2 x(t) = [v(\alpha_1), \ldots, v(\alpha_j), v_t(\beta_1), \ldots, v_t(\beta_k), z_t(\zeta_i), \ldots, z_t(\zeta_l)]^T \tag{7.4}$$

for $\alpha_i, \beta_i \in \Gamma_0$, and $\zeta_i \in \overline{\Omega}$. It is easy to verify that C_2 is not admissible for $\mathbb{T}$ when the state space is X, and for this reason, in [2] the lack-of-robustness result was only given in the input-output setting. In particular, it was shown that if the open-loop transfer function for (7.1), (7.4) is not stable, and there exists a controller (without internal loop) which input-output stabilizes (7.1), (7.4), then this stabilization is not robust with respect to delays—see Theorem 5.12 in [2] for details. Using recent results from Avalos, Lasiecka, and Rebarber [3], we can put this lack of robustness into a state space setting. We use the following result, which also appears in a slightly less general form in [2] (where it was also a corollary of Remark 4).

COROLLARY 7.3. *Suppose that* Σ_p *is a regular linear system with semigroup generator* A*, input space* U *and output space* Y*,* U*, and* Y *are finite-dimensional,* $\sigma(A)$ *is contained in the open left half-plane but* A *is not exponentially stable.*

If there exists a regular stabilizing controller Σ_c *for* Σ_p *(as in Figure* 3*), then the stability of the closed-loop system is not robust with respect to delays.*

Proof. Remark 4 shows that if there exists a stabilizing controller for Σ_p, then $\mathbf{P}$ is unbounded on $\mathbb{C}_0$, where $\mathbf{P}$ is the transfer function of Σ_p. Since $(sI - A)^{-1}$ is analytic on a set containing the closed right half-plane, the same is true for $\mathbf{P}$, in particular, $\mathbf{P}$ is continuous on the closed right half-plane. This, together with its unboundedness implies that

$$\limsup_{|s|\to\infty,\, s\in\mathbb{C}_0} \|\mathbf{P}(s)\|_{\mathcal{L}(X)} = \infty .$$

Now we can apply Theorem 8.5 from [18] to conclude that if the closed-loop system is stable, then its stability is not robust with respect to delays. $\Box$

REMARK 6. *If the feedthrough operator of Σ_p is zero, then Corollary* 7.3 *is true even if Σ_c is not regular. This is because* $\mathbf{PC}$ *will still be regular, where* $\mathbf{C}$ *is the transfer function of Σ_c.*

In [2] this result was not applied to (7.1), (7.4) because at the time that paper was written no natural state space was identified for this system. In Avalos, Lasiecka, and Rebarber [3] a state space is given so that the natural (A, B, C) representation for (7.1), (7.4) is a regular system. This state space is

$$X := \left\{ [z_0, z_1, v_0, v_1] \in \frac{H^{\frac{3}{2}}(\Omega)}{\mathbb{R}} \times \frac{H^{\frac{1}{2}}(\Omega)}{\mathbb{R}} \times H_0^2(\Gamma_0) \times L^2(\Gamma_0), \;\middle|\; z_0 - Nv_1 \in \mathcal{H} \right\},$$

where N is the Neumann map and

$$\mathcal{H} = \left\{ f \in \frac{H^{\frac{3}{2}}(\Omega)}{\mathbb{R}} \;\middle|\; \nabla f \in L^2_{-\frac{1}{2}}(\Omega) \right\}.$$

Here, $L^2_{-\frac{1}{2}}(\Omega)$ denotes the space of functions h on Ω such that $h\varrho^{-\frac{1}{2}} \in L^2(\Omega)$, with $\varrho(x)$ being the distance from x to Γ; see [3] for details.

We can apply Corollary 7.3 to (7.1), (7.4) to conclude the following.

COROLLARY 7.4. *Suppose Σ_p is stabilized (in the state space X given above) by a stabilizing controller Σ_c. Then this stability is not robust with respect to delays.*

Example 2. We consider the linear infinite-dimensional system described by the wave equation on an n-dimensional domain, with mixed boundary control and mixed boundary observation. The bounded domain $\Omega \subset \mathbb{R}^n$ is assumed to have a C^2 boundary Γ, and Ω is locally on one side of Γ. Γ_0 and Γ_1 are nonempty open subsets of Γ such that $\Gamma_0 \cap \Gamma_1 = \emptyset$ and $\overline{\Gamma_0 \cup \Gamma_1} = \Gamma$. We denote by x the space variable $(x \in \Omega)$. The equations of the system are

$$\left\{ \begin{array}{ll} w_{tt}(x,t) = \Delta w(x,t) & \text{on } \Omega \times [0,\infty), \\ w(x,t) = 0 & \text{on } \Gamma_0 \times [0,\infty), \\ \frac{\partial}{\partial\nu} w(x,t) + w_t(x,t) = u(x,t) & \text{on } \Gamma_1 \times [0,\infty), \\ \frac{\partial}{\partial\nu} w(x,t) - w_t(x,t) = y(x,t) & \text{on } \Gamma_1 \times [0,\infty), \\ w(x,0) = w_1(x), \; w_t(x,0) = w_2(x) & \text{on } \Omega, \end{array} \right. \tag{7.5}$$

where u is the input function, and y is the output function. The functions w_1 and w_2 are the initial state of the system. The part Γ_0 of the boundary is just reflecting waves, while the active portion Γ_1 is where both the observation and the control take place. We shall often write $w(t)$ to denote a function of x, meaning that $w(t)(x) = w(x,t)$, and similarly for other functions. The state of the system is

$$z(t) = \begin{bmatrix} w(t) \\ \dot{w}(t) \end{bmatrix}.$$

We introduce the input space U and the state space X by

$$U = L^2(\Gamma_1), \qquad X = H^1_{\Gamma_0}(\Omega) \times L^2(\Omega), \tag{7.6}$$

where $H^1_{\Gamma_0}(\Omega) = \{v \in H^1(\Omega) \mid v|_{\Gamma_0} = 0\}$. The output space is also U. The norm on X is defined by

$$\left\| \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \right\|_X^2 = 2\, \|\nabla w_1\|^2_{L^2(\Omega)} + 2\, \|w_2\|^2_{L^2(\Omega)}. \tag{7.7}$$

The equations (7.5) are well-posed, i.e., they determine a well-posed linear system Σ with the input, state, and output spaces indicated. This can be shown using the general theory of well-posed boundary control systems as developed in Salamon [28]. This well-posedness result, formulated in a different terminology (and for a different but closely related system) is due to Rodriguez-Bernal and Zuazua [24].

The above system is studied in detail in Avalos and Weiss [5], which contains also the proofs of all the other claims that we make in this example. An important fact is that Σ is *conservative.* This means that for every $\tau > 0$,

$$\|z(\tau)\|^2 + \int_0^\tau \|y(t)\|^2\, dt = \|z(0)\|^2 + \int_0^\tau \|u(t)\|^2\, dt, \tag{7.8}$$

and a similar equality holds for the dual system. We may think of u as the "incoming wave" (which brings energy into the system) and of y as the "outgoing wave."

We now introduce several spaces and operators, mainly following Triggiani [32]. More related information can be found in Rodriguez-Bernal and Zuazua [24]. We define the self-adjoint and positive $\Lambda : D(\Lambda) \subset L^2(\Omega) \rightarrow L^2(\Omega)$ by

$$\Lambda w = -\Delta w, \quad D(\Lambda) = \left\{ w \in H^2(\Omega) \;\middle|\; w|_{\Gamma_0} = 0,\ \frac{\partial}{\partial \nu} w|_{\Gamma_1} = 0 \right\}.$$

Then Λ is boundedly invertible and $D(\Lambda^{\frac{1}{2}}) = H^1_{\Gamma_0}(\Omega)$. From Green's theorem and continuous extension we have the middle equality in

$$\|w\|^2_{H^1_{\Gamma_0}(\Omega)} = \int_\Omega |\nabla w|^2 dx = \left\| \Lambda^{\frac{1}{2}} w \right\|^2_{L^2(\Omega)} = \|w\|^2_{D(\Lambda^{\frac{1}{2}})}.$$

The other two equalities above hold by definition. We denote by $H^{-1}_{\Gamma_0}(\Omega)$ the dual of $H^1_{\Gamma_0}$ with respect to the pivot space $L^2(\Omega)$. Then Λ has an extension to a bounded operator $\Lambda : H^1_{\Gamma_0}(\Omega) \rightarrow H^{-1}_{\Gamma_0}(\Omega)$.

We define the Neumann map $N : L^2(\Gamma_1) \rightarrow L^2(\Omega)$ by $Nf = g$ if and only if

$$\Delta g = 0, \quad g|_{\Gamma_0} = 0, \quad \frac{\partial}{\partial \nu} g|_{\Gamma_1} = f.$$

By elliptic theory we have that $N \in \mathcal{L}(L^2(\Gamma_1), H^1_{\Gamma_0}(\Omega))$.

The Dirichlet trace γ_0 satisfies for any continuous function f on Ω

$$\gamma_0 f = f|_{\Gamma_1}.$$

After continuous extension we have that $\gamma_0 \in \mathcal{L}(H^1(\Omega), H^{\frac{1}{2}}(\Gamma_1))$.

The Neumann trace γ_1 satisfies for any C^1 function f on Ω

$$\gamma_1 f = \frac{\partial}{\partial \nu} f|_{\Gamma_1}.$$

Thus, γ_1 is the outward normal derivative restricted to Γ_1. We extend γ_1 to the space of all functions $w \in H^1_{\Gamma_0}(\Omega)$ which satisfy $\Delta w \in L^2(\Omega)$, and then $\gamma_1 w \in H^{-\frac{1}{2}}(\Gamma_1)$. By Green's theorem and by continuous extension, we have that for all $w \in D(\Lambda^{\frac{1}{2}})$,

$$N^*\Lambda w = w|_{\Gamma_1} = \gamma_0 w. \tag{7.9}$$

We define $A : D(A) \subset X \to X$ by

$$A = \begin{bmatrix} 0 & I \\ -\Lambda & -\Lambda N N^* \Lambda \end{bmatrix},$$

$$D(A) = \left\{ \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \begin{matrix} D(\Lambda^{\frac{1}{2}}) \\ \times \\ D(\Lambda^{\frac{1}{2}}) \end{matrix} \;\middle|\; w_1 + NN^*\Lambda w_2 \in D(\Lambda) \right\}.$$

Note that we have the equivalent characterization

$$A \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} w_2 \\ \Delta w_1 \end{bmatrix}, \tag{7.10}$$

$$D(A) = \left\{ \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \begin{matrix} H^1_{\Gamma_0} \\ \times \\ H^1_{\Gamma_0} \end{matrix} \;\middle|\; \begin{matrix} \Delta w_1 \in L^2(\Omega) \\ \\ \gamma_1 w_1 + \gamma_0 w_2 = 0 \end{matrix} \right\}. \tag{7.11}$$

Then A generates a strongly continuous semigroup of contractions on X, which we denote by $\mathbb{T}$. The resolvent of A is compact and $\sigma(A)$ is in the open left half-plane. All these facts about A are proved in [32] and in [24], which also contain relevant earlier references. This A is the semigroup generator of the system Σ.

The control operator B and the observation operator C of Σ are given by

$$B = \begin{bmatrix} 0 \\ \Lambda N \end{bmatrix} \qquad \text{and} \qquad C = \frac{1}{\sqrt{2}} \begin{bmatrix} \gamma_1 & -\gamma_0 \end{bmatrix}.$$

This is proved in Avalos and Weiss [5]. It follows from (7.9) and (7.11) that C can be rewritten in the form $C = \begin{bmatrix} 0 & -2N^*\Lambda \end{bmatrix}$.

The system Σ is regular and its feedthrough operator is $D = 0$. Hence, the transfer function of Σ is $\mathbf{G}(s) = C_\Lambda(sI - A)^{-1}B$, which is analytic with values in $\mathcal{L}(U)$. The fact that Σ is conservative implies that $\mathbf{G} \in H^\infty$ and moreover, $\mathbf{G}(i\omega)$ is a unitary operator for each $\omega \in \mathbb{R}$. In particular, it is easy to check that $\mathbf{G}(0) = I$.

It is not difficult to see that every conservative system is exactly observable if and only if it is exponentially stable. It is shown in [5] that our system Σ is isomorphic to its dual. Thus, Σ is exactly controllable if and only if it is exactly observable if and only if it is exponentially stable. These properties all hold if Γ_1 is sufficiently large, in an appropriate sense. A sharp characterization of such sets Γ_1 was given in Bardos, Lebeau, and Rauch [7]. Earlier, sufficient conditions for Γ_1 to be large enough (in the above sense) were given, for example, in Lasiecka and Triggiani [17].

If the active boundary Γ_1 is "too small," then Σ is not exponentially stable. However, it remains input-output stable ($\mathbf{G} \in H^\infty$) regardless of Γ_1. Then it follows from Remark 4 that the system is not dynamically stabilizable. Thus, the wave equation with the control and observation as in (7.5) is either exponentially stable to begin with, or it is not stabilizable by any controller.

REFERENCES

[1] G. Avalos and I. Lasiecka, *A differential Riccati equation for the active control of a problem in structural acoustics,* J. Optim. Theory Appl., 91 (1996), pp. 695–728.

[2] G. Avalos, I. Lasiecka, and R. Rebarber, *Lack of time-delay robustness for stabilization of a structural acoustics model,* SIAM J. Control Optim., 37 (1999), pp. 1394–1418.

[3] G. Avalos, I. Lasiecka, and R. Rebarber, *Well-posedness of a structural acoustics control model with point observation of the pressure*, J. Differential Equations, to appear.

[4] G. Avalos, I. Lasiecka, and R. Rebarber, *Uniform decay properties of a model in structural acoustics*, J. Math. Pures Appl., to appear.

[5] G. Avalos and G. Weiss, *The Wave Equation as a Conservative Regular Linear System,* in preparation.

[6] H.T. Banks and R.C. Smith, *Feedback control of noise in a 2-D nonlinear structural acoustics model,* Discrete Contin. Dynam. Systems, 1 (1995), pp. 119–149.

[7] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary,* SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[8] R.F. Curtain, *Equivalence of input-output stability and exponential stability for infinite-dimensional systems,* Math. Systems Theory, 21 (1988), pp. 19–48.

[9] R.F. Curtain, *Equivalence of input-output stability and exponential stability*, Systems Control Lett., 12 (1989), pp. 235–239.

[10] R.F. Curtain, G. Weiss, and M. Weiss, *Coprime factorization for regular linear systems,* Automatica, 32 (1996), pp. 1519–1531.

[11] R.F. Curtain, G. Weiss, and M. Weiss, *Stabilization of irrational transfer functions by controllers with internal loop,* in Proceedings of the International Workshop on Operator Theory and Applications, June 2000, Bordeaux, France, Université of Bordeaux 1, France.

[12] R.F. Curtain and H.J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[13] F. Flandoli, I. Lasiecka, and R. Triggiani, *Algebraic Riccati equations with non-smoothing observation arising in hyperbolic and Euler-Bernoulli boundary control problems,* Ann. Mat. Pura Appl., 153 (1988), pp. 307–382.

[14] P. Grabowski and F. Callier, *Admissible observation operators. Semigroup criteria of admissibility,* Integral Equations Operator Theory, 25 (1996), pp. 182–198.

[15] S. Hansen and G. Weiss, *New results on the operator Carleson measure criterion,* IMA J. Math. Control Inform., 14 (1997), pp. 3–32.

[16] B. Jacob and H. Zwart, *Equivalent conditions for stabilizability of infinite-dimensional systems with admissible control operators,* SIAM J. Control and Optim., 37 (1999), pp. 1419–1455.

[17] I. Lasiecka and R. Triggiani, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometric conditions,* Appl. Math. Optim., 25 (1992), pp. 189–224.

[18] H. Logemann, R. Rebarber, and G. Weiss, *Conditions for robustness and nonrobustness of the stability of feedback systems with respect to small delays in the feedback loop,* SIAM J. Control and Optim., 34 (1996), pp. 572–600.

[19] K.A. Morris, *Justification of input/output methods for systems with unbounded control and observation,* IEEE Trans. Automat. Control, 44 (1999), pp. 81–85.

[20] J. Prüss, *On the spectrum of C_0-semigroups,* Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.

[21] R. Ravi and P.P. Khargonekar, *Exponential and input-output stability are equivalent for linear time-varying systems,* Sādhanā 18 (1993), pp. 31–37.

[22] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems,* IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.

[23] R. Rebarber and H.J. Zwart, *Open-loop stabilizability of infinite-dimensional systems,* Math. Control Signals Systems, 11 (1998), pp. 129–160.

[24] A. Rodríguez-Bernal and E. Zuazua, *Parabolic singular limit of a wave equation with localized boundary damping,* Discrete Contin. Dynam. Systems, 1 (1995), pp. 303–346.

[25] D.L. Russell, *A general framework for the study of indirect damping mechanisms in elastic systems,* J. Math. Anal. Appl., 173 (1993), pp. 339–358.

[26] D.L. Russell and G. Weiss, *A general necessary condition for exact observability,* SIAM J. Control Optim., 32 (1994), pp. 1–23.

[27] D. Salamon, *Realization theory in Hilbert space,* Math. Systems Theory, 21 (1989), pp. 147–164.

[28] D. Salamon, *Infinite dimensional systems with unbounded control and observation: A functional analytic approach,* Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.

[29] O.J. Staffans, *Quadratic optimal control of stable well-posed linear systems,* Trans. Amer. Math. Soc., 349 (1997), pp. 3679–3715.

[30] O.J. Staffans, *Coprime factorizations and well-posed linear systems,* SIAM J. Control Optim., 36 (1998), pp. 1268–1292.

[31] O.J. Staffans, *Quadratic optimal control of well-posed linear systems,* SIAM J. Control Optim., 37 (1999), pp. 131–164.

[32] R. Triggiani, *Wave equation on a bounded domain with boundary dissipation: An operator approach,* J. Math. Anal. Appl., 137 (1989), pp. 438–461.

[33] G. Weiss, *Admissibility of unbounded control operators,* SIAM J. Control Optim., 27 (1989), pp. 527–545.

[34] G. Weiss, *Weak L^p-stability of a linear semigroup on a Hilbert space implies exponential stability,* J. Differential Equations, 76 (1988), pp. 269–285.

[35] G. Weiss, *Admissible observation operators for linear semigroups,* Israel J. Math., 65 (1989), pp. 17–43.

[36] G. Weiss, *Transfer functions of regular linear systems, Part* I: *Characterizations of regularity,* Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[37] G. Weiss, *Regular linear systems with feedback,* Math. Control Signals Systems, 7 (1994), pp. 23–57.

[38] G. Weiss, *Two conjectures on the admissibility of control operators*, in Estimation and Control of Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser-Verlag, Basel, 1991, pp. 367–378.

[39] G. Weiss, O.J. Staffans, and F. Callier, *Transfer Functions of Regular Linear Systems, Part* II: *Inversions and Duality,* in preparation.

[40] G. Weiss and R.F. Curtain, *Dynamic stabilization of regular linear systems,* IEEE Trans. Automat. Control, 42 (1997), pp. 4–21.

[41] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems,* Math. Control Signals Systems, 10 (1997), pp. 287–330.

[42] H.J. Zwart, *Linear quadratic optimal control for abstract linear systems*, in Modelling and Optimization of Distributed Parameter Systems; Applications to Engineering, K. Malanowski, Z. Nahorski, and M. Peszynska, eds., Chapman & Hall, London, 1996.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1233–1295

A TWOFOLD SPLINE APPROXIMATION FOR FINITE HORIZON LQG CONTROL OF HEREDITARY SYSTEMS*

A. GERMANI†, C. MANES†, AND P. PEPE†

Abstract. In this paper an approximation scheme is developed for the solution of the linear quadratic Gaussian (LQG) control on a finite time interval for hereditary systems with multiple noncommensurate delays and distributed delay. The solution here proposed is achieved by means of two approximating subspaces: the first one to approximate the Riccati equation for control and the second one to approximate the filtering equations. Since the approximating subspaces have finite dimension, the resulting equations can be implemented. The convergence of the approximated control law to the optimal one is proved. Simulation results are reported on a wind tunnel model, showing the high performance of the method.

Key words. hereditary systems, linear quadratic Gaussian regulator, infinite dimensional systems, Galerkin spline approximation

AMS subject classifications. 93E11, 93E20, 93E25

PII. S0363012998337461

1. Introduction.

It is well known that the solution of the linear quadratic regulation problem and of the optimal Gaussian filtering problem for linear delay systems is found in terms of infinite dimensional operators [7, 8, 9, 10, 11, 12, 13, 17, 23, 31, 36, 37, 39]. On the other hand, implementation of a control/filtering scheme in this case requires a finite dimensional approximation of such operators.

Although much attention has been devoted to separately developing an approximation theory for the linear quadratic (LQ) regulation [4, 11, 12, 16, 24, 26, 27, 30, 33, 40] and the optimal Gaussian filtering [14, 20] of delay systems, the approximation problem of the overall linear quadratic Gaussian (LQG) regulator has not been conveniently treated in the literature.

The averaging approximation scheme has been used in [24], for both the finite and infinite horizon LQ problem of delay systems, and convergence results are obtained by considering a conjecture, later proved to be true [41], that is the question of whether the sequence of approximating systems gives uniformly exponentially stable systems for sufficiently large indexes if the underlying retarded functional differential equation is stable.

The spline approximation scheme developed in [3] has been applied to the LQ problem of delay systems in [4]. Although numerical simulations show better performance than the averaging scheme, no theoretical convergence results are so far available. In [6], it is proved that the adjoint of the approximate semigroup governing the system does not converge in a strong way to the adjoint of such semigroup. As a consequence, the main hypothesis which guarantees the convergence results in [24] cannot be satisfied, and therefore this spline approximation scheme cannot be safely applied.

A new spline approximation scheme has been developed in [27], for the LQ prob-

*Received by the editors April 15, 1998; accepted for publication (in revised form) July 6, 2000; published electronically December 13, 2000. This work was supported by ASI (Agenzia Spaziale Italiana). Preliminary results have been published in [19, 21].

http://www.siam.org/journals/sicon/39-4/33746.html

†Dipartimento di Ingegneria Elettrica, Università degli Studi dell'Aquila, 67040 Poggio di Roio, L'Aquila, Italy (germani@ing.univaq.it, manes@ing.univaq.it, pepe@ing.univaq.it).

lem of delay systems with any number of pure delay terms, assuming the absolute continuity of the kernel in the distributed delay integral. Theoretical convergence results are obtained in the finite horizon case, as this approximation scheme does not guarantee the uniform exponential stability of the approximate semigroups. However, it is proved in [28, 29] that, in the case of commensurate delays and without distributed delay, a weaker condition is sufficient to obtain the strong convergence of the approximated LQ algebraic Riccati equation solution. The authors call this condition *uniform output stability.* It is proved that the spline approximation scheme developed in [27] does satisfy this condition, so that the above convergence result is available for the infinite horizon case. But, as pointed out by Morris on page 9 of paper [36], the convergence properties of this approximation scheme are not sufficient to ensure convergence of the closed loop response.

In [40], a piecewise linear approximation theory has been developed for the finite and infinite horizon LQ of general delay systems. Theoretical convergence results are obtained both in the finite and infinite horizon cases, as the condition of uniform exponential stability is verified.

In [32] error estimates are established for the approximation of delay systems by means of the averaging scheme. In [26] a scheme using first order splines is developed satisfying the uniform exponential stability condition, and error estimates are established too, as is done in [32] for the averaging scheme. Such a scheme uses the classic averaging subspace of piecewise constant functions to define the approximated system equation, but defines the approximated infinitesimal generator in that subspace not in the usual averaging methodology but by using an inverse projector from such subspace to the subspace built up using splines. Such a scheme, which is a mixed averaging spline one, is used in [26] for the infinite horizon LQ problem of general hereditary systems.

The matter of uniform exponential stability for spline approximation schemes has been investigated in [15], in the scalar open loop case. There the real eigenvalue (unique if the coefficient on delay term is positive, in the hereditary equation) of the infinitesimal generator of the semigroup governing the system is used, in order to define a particular inner product, by which Galerkin spline approximations [3] preserve the uniform exponential stability of the approximated semigroups. How this can be applied to optimal multivariables regulator problems is an open and interesting question.

In the synthesis of approximate optimal controllers developed by all above approximation schemes [4, 24, 26, 27, 40] it is assumed that the system state is completely accessible. Moreover, the approximated control input is generated by a finite rank feedback operator applied to the true state in the delay time interval. From an engineering point of view, the resulting controller is still infinite dimensional and therefore not directly implementable.

The synthesis of finite dimensional dynamic output feedback compensators for hereditary systems in a deterministic setting is considered in paper [25, section 4.2]. The proposed controller is composed of an observer and of a feedback control law from the observed state. Both the gains, for the finite dimensional observer and control, are obtained by approximating the solutions of two algebraic Riccati equations. The resulting controller resembles the solution of an LQG problem, although no reference to an optimal stochastic control problem is made in the paper. The main tool is the use of the averaging approximation scheme [2, 24] and the main result is the stability of the overall closed loop system.

In [35] the same problem is investigated with reference to a general class of deterministic distributed systems.

An approximation theory that provides an implementable scheme for the filtering problem of systems evolving on Hilbert spaces has been studied in [14, 18, 20, 22]. This theory has been successfully applied to delay systems.

In the literature the case with one pure delay term is usually completely reported [2, 24, 28, 29, 40] and the general case with multiple noncommensurate delays is usually just briefly indicated. However, the extension of all results to the general case is not straightforward [2, 3, 24, 40] or even unfeasible [28, 29].

As a final point of this bibliographic review, we must stress the existence of a large amount of spline approximation schemes [3, 4, 26, 27, 40] for the deterministic optimal quadratic state regulator (LQ problem), where the control gain operator is approximated by approximating the relevant Riccati equation. In principle, the same approximation schemes could be adapted for approximating the covariance operator defined by the solution of the dual Riccati equation, and the Kalman filter equation that solves the LQG problem in the stochastic setting. On the other hand, the applicability of such schemes to the case of stochastic delay systems with partial noisy state observations is not a trivial question and it has not been investigated up to now, and the main problem of proving the convergence remains unsolved.

The control problem with partial state observation has been treated in literature employing the averaging scheme in a deterministic setting [25]. On the other hand, a known result [4] is the superiority of spline approximation schemes with respect to averaging ones, with respect to numerical convergence rate.

On the basis of these considerations the aim of this paper is to define a finite dimensional scheme that approximates the solution of the finite horizon linear quadratic Gaussian control problem for stochastic delay systems with partial observations. The resulting implementable scheme has the following features:

(i) the optimal closed loop response of the LQG problem can be approximated with arbitrarily small error;

(ii) the scheme can be applied also in the LQ problem;

(iii) the approximation method is based on splines and not on averaging;

(iv) the matrices that implement the approximation of the optimal filter-controller scheme are easily parametrized as a function of the approximation order and can be easily computed;

(v) the scheme allows one to deal with general hereditary systems, that is, with multiple noncommensurate delays and distributed delay;

(vi) simultaneous approximation of a semigroup and of its adjoint is not required, so that problems arising from nondensity of the intersection of the respective generator domains are avoided;

(vii) the scheme allows a quite natural extension to be used for the solution of the infinite horizon LQG problem;

(viii) the scheme has nice numerical properties, in that it shows good performances even with a low finite dimensional approximation order;

(ix) the scheme allows one to get a faster convergence of the approximation by increasing the order of the spline degree.

Of course for most of the above-mentioned points, the scientific literature offers effective algorithms. Nevertheless, the problem of considering all these issues at the same time remains an interesting point.

The paper is organized as follows. In section 2 stochastic hereditary systems are written in state-space form and the infinitesimal generator of the adjoint of the semigroup that governs the system is studied. It is proved that such an operator has a deeply different structure if a weighted inner product is used instead of the usual one. In section 3 the finite horizon LQG is presented, and theorems for a suitable approximation scheme are proved. In section 4 an approximation scheme which satisfies hypotheses of section 3 is described for the general case. In section 5 matrices which represent finite dimensional linear operators are calculated to implement the method. In section 6 the infinite horizon case is addressed. In section 7 simulation results are reported, showing the effectiveness of the proposed method. Section 8 contains the conclusions.

2. Stochastic delay systems. In this paper we deal with the class of those dynamical systems that in technical literature are generally known as *linear delay systems*, sometimes also called *hereditary systems.* When state and observation noise are present, these are described, for $t \geq 0$, by stochastic equations of the type

$$\begin{aligned} \dot{\boldsymbol{z}}(t) &= \boldsymbol{A}_0\boldsymbol{z}(t) + \sum_{h=1}^{\delta} \boldsymbol{A}_h\boldsymbol{z}(t-r_h) \\ &\qquad + \int_{-r}^{0} \boldsymbol{A}_{01}(\vartheta)\boldsymbol{z}(t+\vartheta)d\vartheta + \boldsymbol{B}_0\boldsymbol{u}(t) + \boldsymbol{F}_0\boldsymbol{\omega}(t), \\ \boldsymbol{y}(t) &= \boldsymbol{C}_0\boldsymbol{z}(t) + \boldsymbol{G}\boldsymbol{\omega}(t) \end{aligned} \tag{2.1}$$

with $\boldsymbol{z}(t) \in \mathbb{R}^N$, $\boldsymbol{u}(t) \in \mathbb{R}^p$, $\boldsymbol{y}(t) \in \mathbb{R}^q$, $\boldsymbol{\omega}(t) \in \mathbb{R}^s$, $r_\delta = r > r_{\delta-1} > \cdots r_1 > r_0 = 0$, $\boldsymbol{A}_h \in \mathbb{R}^{N\times N}$, $\boldsymbol{A}_{01} \in L_2([-r,0];\mathbb{R}^{N\times N})$, $\boldsymbol{B}_0 \in \mathbb{R}^{N\times p}$, $\boldsymbol{C}_0 \in \mathbb{R}^{q\times N}$, $\boldsymbol{G} \in \mathbb{R}^{q\times s}$, $\boldsymbol{F}_0 \in \mathbb{R}^{N\times s}$.

The noise $\boldsymbol{\omega}$ belongs to the Hilbert space $L_2([0,t_f];\mathbb{R}^s)$ equipped with the standard Gaussian cylinder measure (this corresponds to model $\boldsymbol{\omega}$ as a white-noise process [1]). Independence of state and observation noises is assumed, that is, $\boldsymbol{F}_0\boldsymbol{G}^{\mathrm{T}} = \boldsymbol{0}$ and, without loss of generality, $\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}} = \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ denotes the identity matrix in $\mathbb{R}^{q\times q}$.

The variable $\boldsymbol{z}$ in the interval $[-r,0]$ is assumed to be generated as follows:

$$\boldsymbol{z}(\vartheta) = \bar{\boldsymbol{z}}(\vartheta) + \int_{-r}^{0} k(\vartheta,\tau)\bar{\boldsymbol{\omega}}(\tau)d\tau, \quad \vartheta \in [-r,0], \tag{2.2}$$

where $\bar{\boldsymbol{z}}$ is absolutely continuous with derivative in $L_2([-r,0];\mathbb{R}^N)$ and the process $\bar{\boldsymbol{\omega}}$, independent of $\boldsymbol{\omega}$, belongs to the Hilbert space $L_2([-r,0];\mathbb{R}^{\bar{s}})$ equipped with the standard Gaussian cylinder measure, and the kernel $k(\vartheta,\tau)$ is integrable for $\tau \in [-r,0]$.

As is well known, system (2.1) can be rewritten in state-space form in the Hilbert space $\boldsymbol{M}_2 = \mathbb{R}^N \times L_2([-r,0];\mathbb{R}^N)$, endowed with the following weighted inner product [3]:

$$\left(\begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix}, \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{bmatrix} \right)_{\boldsymbol{M}_2} = \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{y}_0 + \int_{-r}^{0} \boldsymbol{x}_1^{\mathrm{T}}(\vartheta)\boldsymbol{y}_1(\vartheta)g(\vartheta)d\vartheta, \tag{2.3}$$

where $g(\vartheta)$ is the piecewise constant nondecreasing function defined as

$$g(\vartheta) = \chi_{[-r_\delta,-r_{\delta-1}]}(\vartheta) + \sum_{j=1}^{\delta-1}(\delta-j+1)\chi_{(-r_j,-r_{j-1}]}(\vartheta), \tag{2.4}$$

where χ_S denotes the characteristic function of the interval S.

Here and in the following the standard assumption is made that summations vanish when the upper limit is smaller than the lower one (e.g., $\delta = 1$ in (2.4)).

In this paper, for the sake of brevity and whenever it does not cause confusion, the space $L_2([-r,0];\mathbb{R}^N)$ will be simply indicated as L_2. In the same way C^k will denote the space $C^k([-r,0];\mathbb{R}^N)$ of functions with values in $\mathbb{R}^N$ that have continuous derivatives until order k, while the symbol $W^{1,2}$ will indicate the space of absolutely continuous functions from $[-r,0]$ in $\mathbb{R}^N$, with derivative in L_2.

In $\boldsymbol{M}_2$ the system (2.1), (2.2) assumes the form

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{F}\boldsymbol{\omega}(t), \quad \boldsymbol{x}(0) = \begin{bmatrix} \bar{\boldsymbol{z}}(0) \\ \bar{\boldsymbol{z}} \end{bmatrix} + \begin{bmatrix} \mathcal{L}_0 \\ \mathcal{L}_1 \end{bmatrix} \bar{\boldsymbol{\omega}}, \tag{2.5}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{G}\boldsymbol{\omega}(t), \tag{2.6}$$

where $\boldsymbol{A} : \mathcal{D}(\boldsymbol{A}) \mapsto \boldsymbol{M}_2$ is defined as

$$\boldsymbol{A}\begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_0\boldsymbol{x}_0 + \sum_{h=1}^{\delta} \boldsymbol{A}_h \boldsymbol{x}_1(-r_h) + \int_{-r}^{0} \boldsymbol{A}_{01}(\vartheta)\boldsymbol{x}_1(\vartheta)d\vartheta \\ \frac{d}{d\vartheta}\boldsymbol{x}_1 \end{bmatrix} \tag{2.7}$$

with domain

$$\mathcal{D}(\boldsymbol{A}) = \left\{ \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix} \middle| \begin{matrix} \boldsymbol{x}_0 \in \mathbb{R}^N \\ \boldsymbol{x}_1 \in W^{1,2} \end{matrix} \quad \boldsymbol{x}_0 = \boldsymbol{x}_1(0) \right\}, \tag{2.8}$$

and the linear operators $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{F}$ are defined as

$$\boldsymbol{B} : \mathbb{R}^p \mapsto \boldsymbol{M}_2, \qquad \boldsymbol{B}\boldsymbol{u}(t) = \begin{bmatrix} \boldsymbol{B}_0\boldsymbol{u}(t) \\ \boldsymbol{0} \end{bmatrix}, \tag{2.9}$$

$$\boldsymbol{C} : \boldsymbol{M}_2 \mapsto \mathbb{R}^q, \qquad \boldsymbol{C}\begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix} = \boldsymbol{C}_0\boldsymbol{x}_0, \tag{2.10}$$

$$\boldsymbol{F} : \mathbb{R}^s \mapsto \boldsymbol{M}_2, \qquad \boldsymbol{F}\boldsymbol{\omega}(t) = \begin{bmatrix} \boldsymbol{F}_0 \ \boldsymbol{\omega}(t) \\ \boldsymbol{0} \end{bmatrix}. \tag{2.11}$$

The Hilbert–Schmidt operator $\mathcal{L} = \left[\begin{smallmatrix} \mathcal{L}_0 \\ \mathcal{L}_1 \end{smallmatrix}\right]$, which defines the stochastic initial state $\boldsymbol{x}(0)$, derives from definition (2.2) and is defined as follows:

$$\begin{aligned} &\mathcal{L}_0 : \ L_2([-r,0];\mathbb{R}^{\bar{s}}) \mapsto \mathbb{R}^N; \qquad && \mathcal{L}_0\boldsymbol{\omega} = \int_{-r}^{0} k(0,\tau)\bar{\boldsymbol{\omega}}(\tau)d\tau, \\ &\mathcal{L}_1 : \ L_2([-r,0];\mathbb{R}^{\bar{s}}) \mapsto W^{1,2}; \qquad && \mathcal{L}_1\boldsymbol{\omega}(\vartheta) = \int_{-r}^{0} k(\vartheta,\tau)\bar{\boldsymbol{\omega}}(\tau)d\tau. \end{aligned} \tag{2.12}$$

The mean value and nuclear covariance of the initial state $\boldsymbol{x}_0$ are as follows:

$$\bar{\boldsymbol{x}}_0 = \begin{bmatrix} \bar{\boldsymbol{z}}(0) \\ \bar{\boldsymbol{z}} \end{bmatrix}, \qquad \boldsymbol{P}_0 = \mathcal{L}\mathcal{L}^*. \tag{2.13}$$

Remark 2.1. Note that the weighted scalar product (2.3), (2.4) assures that there exist real α such that $\boldsymbol{A}-\alpha\boldsymbol{I}$ has the nice property to be dissipative [3]. This property is used in the paper to prove the convergence of the approximation scheme.

For the reader's convenience, the definitions of some operators related to the system (2.5), (2.6) that will be extensively used in the paper are reported below.

PROPOSITION 2.2. *The operators $\boldsymbol{B}^*$, $\boldsymbol{C}^*$, $\boldsymbol{F}^*$, $\boldsymbol{B}\boldsymbol{B}^*$, $\boldsymbol{C}^*\boldsymbol{C}$, $\boldsymbol{F}\boldsymbol{F}^*$, and $\boldsymbol{A}^*$ are as follows:*

$$\boldsymbol{B}^* : \boldsymbol{M}_2 \mapsto \mathbb{R}^p, \qquad \boldsymbol{B}^* \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix} = \boldsymbol{B}_0^{\mathrm{T}} \boldsymbol{x}_0; \tag{2.14}$$

$$\boldsymbol{C}^* : \mathbb{R}^q \mapsto \boldsymbol{M}_2, \qquad \boldsymbol{C}^* \boldsymbol{y} = \begin{bmatrix} \boldsymbol{C}_0^{\mathrm{T}} \boldsymbol{y} \\ 0 \end{bmatrix}; \tag{2.15}$$

$$\boldsymbol{F}^* : \boldsymbol{M}_2 \mapsto \mathbb{R}^s, \qquad \boldsymbol{F}^* \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_0^{\mathrm{T}} \boldsymbol{x}_0 \\ \boldsymbol{0} \end{bmatrix}; \tag{2.16}$$

$$\boldsymbol{B}\boldsymbol{B}^* : \boldsymbol{M}_2 \mapsto \boldsymbol{M}_2, \qquad \boldsymbol{B}\boldsymbol{B}^* \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_0 \boldsymbol{B}_0^{\mathrm{T}} \boldsymbol{x}_0 \\ 0 \end{bmatrix}; \tag{2.17}$$

$$\boldsymbol{C}^*\boldsymbol{C} : \boldsymbol{M}_2 \mapsto \boldsymbol{M}_2, \qquad \boldsymbol{C}^*\boldsymbol{C} \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_0^{\mathrm{T}} \boldsymbol{C}_0 \boldsymbol{x}_0 \\ 0 \end{bmatrix}; \tag{2.18}$$

$$\boldsymbol{F}\boldsymbol{F}^* : \boldsymbol{M}_2 \mapsto \boldsymbol{M}_2, \qquad \boldsymbol{F}\boldsymbol{F}^* \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_0 \boldsymbol{F}_0^{\mathrm{T}} \boldsymbol{x}_0 \\ 0 \end{bmatrix}; \tag{2.19}$$

$$\begin{aligned} &\boldsymbol{A}^* : \mathcal{D}(\boldsymbol{A}^*) \mapsto \boldsymbol{M}_2, \\ &\boldsymbol{A}^* \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{bmatrix} = \begin{bmatrix} \delta\, \boldsymbol{y}_1(0) + \boldsymbol{A}_0^{\mathrm{T}} \boldsymbol{y}_0 \\ \dfrac{1}{g} \boldsymbol{A}_{01}^{\mathrm{T}} \boldsymbol{y}_0 - \dfrac{d}{d\vartheta} \left(\boldsymbol{y}_1 - \displaystyle\sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_1) \chi_{[-r,-r_j]} \right) \end{bmatrix}, \end{aligned} \tag{2.20}$$

with dense domain

$$\mathcal{D}(\boldsymbol{A}^*) = \left\{ \begin{bmatrix} \boldsymbol{y}_0 \\ \\ \boldsymbol{y}_1 \end{bmatrix} \middle| \begin{array}{c} \boldsymbol{y}_0 \in \mathbb{R}^N, \quad \boldsymbol{A}_\delta^{\mathrm{T}} \boldsymbol{y}_0 = \boldsymbol{y}_1(-r), \\ \left(\boldsymbol{y}_1 - \displaystyle\sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_1) \chi_{[-r,-r_j]} \right) \in W^{1,2} \end{array} \right\}, \tag{2.21}$$

where

$$\boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_1) = \frac{\boldsymbol{y}_1(-r_j) - \boldsymbol{A}_j^{\mathrm{T}} \boldsymbol{y}_0}{\delta - j + 1}, \qquad j = 1, \ldots, \delta - 1. \tag{2.22}$$

The proof that the operator defined by (2.20), (2.21), (2.22) is in fact the adjoint of operator $\boldsymbol{A}$ is reported in appendix.

Remark 2.3. The difference between the case of just one pure delay and of multiple pure delays is given by summations in (2.20), (2.21), which vanish in the first case and complicate the analysis very much in the second one.

3. The finite horizon LQG for delay systems. In this section the problem of defining a feedback control law for the stochastic delay system (2.1), (2.2) is considered. In particular we are interested in the problem of synthesizing the control law that minimizes the cost functional

$$J_f(\boldsymbol{u}) = \int_0^{t_f} E[\boldsymbol{z}^{\mathrm{T}}(t)\boldsymbol{Q}_0\boldsymbol{z}(t) + \boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{u}(t)]dt, \tag{3.1}$$

with $0 < t_f < \infty$, where matrix $\boldsymbol{Q}_0$ is symmetric nonnegative definite. It can be readily recognized that the functional (3.1) admits the following representation in $\boldsymbol{M}_2$:

$$J_f(\boldsymbol{u}) = \int_0^{t_f} E[\big(\boldsymbol{Q}\boldsymbol{x}(t), \boldsymbol{x}(t)\big) + \boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{u}(t)]dt, \tag{3.2}$$

where $\boldsymbol{Q} : \boldsymbol{M}_2 \mapsto \boldsymbol{M}_2$ is defined as

$$\boldsymbol{Q}\begin{bmatrix}\boldsymbol{x}_0\\ \boldsymbol{x}_1\end{bmatrix} = \begin{bmatrix}\boldsymbol{Q}_0\boldsymbol{x}_0\\ \boldsymbol{0}\end{bmatrix} \tag{3.3}$$

and $\boldsymbol{x}(t)$ satisfies system equations (2.5), (2.6). The solution of this problem, as is well known, is the classical LQG controller given by the following equations [1]:

$$\boldsymbol{u}(t) = -\boldsymbol{B}^*\boldsymbol{R}(t_f - t)\hat{\boldsymbol{x}}(t), \tag{3.4}$$

$$\boldsymbol{R}(t) = \int_0^t \boldsymbol{T}^*(t-\tau)[\boldsymbol{Q} - \boldsymbol{R}(\tau)\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}(\tau)]\boldsymbol{T}(t-\tau)d\tau, \tag{3.5}$$

$$\hat{\boldsymbol{x}}(t) = \boldsymbol{T}(t)\hat{\boldsymbol{x}}_0 + \int_0^t \boldsymbol{T}(t-\tau)\left[\boldsymbol{P}(\tau)\boldsymbol{C}^*[\boldsymbol{y}(\tau) - \boldsymbol{C}\hat{\boldsymbol{x}}(\tau)] + \boldsymbol{B}\boldsymbol{u}(\tau)\right]d\tau, \tag{3.6}$$

$$\boldsymbol{P}(t) = \boldsymbol{T}(t)\boldsymbol{P}_0\boldsymbol{T}^*(t) + \int_0^t \boldsymbol{T}(t-\tau)[\boldsymbol{F}\boldsymbol{F}^* - \boldsymbol{P}(\tau)\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{P}(\tau)]\boldsymbol{T}^*(t-\tau)d\tau, \tag{3.7}$$

where $\boldsymbol{T}(t)$ is the semigroup governing the system, that is, the semigroup generated by the operator $\boldsymbol{A}$ in (2.7), (2.8), and $\hat{\boldsymbol{x}}_0$ and $\boldsymbol{P}_0$ are the expected value and the covariance operator of the initial state $\boldsymbol{x}(0)$ in $\boldsymbol{M}_2$, respectively. The solution given by these equations is a very important result only from a theoretical point of view. For our purposes we need to recall that the solutions of the Riccati equations (3.5), (3.7) evolve in the Hilbert space of Hilbert–Schmidt operators and moreover, for every $t_f < \infty$, there exist constants K_P and K_R such that [20]

$$\begin{aligned} \sup_{t\in[0,t_f]} \|\boldsymbol{P}(t)\|_{H.S.} &= K_P < \infty, \\ \sup_{t\in[0,t_f]} \|\boldsymbol{R}(t)\|_{H.S.} &= K_R < \infty, \end{aligned} \tag{3.8}$$

where, as usual, $\|\cdot\|_{H.S.}$ denotes the Hilbert–Schmidt norm [1].

In engineering applications, due to its infinite dimensional nature, such a solution is not directly implementable. Therefore it becomes important to investigate when such a solution admits a finite dimensional approximation.

Throughout the paper, given a Hilbert space $\mathcal{X}$ and a closed subspace $\mathcal{S} \subset \mathcal{X}$, the orthogonal projection operator from $\mathcal{X}$ to $\mathcal{S}$ will be denoted as $\boldsymbol{\Pi}_{\mathcal{S}}$.

In the next lemma, the linear space of bounded operators on a Hilbert space H is denoted $L(H)$.

LEMMA 3.1. *Let H_1, H_2 be separable Hilbert spaces. Let $\{G_m(t),\ t \in [0,t_f]\}$ be a sequence of strongly continuous $L(H_2)$ valued functions, strongly convergent to $\{G(t),\ t \in [0,t_f]\}$, uniformly on $[0,t_f]$. Let K be a compact subset in the Hilbert space of Hilbert–Schmidt operators mapping H_1 to H_2.*

Then $\|G_m(t)N - G(t)N\|_{H.S.}$ converges to zero, uniformly with respect to $N \in K$ and $t \in [0,t_f]$.

Proof. See [20]. □

LEMMA 3.2. *Let H_1 and H_2 be separable Hilbert spaces. Let $G(t)$ be a semigroup on H_2 and $G_n(t)$ a sequence of semigroups on H_2 strongly convergent to $G(t)$ uniformly with respect to $t \in [0,t_f]$. For $0 \le \tau \le t$, let $\Gamma(t,\tau)$ be the mild evolution operator*

$$\Gamma(t,\tau) = G(t-\tau) + \int_\tau^t G(t-\vartheta)Op(\vartheta)\Gamma(\vartheta,\tau)d\vartheta, \tag{3.9}$$

where $Op \in C\left([0,t_f];L(H_2)\right)$ and let $\Gamma_n(t,\tau)$ be the sequence of mild evolution operators

$$\Gamma_n(t,\tau) = G_n(t-\tau) + \int_\tau^t G_n(t-\vartheta)Op_n(\vartheta)\Gamma_n(\vartheta,\tau)d\vartheta, \tag{3.10}$$

where $Op_n \in C\left([0,t_f];L(H_2)\right)$ converges pointwise strongly to Op, uniformly in $[0,t_f]$. Let K be a compact subset in the Hilbert space of Hilbert–Schmidt operators mapping H_1 to H_2.

Then $\|\Gamma(t,\tau)N - \Gamma_n(t,\tau)N\|_{H.S.}$ converges to zero, uniformly with respect to $N \in K$ and $0 \le \tau \le t \le t_f$.

Proof. It is

$$\begin{aligned}\|\Gamma(t,\tau)N - \Gamma_n(t,\tau)N\|_{H.S.} &\le \|G(t-\tau)N - G_n(t-\tau)N\|_{H.S.}\\ &+ \int_\tau^t \|G(t-\vartheta)Op(\vartheta)\| \cdot \|\Gamma(\vartheta,\tau)N - \Gamma_n(\vartheta,\tau)N\|_{H.S.}d\vartheta\\ &+ \int_\tau^t \|G(t-\vartheta)Op(\vartheta) - G_n(t-\vartheta)Op_n(\vartheta)\|\\ &\qquad \cdot\|\Gamma_n(\vartheta,\tau)N - \Gamma(\vartheta,\tau)N\|_{H.S.}d\vartheta\\ &+ \int_\tau^t \|(G(t-\vartheta)Op(\vartheta) - G_n(t-\vartheta)Op_n(\vartheta))\Gamma(\vartheta,\tau)N\|_{H.S.}d\vartheta.\end{aligned} \tag{3.11}$$

Let M be a positive real such that

$$\begin{aligned} M &\ge \sup_{(t,\vartheta)\in[0,t_f]\times[0,t_f]} \|G(t)\|\|Op(\vartheta)\|,\\ M &\ge \sup_{(t,\vartheta,n)\in[0,t_f]\times[0,t_f]\times Z^+} \|G_n(t)\|\|Op_n(\vartheta)\|.\end{aligned} \tag{3.12}$$

Then

$$\begin{aligned}\|\Gamma(t,\tau)N - &\Gamma_n(t,\tau)N\|_{H.S.} \leq \|G(t-\tau)N - G_n(t-\tau)N\|_{H.S.}\\ &+\int_\tau^t \|(G(t-\vartheta)Op(\vartheta) - G_n(t-\vartheta)Op_n(\vartheta))\Gamma(\vartheta,\tau)N\|_{H.S.}d\vartheta\\ &+3M\int_\tau^t \|\big(\Gamma_n(\vartheta,\tau) - \Gamma(\vartheta,\tau)\big)N\|_{H.S.}d\vartheta.\end{aligned} \tag{3.13}$$

Applying the Gronwall's inequality,

$$\begin{aligned}\|\Gamma(t,\tau)N - &\Gamma_n(t,\tau)N\|_{H.S.} \leq e^{3Mt_f}\Big(\|G(t-\tau)N - G_n(t-\tau)N\|_{H.S.}\\ &+\int_\tau^t \big\|\big(G(t-\vartheta)Op(\vartheta) - G_n(t-\vartheta)Op_n(\vartheta)\big)\Gamma(\vartheta,\tau)N\big\|_{H.S.}d\vartheta\Big)\\ &\leq e^{3Mt_f}\Big(\|G(t-\tau)N - G_n(t-\tau)N\|_{H.S.}\\ &+\int_\tau^t \big\|\big(G(t-\vartheta) - G_n(t-\vartheta)\big)Op(\vartheta)\Gamma(\vartheta,\tau)N\big\|_{H.S.}d\vartheta\\ &+\int_\tau^t M\big\|\big(Op(\vartheta) - Op_n(\vartheta)\big)\Gamma(\vartheta,\tau)N\big\|_{H.S.}d\vartheta\Big).\end{aligned} \tag{3.14}$$

Since the set of operators $\{Op(\vartheta)\Gamma(t,\tau)N, \vartheta\in[0,t_f], 0\leq\tau\leq t\leq t_f\}$ and the set $\{\Gamma(t,\tau)N,\ 0\leq\tau\leq t\leq t_f\}$ are compact in the Hilbert space of Hilbert–Schmidt operators mapping H_1 to H_2, by Lemma 3.1 the right-hand side of inequality (3.14) tends to zero for $n\to\infty$, and the lemma is proved. □

THEOREM 3.3. *Let Ψ_n and Ψ'_n be sequences of finite dimensional subspaces of $\boldsymbol{M}_2$ contained in $\mathcal{D}(\boldsymbol{A})$ and in $\mathcal{D}(\boldsymbol{A}^*)$, respectively. Let $\boldsymbol{\Pi}_{\Psi_n}:\boldsymbol{M}_2\mapsto\Psi_n$ and $\boldsymbol{\Pi}_{\Psi'_n}:\boldsymbol{M}_2\mapsto\Psi'_n$ be the sequences of orthoprojection operators in Ψ_n and Ψ'_n, respectively. Let $\boldsymbol{T}_{\Psi_n}(t)$ be the semigroup generated by the operator $\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{A}\boldsymbol{\Pi}_{\Psi_n}:\boldsymbol{M}_2\mapsto\Psi_n$ and $\boldsymbol{T}^*_{\Psi'_n}(t)$ the semigroup generated by the operator $\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{A}^*\boldsymbol{\Pi}_{\Psi'_n}:\boldsymbol{M}_2\mapsto\Psi'_n$. Let $\boldsymbol{P}_n(t)$ and $\boldsymbol{R}_n(t)$ be the solutions of the finite dimensional differential Riccati equations*

$$\begin{aligned}\dot{\boldsymbol{P}}_n(t) &= \boldsymbol{\Pi}_{\Psi_n}\boldsymbol{A}\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}_n(t) + \boldsymbol{P}_n(t)(\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{A}\boldsymbol{\Pi}_{\Psi_n})^*\\ &\qquad -\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}_n(t) + \boldsymbol{\Pi}_{\Psi_n}\boldsymbol{F}\boldsymbol{F}^*\boldsymbol{\Pi}_{\Psi_n},\\ \boldsymbol{P}_n(0) &= \boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}_0\boldsymbol{\Pi}_{\Psi_n},\end{aligned} \tag{3.15}$$

$$\begin{aligned}\dot{\boldsymbol{R}}_n(t) &= \boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{A}^*\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{R}_n(t) + \boldsymbol{R}_n(t)(\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{A}^*\boldsymbol{\Pi}_{\Psi'_n})^*\\ &\qquad -\boldsymbol{R}_n(t)\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{R}_n(t) + \boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{Q}\boldsymbol{\Pi}_{\Psi'_n},\\ \boldsymbol{R}_n(0) &= \boldsymbol{0}.\end{aligned} \tag{3.16}$$

Assume the following hypotheses are satisfied:

(Hp$_1$) $\boldsymbol{\Pi}_{\Psi_n}$ *converges strongly to the identity operator;*

(Hp$_2$) $\boldsymbol{\Pi}_{\Psi'_n}$ *converges strongly to the identity operator;*

(Hp$_3$) $\boldsymbol{T}_{\Psi_n}(t)$ *converges strongly to* $\boldsymbol{T}(t)$ *uniformly in* $[0,t_f]$*;*

(Hp$_4$) $\boldsymbol{T}^*_{\Psi'_n}(t)$ *converges strongly to* $\boldsymbol{T}^*(t)$ *uniformly in* $[0,t_f]$.

Then

$$\begin{aligned} &\|\boldsymbol{P}_n(t) - \boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}\|_{H.S.} \to 0, \\ &\|\boldsymbol{R}_n(t) - \boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{R}(t)\boldsymbol{\Pi}_{\Psi'_n}\|_{H.S.} \to 0, \end{aligned} \qquad \textit{uniformly in} \quad [0, t_f]. \tag{3.17}$$

Proof. See the proof of Theorem 3 in [20]. □

Remark 3.4. Note that, with the given definitions, in general the semigroup $\boldsymbol{T}^*_{\Psi'_n}(t)$ generated by the operator $\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{A}^*\boldsymbol{\Pi}_{\Psi'_n}$ is different from the semigroup $\boldsymbol{T}^*_{\Psi_n}(t)$, the adjoint of the semigroup generated by $\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{A}\boldsymbol{\Pi}_{\Psi_n}$.

LEMMA 3.5. *Let Ψ_n and Ψ'_n be sequences of finite dimensional subspaces of $\boldsymbol{M}_2$ contained in $\mathcal{D}(\boldsymbol{A})$ and in $\mathcal{D}(\boldsymbol{A}^*)$, respectively. Let $\boldsymbol{\Pi}_{\Psi_n} : \boldsymbol{M}_2 \mapsto \Psi_n$ and $\boldsymbol{\Pi}_{\Psi'_n} : \boldsymbol{M}_2 \mapsto \Psi'_n$ be the corresponding sequences of orthoprojection operators. Let $\mathcal{H} = \boldsymbol{M}_2 \times \boldsymbol{M}_2$ and $\mathcal{H}_n = \boldsymbol{M}_2 \times \Psi_n$. Consider the following operators:*

$$\mathcal{A} = \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & \boldsymbol{A} \end{bmatrix} : \mathcal{D}(\boldsymbol{A}) \times \mathcal{D}(\boldsymbol{A}) \mapsto \mathcal{H}, \tag{3.18}$$

$$\mathcal{A}_n = \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & \boldsymbol{\Pi}_{\Psi_n}\boldsymbol{A}\boldsymbol{\Pi}_{\Psi_n} \end{bmatrix} : \mathcal{D}(\boldsymbol{A}) \times \boldsymbol{M}_2 \mapsto \mathcal{H}_n, \tag{3.19}$$

$$\boldsymbol{D}(t) = \begin{bmatrix} 0 & -\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}(t_f - t) \\ \boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{C} & -\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}(t_f - t) - \boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{C} \end{bmatrix} : \mathcal{H} \mapsto \mathcal{H}, \tag{3.20}$$

$$\boldsymbol{D}_n(t) = \begin{bmatrix} 0 & -\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}_n(t_f - t)\boldsymbol{\Pi}_{\Psi'_n} \\ \boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C} & -\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}_n(t_f - t)\boldsymbol{\Pi}_{\Psi'_n} - \boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n} \end{bmatrix} :$$

$$\mathcal{H} \mapsto \mathcal{H}_n, \tag{3.21}$$

$$\boldsymbol{O}(t) = \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{G} \end{bmatrix} : \mathbb{R}^s \mapsto \mathcal{H}, \tag{3.22}$$

$$\boldsymbol{O}_n(t) = \begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{G} \end{bmatrix} : \mathbb{R}^s \mapsto \mathcal{H}_n. \tag{3.23}$$

Let $\boldsymbol{S}(t)$ and $\boldsymbol{S}_n(t)$ be the semigroups generated by operators $\mathcal{A}$ and $\mathcal{A}_n$, respectively.

Let $\boldsymbol{\Pi}_{\mathcal{H}_n}$ be the following sequence of orthoprojection operators, strongly converging to identity,

$$\boldsymbol{\Pi}_{\mathcal{H}_n} = \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \boldsymbol{\Pi}_{\Psi_n} \end{bmatrix} : \mathcal{H} \mapsto \mathcal{H}_n \tag{3.24}$$

Assume that assumptions Hp_1–Hp_4 of Theorem 3.3 *are satisfied.*

Then

(Th$_1$) $\boldsymbol{S}_n(t)$*converges strongly to* $\boldsymbol{S}(t)$ *uniformly in* $[0, t_f]$;

(Th$_2$) $\|\boldsymbol{\Pi}_{\mathcal{H}_n}\boldsymbol{O}(t) - \boldsymbol{O}_n(t)\|_{H.S.} \to 0$ *uniformly in* $[0, t_f]$;

(Th$_3$) $\|\boldsymbol{\Pi}_{\mathcal{H}_n}\boldsymbol{D}(t) - \boldsymbol{D}_n(t)\|_{H.S.} \to 0$ *uniformly in* $[0, t_f]$.

Proof. Thesis Th$_1$ is an immediate consequence of hypothesis Hp$_3$.

As far as Th$_2$ is concerned, we have

$$\begin{aligned}\|\boldsymbol{\Pi}_{\mathcal{H}_n}\boldsymbol{O}(t)-\boldsymbol{O}_n(t)\|_{H.S.}&=\|\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{G}-\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{G}\|_{H.S.}\\ &\le \Big\|\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{G}-\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{G}\\ &\qquad +\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{G}-\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{G}\|_{H.S.}\\ &\le \|\boldsymbol{\Pi}_{\Psi_n}\|\|\boldsymbol{P}(t)\|_{H.S.}\|\boldsymbol{C}^*\boldsymbol{G}-\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{G}\|_{H.S.}\\ &\qquad +\|\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}-\boldsymbol{P}_n(t)\|_{H.S.}\|\boldsymbol{\Pi}_{\Psi_n}\|\|\boldsymbol{C}^*\boldsymbol{G}\|.\end{aligned}\tag{3.25}$$

From (3.25) it follows that $\|\boldsymbol{\Pi}_{\mathcal{H}_n}\boldsymbol{O}(t)-\boldsymbol{O}_n(t)\|_{H.S.}\to 0$ uniformly with respect to $t\in[0,t_f]$ because of the boundedness of $\|\boldsymbol{P}(t)\|_{H.S}$ and the uniform convergence of $\boldsymbol{P}_n(t)$ stated in Theorem 3.3.

As for thesis Th$_3$, it is

$$\boldsymbol{\Pi}_{\mathcal{H}_n}\boldsymbol{D}(t)-\boldsymbol{D}_n(t)=\begin{bmatrix}0 & \boldsymbol{O}p^n_{1,2}(t)\\ \boldsymbol{O}p^n_{2,1}(t) & \boldsymbol{O}p^n_{2,2}(t)\end{bmatrix},\tag{3.26}$$

where

$$\begin{aligned}\boldsymbol{O}p^n_{1,2}(t)&=\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}_n(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}-\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}(t_f-t);\\ \boldsymbol{O}p^n_{2,1}(t)&=\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{C}-\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C};\\ \boldsymbol{O}p^n_{2,2}(t)&=\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}_n(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}\\ &\qquad+\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n}-\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{R}(t_f-t)-\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{C}.\end{aligned}\tag{3.27}$$

To prove Th$_3$ it is sufficient to prove that the three operators in (3.27) converge uniformly to zero in the $H.S.$ norm. Let us start with operator $\boldsymbol{O}p^n_{1,2}(t)$. We have

$$\begin{aligned}\|\boldsymbol{O}p^n_{1,2}(t)\|&\le\|\boldsymbol{B}\boldsymbol{B}^*\|\big\|\boldsymbol{R}_n(t_f-t)-\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{R}(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}\big\|_{H.S.}\|\boldsymbol{\Pi}_{\Psi'_n}\|\\ &\qquad+\|\boldsymbol{B}\boldsymbol{B}^*\|\big\|\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{R}(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}-\boldsymbol{R}(t_f-t)\big\|_{H.S.}.\end{aligned}\tag{3.28}$$

Moreover, from the uniform convergence of $\boldsymbol{R}_n(t)$ stated in Theorem 3.3, $\boldsymbol{R}$ being self-adjoint, and for Lemma 3.1, by

$$\begin{aligned}&\|\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{R}(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}-\boldsymbol{R}(t_f-t)\|_{H.S.}\\ &\quad\le\|\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{R}(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}-\boldsymbol{R}(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}+\boldsymbol{R}(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}-\boldsymbol{R}(t_f-t)\|_{H.S.}\\ &\quad\le 2\|\boldsymbol{\Pi}_{\Psi'_n}\boldsymbol{R}(t_f-t)-\boldsymbol{R}(t_f-t)\|,\end{aligned}\tag{3.29}$$

it follows that $\|\boldsymbol{O}p^n_{1,2}(t)\|_{H.S.}\to 0$ uniformly in $[0,t_f]$. Consider now the term $\boldsymbol{O}p^n_{2,1}(t)$. Its Hilbert–Schmidt norm satisfies

$$\begin{aligned}&\|\boldsymbol{O}p^n_{2,1}(t)\|_{H.S.}\\ &\quad\le\ \big\|\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)-\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}+\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}-\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\big\|_{H.S.}\|\boldsymbol{C}^*\boldsymbol{C}\|\\ &\quad\le\ \Big(\|\boldsymbol{P}(t)-\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}\|_{H.S.}+\|(\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}-\boldsymbol{P}_n(t))\boldsymbol{\Pi}_{\Psi_n}\|_{H.S.}\Big)\|\boldsymbol{C}^*\boldsymbol{C}\|.\end{aligned}\tag{3.30}$$

Since, by Lemma 3.1, $\|\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)-\boldsymbol{P}(t)\|_{H.S.}\to 0$ uniformly and $\|(\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}-\boldsymbol{P}_n(t))\boldsymbol{\Pi}_{\Psi_n}\|_{H.S.}\to 0$ by Theorem 3.3, it follows that the norm of $\boldsymbol{Op}^n_{2,1}(t)$ tends to zero uniformly in $[0,t_f]$.

It remains to prove that $\|\boldsymbol{Op}^n_{2,2}(t)\|_{H.S.}\to 0$ uniformly.

$$\begin{aligned}\|\boldsymbol{Op}^n_{2,2}(t)\|_{H.S.}\le \|\boldsymbol{\Pi}_{\Psi_n}\|\big\|\boldsymbol{BB}^*\boldsymbol{R}(t_f-t)-\boldsymbol{BB}^*\boldsymbol{R}_n(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}\big\|_{H.S.}&\\ +\big\|\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{C}-\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n}\big\|_{H.S.}&\end{aligned}\tag{3.31}$$

Uniform convergence to zero of $\|\boldsymbol{BB}^*\boldsymbol{R}(t_f-t)-\boldsymbol{BB}^*\boldsymbol{R}_n(t_f-t)\boldsymbol{\Pi}_{\Psi'_n}\|_{H.S.}$ has already been proved. Moreover,

$$\begin{aligned}&\|\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{C}-\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n}\|_{H.S.}\\ &\quad\le\ \|\boldsymbol{\Pi}_{\Psi_n}\|\|\boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{C}-\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n}\|_{H.S.}\\ &\qquad+\big\|\big(\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}-\boldsymbol{P}_n(t)\big)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\big\|_{H.S.}\|\boldsymbol{\Pi}_{\Psi_n}\|.\end{aligned}\tag{3.32}$$

Again, as proved in [20], the term $\|(\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{\Pi}_{\Psi_n}-\boldsymbol{P}_n(t))\boldsymbol{\Pi}_{\Psi_n}\|_{H.S.}\to 0$ uniformly and thanks to Lemma 3.1 also $\|\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{P}(t)-\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\|_{H.S.}\to 0$ uniformly. From (3.32) it follows that $\|\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{P}(t)\boldsymbol{C}^*\boldsymbol{C}-\boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n}\|_{H.S.}\to 0$ uniformly, so that $\|\boldsymbol{Op}^n_{2,2}(t)\|_{H.S.}\to 0$ uniformly in $[0,t_f]$, and the lemma is proved. □

LEMMA 3.6. *Let $\boldsymbol{U}(t,\tau), 0\le\tau\le t$, be the mild evolution operator*

$$\boldsymbol{U}(t,\tau)=\boldsymbol{S}(t-\tau)+\int_\tau^t \boldsymbol{S}(t-\vartheta)\boldsymbol{D}(\vartheta)\boldsymbol{U}(\vartheta,\tau)d\vartheta.\tag{3.33}$$

Let $\{\boldsymbol{U}_n(t,\tau)\}$ be the sequence of mild evolution operators

$$\boldsymbol{U}_n(t,\tau)=\boldsymbol{S}_n(t-\tau)+\int_\tau^t \boldsymbol{S}_n(t-\vartheta)\boldsymbol{D}_n(\vartheta)\boldsymbol{U}_n(\vartheta,\tau)d\vartheta.\tag{3.34}$$

Then $\{\boldsymbol{U}_n(t,\tau)\}$ converges strongly to $\boldsymbol{U}(t,\tau)$ uniformly in $0\le\tau\le t\le t_f$, that is, given any $X\in\mathcal{H}$,

$$\lim_{n\to\infty}\sup_{0\le\tau\le t\le t_f}\|\boldsymbol{U}_n(t,\tau)X-\boldsymbol{U}(t,\tau)X\|=0.\tag{3.35}$$

Proof. Let us denote by $g(t,\tau)$ and $g_n(t,\tau)$ the quantities

$$g(t,\tau)=\boldsymbol{U}(t,\tau)X,\tag{3.36}$$

$$g_n(t,\tau)=\boldsymbol{U}_n(t,\tau)X,\tag{3.37}$$

from which, denoting the approximation error by $e_n(t,\tau)$

$$e_n(t,\tau)=g(t,\tau)-g_n(t,\tau),\tag{3.38}$$

we have

$$\begin{aligned}e_n(t,\tau)=\boldsymbol{S}(t-\tau)X+\int_\tau^t \boldsymbol{S}(t-\vartheta)\boldsymbol{D}(\vartheta)g(\vartheta,\tau)d\vartheta&\\ -\boldsymbol{S}_n(t-\tau)X-\int_\tau^t \boldsymbol{S}_n(t-\vartheta)\boldsymbol{D}_n(\vartheta)g_n(\vartheta,\tau)d\vartheta,&\end{aligned}\tag{3.39}$$

and therefore

$$\begin{aligned} e_n(t,\tau) &= (\boldsymbol{S}(t-\tau) - \boldsymbol{S}_n(t-\tau))X \\ &\quad + \int_\tau^t \Big(\boldsymbol{S}(t-\vartheta)\boldsymbol{D}(\vartheta)g(\vartheta,\tau) - \boldsymbol{S}_n(t-\vartheta)\boldsymbol{D}_n(\vartheta)g_n(\vartheta,\tau)\Big)d\vartheta \end{aligned} \tag{3.40}$$

from which

$$\begin{aligned} \|e_n(t,\tau)\| &\le \|(\boldsymbol{S}(t-\tau) - \boldsymbol{S}_n(t-\tau))X\| \\ &\quad + \int_\tau^t \|\boldsymbol{S}(t-\vartheta)\boldsymbol{D}(\vartheta) - \boldsymbol{S}_n(t-\vartheta)\boldsymbol{D}(\vartheta)\|\|g(\vartheta,\tau)\|d\vartheta \\ &\quad + \int_\tau^t \|\boldsymbol{S}_n(t-\vartheta)\|\|\boldsymbol{\Pi}_{\mathcal{H}_n}\boldsymbol{D}(\vartheta) - \boldsymbol{D}_n(\vartheta)\|\|g(\vartheta,\tau)\|d\vartheta \\ &\quad + \int_\tau^t \|\boldsymbol{S}_n(t-\vartheta)\boldsymbol{D}_n(\vartheta)\|\|e_n(\vartheta,\tau)\|d\vartheta. \end{aligned} \tag{3.41}$$

Now, given $\epsilon > 0$, by Lemma 3.1 there exists an integer $\nu_{\epsilon,X}$ such that, for all $n > \nu_{\epsilon,X}$, we have

$$\|e_n(t,\tau)\| \le \epsilon + \bar{S}\bar{D}\int_\tau^t \|e_n(\vartheta,\tau)\|d\vartheta, \tag{3.42}$$

where

$$\begin{aligned} \bar{S} &= \sup_{n,t\in[0,t_f]} \|\boldsymbol{S}_n(t)\|, \\ \bar{D} &= \sup_{n,t\in[0,t_f]} \|\boldsymbol{D}_n(t)\|. \end{aligned} \tag{3.43}$$

By Gronwall's lemma,

$$\|e_n(t,\tau)\| \le \epsilon e^{\bar{S}\bar{D}(t-\tau)}, \tag{3.44}$$

and therefore

$$\sup_{0\le\tau\le t\le t_f} \|e_n(t,\tau)\| \le \epsilon e^{\bar{S}\bar{D}t_f}. \tag{3.45}$$

This concluded the proof. □

Now, the main theorem can be given.

THEOREM 3.7. *Using the same hypotheses of Theorem* 3.3, *let* $\boldsymbol{u}_n(t)$ *be the input obtained by the following finite dimensional equations:*

$$\begin{aligned} &\dot{\hat{\boldsymbol{x}}}_n(t) = \boldsymbol{\Pi}_{\Psi_n}\boldsymbol{A}\boldsymbol{\Pi}_{\Psi_n}\hat{\boldsymbol{x}}_n(t) + \boldsymbol{\Pi}_{\Psi_n}\boldsymbol{B}\boldsymbol{u}_n(t) + \boldsymbol{P}_n(t)\boldsymbol{\Pi}_{\Psi_n}\boldsymbol{C}^*\big(\boldsymbol{y}(t) - \boldsymbol{C}\boldsymbol{\Pi}_{\Psi_n}\hat{\boldsymbol{x}}_n(t)\big), \\ &\hat{\boldsymbol{x}}_n(0) = \boldsymbol{\Pi}_{\Psi_n}\hat{\boldsymbol{x}}(0), \end{aligned} \tag{3.46}$$

$$\boldsymbol{u}_n(t) = -\boldsymbol{B}^*\boldsymbol{R}_n(t_f - t)\boldsymbol{\Pi}_{\Psi'_n}\hat{\boldsymbol{x}}_n(t), \tag{3.47}$$

where $\boldsymbol{P}_n$ *and* $\boldsymbol{R}_n$ *are given by* (3.15) *and* (3.16). *Let* $\boldsymbol{u}(t)$ *be the optimal input,* $\hat{\boldsymbol{x}}(t)$ *the optimal estimated state,* $\boldsymbol{x}_n(t)$ *and* $\boldsymbol{x}(t)$ *the actual state evolving when* $\boldsymbol{u}_n(t)$ *and* $\boldsymbol{u}(t)$ *are applied to system* (2.1), (2.2), *respectively.*

Then

$$\lim_{n\to\infty} E\|\boldsymbol{x}_n - \boldsymbol{x}\|^2_{L_2([0,t_f];\boldsymbol{M}_2)} = 0, \tag{3.48}$$

$$\lim_{n\to\infty} E\|\hat{\boldsymbol{x}}_n - \hat{\boldsymbol{x}}\|^2_{L_2([0,t_f];\boldsymbol{M}_2)} = 0, \tag{3.49}$$

$$\lim_{n\to\infty} E\|\boldsymbol{u}_n - \boldsymbol{u}\|^2_{L_2([0,t_f];\mathbb{R}^p)} = 0, \tag{3.50}$$

$$\lim_{n\to\infty} |J_f(\boldsymbol{u}_n) - J_f(\boldsymbol{u})| = 0. \tag{3.51}$$

Proof. Let $\boldsymbol{X} = \left[\begin{smallmatrix}\boldsymbol{x}\\ \hat{\boldsymbol{x}}\end{smallmatrix}\right]$ and $\boldsymbol{X}_n = \left[\begin{smallmatrix}\boldsymbol{x}_n\\ \hat{\boldsymbol{x}}_n\end{smallmatrix}\right]$. It is

$$\begin{aligned}\dot{\boldsymbol{X}}(t) &= \mathcal{A}\boldsymbol{X}(t) + \boldsymbol{D}(t)\boldsymbol{X}(t) + \boldsymbol{O}(t)\boldsymbol{\omega}(t),\\ \dot{\boldsymbol{X}}_n(t) &= \mathcal{A}_n\boldsymbol{X}_n(t) + \boldsymbol{D}_n(t)\boldsymbol{X}_n(t) + \boldsymbol{O}_n(t)\boldsymbol{\omega}(t),\\ \boldsymbol{X}(0) &= \begin{bmatrix}\boldsymbol{x}(0)\\ \hat{\boldsymbol{x}}(0)\end{bmatrix}, \qquad \boldsymbol{X}_n(0) = \begin{bmatrix}\boldsymbol{x}(0)\\ \boldsymbol{\Pi}_{\Psi_n}\hat{\boldsymbol{x}}(0)\end{bmatrix},\end{aligned} \tag{3.52}$$

where $\mathcal{A}, \mathcal{A}_n, \boldsymbol{D}(t), \boldsymbol{D}_n(t), \boldsymbol{O}(t), \boldsymbol{O}_n(t)$ have been defined in Lemma 3.5.

Let $\boldsymbol{S}(t)$, $\boldsymbol{S}_n(t)$, $\boldsymbol{U}(t,\tau)$, $\boldsymbol{U}_n(t,\tau)$ be as in Lemmas 3.5, 3.6. We have

$$\boldsymbol{X}(t) = \boldsymbol{S}(t)\boldsymbol{X}(0) + \int_0^t \boldsymbol{S}(t-\tau)\big(\boldsymbol{D}(\tau)\boldsymbol{X}(\tau) + \boldsymbol{O}(\tau)\boldsymbol{\omega}(\tau)\big)d\tau, \tag{3.53}$$

$$\boldsymbol{X}_n(t) = \boldsymbol{S}_n(t)\boldsymbol{X}_n(0) + \int_0^t \boldsymbol{S}_n(t-\tau)\big(\boldsymbol{D}_n(\tau)\boldsymbol{X}_n(\tau) + \boldsymbol{O}_n(\tau)\boldsymbol{\omega}(\tau)\big)d\tau, \tag{3.54}$$

which can be rewritten as

$$\boldsymbol{X}(t) = \boldsymbol{U}(t,0)\boldsymbol{X}(0) + \int_0^t \boldsymbol{U}(t,\tau)\boldsymbol{O}(\tau)\boldsymbol{\omega}(\tau)d\tau, \tag{3.55}$$

$$\boldsymbol{X}_n(t) = \boldsymbol{U}_n(t,0)\boldsymbol{X}_n(0) + \int_0^t \boldsymbol{U}_n(t,\tau)\boldsymbol{O}_n(\tau)\boldsymbol{\omega}(\tau)d\tau. \tag{3.56}$$

Let us introduce the Hilbert spaces

$$\boldsymbol{W}_{\boldsymbol{X},t} = L_2([0,t];\mathcal{H}), \qquad \boldsymbol{W}_{\omega,t} = L_2([0,t];\mathbb{R}^s), \tag{3.57}$$

and define the operators

$$\begin{aligned}&\boldsymbol{L}_t : \boldsymbol{W}_{\omega,t} \mapsto \boldsymbol{W}_{\boldsymbol{X},t},\\ &f = \boldsymbol{L}_t\, g, \qquad f(\tau) = \int_0^\tau \boldsymbol{U}(\tau,\vartheta)\boldsymbol{O}(\vartheta)g(\vartheta)d\vartheta,\end{aligned} \tag{3.58}$$

$$\begin{aligned}&\boldsymbol{L}_{t,n} : \boldsymbol{W}_{\omega,t} \mapsto \boldsymbol{W}_{\boldsymbol{X},t},\\ &f = \boldsymbol{L}_{t,n}\, g, \qquad f(\tau) = \int_0^\tau \boldsymbol{U}_n(\tau,\vartheta)\boldsymbol{O}_n(\vartheta)g(\vartheta)d\vartheta,\end{aligned} \tag{3.59}$$

and the functions

$$\boldsymbol{X}_0 : \quad \boldsymbol{X}_0(\tau) = \boldsymbol{U}(\tau,0)\boldsymbol{X}(0), \tag{3.60}$$

$$\boldsymbol{X}_{0,n} : \quad \boldsymbol{X}_{0,n}(\tau) = \boldsymbol{U}_n(\tau,0)\boldsymbol{X}_n(0). \tag{3.61}$$

In the space $\boldsymbol{W}_{\boldsymbol{X},t}$, (3.55), (3.56) can be expressed as

$$\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{L}_t\boldsymbol{\omega}, \tag{3.62}$$

$$\boldsymbol{X}_n = \boldsymbol{X}_{0,n} + \boldsymbol{L}_{t,n}\boldsymbol{\omega}, \tag{3.63}$$

so that

$$\begin{aligned} E\|\boldsymbol{X} - \boldsymbol{X}_n\|^2_{\boldsymbol{W}_{\boldsymbol{X},t}} &= E\|\boldsymbol{X}_0 - \boldsymbol{X}_{0,n} + (\boldsymbol{L}_t - \boldsymbol{L}_{t,n})\boldsymbol{\omega}\|^2_{\boldsymbol{W}_{\boldsymbol{X},t}} \\ &\le 2E\|\boldsymbol{X}_0 - \boldsymbol{X}_{0,n}\|^2_{\boldsymbol{W}_{\boldsymbol{X},t}} + 2E\|(\boldsymbol{L}_t - \boldsymbol{L}_{t,n})\boldsymbol{\omega}\|^2_{\boldsymbol{W}_{\boldsymbol{X},t}}. \end{aligned} \tag{3.64}$$

The first term in the right-hand side goes to zero thanks to Lemma 3.6.

For the second term we have

$$\begin{aligned} E\|(\boldsymbol{L}_t - \boldsymbol{L}_{t,n})\boldsymbol{\omega}\|^2_{\boldsymbol{W}_{(\boldsymbol{X},t)}} &= \|\boldsymbol{L}_t - \boldsymbol{L}_{t,n}\|^2_{H.S.} \\ &= \int_0^t \int_0^\tau \left\|\boldsymbol{U}(\tau,\vartheta)\boldsymbol{O}(\vartheta) - \boldsymbol{U}_n(\tau,\vartheta)\boldsymbol{O}_n(\vartheta)\right\|^2_{H.S.} d\vartheta d\tau \\ &\le 2\int_0^t \int_0^\tau \left\|\big(\boldsymbol{U}(\tau,\vartheta) - \boldsymbol{U}_n(\tau,\vartheta)\big)\boldsymbol{O}(\vartheta)\right\|^2_{H.S.} d\vartheta d\tau \\ &\quad + 2\int_0^t \int_0^\tau \left\|\boldsymbol{U}_n(\tau,\vartheta)\big(\boldsymbol{O}(\vartheta) - \boldsymbol{O}_n(\vartheta)\big)\right\|^2_{H.S.} d\vartheta d\tau \\ &\le \sup_{0\le\vartheta\le\tau\le t} \ \sup_{\boldsymbol{M}\in\{\boldsymbol{O}(\vartheta),\vartheta\in[0,t]\}} \left\|\big(\boldsymbol{U}(\tau,\vartheta) - \boldsymbol{U}_n(\tau,\vartheta)\big)M\right\|^2_{H.S.} t^2 \\ &\quad + \sup_{0\le\vartheta\le\tau\le t, \ n\in Z^+} \|\boldsymbol{U}_n(\tau,\vartheta)\| t^2 \sup_{\vartheta\in[0,t]} \|\boldsymbol{O}(\vartheta) - \boldsymbol{O}_n(\vartheta)\|^2_{H.S.} \end{aligned} \tag{3.65}$$

which goes to zero by using Lemmas 3.1, 3.5, and 3.6. This concludes the proof. □

4. The approximation scheme. In this section, we will derive the approximation scheme for the *LQG* controller (3.4)–(3.7). The first step is the definition of the sequences Ψ_n and Ψ'_n of subspaces approximating $\mathcal{D}(\boldsymbol{A})$ and $\mathcal{D}(\boldsymbol{A}^*)$. This is made by a suitable definition of basis vectors for subspaces $\Psi_n \subset \mathcal{D}(\boldsymbol{A})$ and $\Psi'_n \subset \mathcal{D}(\boldsymbol{A}^*)$. In order to avoid confusion with the general settings in section 3, the forthcoming choice for Ψ_n and Ψ'_n will be denoted by Φ_n and Φ'_n respectively. In [3] the dynamics of linear delay systems is approximated using classical first order splines uniformly distributed over the interval $[-r, 0]$. With this choice the computation of matrix representation of the approximated operators is quite complex due to the fact that in general, for a given number n of subintervals of $[-r, 0]$, the delay instants $-r_j$ do not coincide with knots of splines.

It is useful to define a multi-index $s = (n_1, \ldots, n_\delta)$ that characterizes the partition of each interval $[-r_i, -r_{i-1}]$, for $i = 1, \ldots, \delta$, into n_i subintervals of length $(r_i - r_{i-1})/n_i$, in which $n_i + 1$ classical first order splines are considered (see Figure 1), numbered from 0 to n_i.

DEFINITION 4.1. *A sequence $\{s_n\}$ of multi-indexes, defined for $n = 1, 2, \ldots$, where n is the lowest of indexes n_j of the multi-index (i.e., $n = \min\{s_n\}$), is denoted a test sequence if there exists a constant $\bar{c}$ such that for each n it is* $\max\{s_n\}/n \le \bar{c}$.

Let $t^i_j = -r_{i-1} - (r_i - r_{i-1})j/n_i$, for $j = 0, 1, \ldots, n_i$, $i = 1, \ldots, \delta$. Let *spline*^i_j be the spline j of interval i, that is, the spline with knot in t^i_j. Let ϕ_k, $k = 1, 2, \ldots, N$, be the canonical base in $\mathbb{R}^N$.

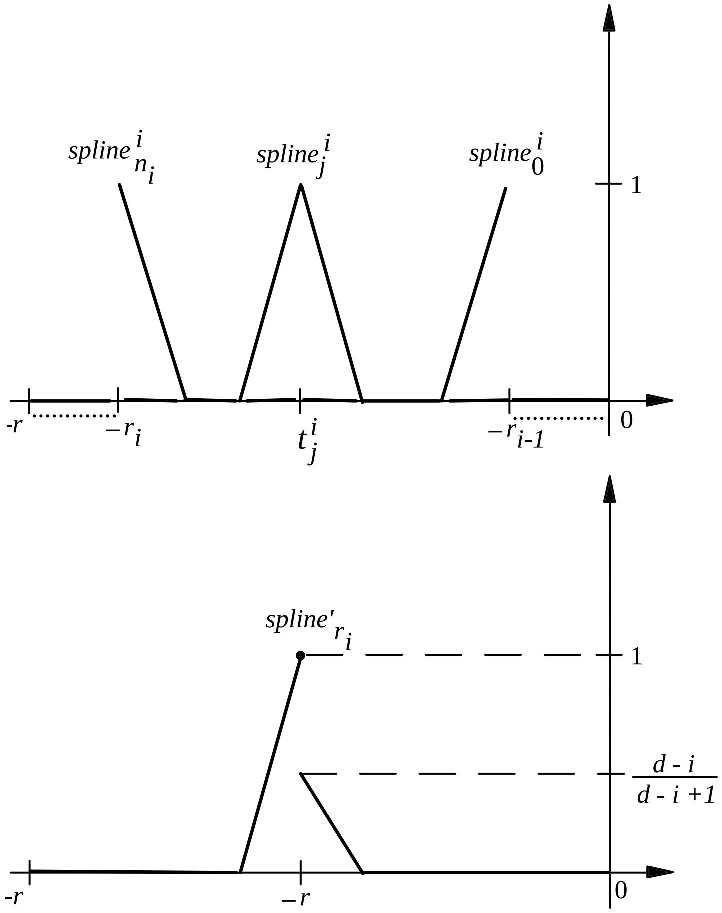


FIG. 1. *First order splines used for the approximation schemes.*

The approximating subspace Φ_n and Φ'_n are defined as follows.

DEFINITION 4.2. *For any given multi-index s_n of a test sequence let Φ_n be the subspace of linear combinations of vectors v^i_h, $v^{r_i}_k$ defined as follows:*

$$v^1_k = \begin{bmatrix} \phi_k \\ \phi_k spline^1_0 \end{bmatrix}, \quad k = 1, \dots, N, \tag{4.1}$$

$$v^i_{jN+k} = \begin{bmatrix} 0 \\ \phi_k spline^i_j \end{bmatrix}, \quad k = 1, \dots, N, \quad j = 1, \dots, n_i - 1, \tag{4.2}$$

$$v^\delta_{n_\delta N+k} = \begin{bmatrix} 0 \\ \phi_k spline^\delta_{n_\delta} \end{bmatrix}, \quad k = 1, \dots, N, \tag{4.3}$$

$$v^{r_i}_k = \begin{bmatrix} 0 \\ \phi_k spline_{r_i} \end{bmatrix}, \quad \begin{array}{l} i = 1, \dots, \delta - 1, \\ k = 1, \dots, N, \end{array} \tag{4.4}$$

where

$$spline_{r_i} = spline^i_{n_i} \cdot \chi_{(-r_i,-r_{i-1}]} + spline^{i+1}_0, \quad i = 1, \ldots, \delta - 1. \tag{4.5}$$

Definition 4.3. *For any given multi-index s_n of a test sequence let Φ'_n be the subspace of linear combinations of vectors w^i_h, $w^{r_i}_k$, w'_k defined as follows:*

$$w^1_k = \begin{bmatrix} 0 \\ \phi_k spline^1_0 \end{bmatrix}, \qquad k = 1, \ldots, N, \tag{4.6}$$

$$w^i_{jN+k} = v^i_{jN+k}, \qquad k = 1, \ldots, N, \qquad \begin{array}{l} i = 1, \ldots, \delta, \\ j = 1, \ldots, n_i - 1, \end{array} \tag{4.7}$$

$$w^{r_j}_k = \begin{bmatrix} 0 \\ \phi_k spline'_{r_j} \end{bmatrix}, \quad j = 1, \ldots, \delta - 1, \quad k = 1, \ldots, N, \tag{4.8}$$

where

$$spline'_{r_j} = \frac{(\delta - j)}{(\delta - j + 1)} spline^j_{n_j} \cdot \chi_{(-r_j,-r_{j-1}]} + spline^{j+1}_0, \tag{4.9}$$

$$spline''_{r_j} = spline^j_{n_j} \cdot \chi_{(-r_j,-r_{j-1}]}, \qquad j = 1, \ldots, \delta - 1, \tag{4.10}$$

$$w'_k = \begin{bmatrix} \phi_k \\ a_{\delta k} spline^\delta_{n_\delta} + \sum_{j=1}^{\delta-1} \frac{1}{(\delta-j+1)} a_{jk} spline''_{r_j} \end{bmatrix}, \quad k = 1, \ldots, N, \tag{4.11}$$

where a_{jk} is the k column of matrix $\boldsymbol{A}^{\mathrm{T}}_j$ for $j = 1, \ldots, \delta$.

Theorem 4.4. *For each multi-index s_n of a test sequence, it is $\Phi_n \subset \mathcal{D}(\boldsymbol{A})$ and $\Phi'_n \subset \mathcal{D}(\boldsymbol{A}^*)$.*

Proof. It is immediate to verify that $\Phi_n \subset \mathcal{D}(\boldsymbol{A})$. A more detailed proof is required to show that $\Phi'_n \subset \mathcal{D}(\boldsymbol{A}^*)$. To this aim it is sufficient to verify that each vector w belongs to it. It is easy to check that vectors w^i_{jN+k}, for $i = 1, \ldots, \delta$, $k = 1, \ldots, N$, $j = 1, \ldots, n_i - 1$, and vectors w^1_k, for $k = 1, \ldots, N$, belong to $\boldsymbol{D}(\boldsymbol{A}^*)$. Let us consider now the vectors $w^{r_i}_k$, for $i = 1, \ldots, \delta - 1$, $k = 1, \ldots, N$ (as usual, we shall indicate the part in $\mathbb{R}^N$ by using the subscript 0 and the part in L_2 by using the subscript 1).

From the definition, for each k it is

$$\begin{aligned} &(w^{r_i}_k)_0 = 0, \\ &(w^{r_i}_k)_1(-r_j) = 0, \quad i, j = 1, \ldots, \delta - 1, \ i \neq j, \\ &(w^{r_i}_k)_1(-r) = 0, \quad i = 1, \ldots, \delta - 1, \end{aligned} \tag{4.12}$$

so that for $k = 1, \ldots, N$, $\ i = 1, \ldots, \delta - 1$,

$$(w^{r_i}_k)_1(-r) = \boldsymbol{A}^{\mathrm{T}}_\delta (w^{r_i}_k)_0 \tag{4.13}$$

and

$$\begin{aligned} \sum_{j=1}^{\delta-1} \boldsymbol{k}_j\left((w^{r_i}_k)_0, (w^{r_i}_k)_1\right) \chi_{[-r,-r_j]} &= \sum_{j=1}^{\delta-1} \frac{(w^{r_i}_k)_1(-r_j) - \boldsymbol{A}^{\mathrm{T}}_j (w^{r_i}_k)_0}{\delta - j + 1} \chi_{[-r,-r_j]} \\ &= \frac{(w^{r_i}_k)_1(-r_i)}{\delta - i + 1} \chi_{[-r,-r_i]}. \end{aligned} \tag{4.14}$$

So it is only to be verified that for $i = 1, \ldots, \delta - 1,\ k = 1, \ldots, N$,

$$(w_k^{r_i})_1 - \frac{1}{(\delta - i + 1)} (w_k^{r_i})_1 (-r_i) \chi_{[-r,-r_i]} \in W^{1,2}. \tag{4.15}$$

Since

$$spline'_{r_i}(-r_i) - \frac{1}{\delta - i + 1} = \frac{\delta - i}{\delta - i + 1} = \lim_{\vartheta \to -r_i^+} spline'_{r_i}(\vartheta), \tag{4.16}$$

(4.15) is clearly true.

For vectors w'_k defined in (4.11), for $k = 1, \ldots, N$, it is

$$(w'_k)_1(-r) = a_{\delta k} = \boldsymbol{A}_\delta^{\mathrm{T}} \phi_k = \boldsymbol{A}_\delta^{\mathrm{T}} (w'_k)_0. \tag{4.17}$$

It is also

$$(w'_k)_1(-r_i) = 0, \quad i = 1, \ldots, \delta - 1, \tag{4.18}$$

and therefore

$$\begin{aligned} \sum_{j=1}^{\delta-1} \boldsymbol{k}_j \left((w'_k)_0, (w'_k)_1\right) \chi_{[-r,-r_j]} &= \sum_{j=1}^{\delta-1} \frac{(w'_k)_1 (-r_j) - \boldsymbol{A}_j^{\mathrm{T}} (w'_k)_0}{\delta - j + 1} \chi_{[-r,-r_j]} \\ &= \sum_{j=1}^{\delta-1} \frac{-\boldsymbol{A}_j^{\mathrm{T}} (w'_k)_0}{(\delta - j + 1)} \chi_{[-r,-r_j]} = \sum_{j=1}^{\delta-1} \frac{-a_{jk}}{(\delta - j + 1)} \chi_{[-r,-r_j]}. \end{aligned} \tag{4.19}$$

From

$$\lim_{\vartheta \mapsto -r_j^+} (w'_k)_1 (\vartheta) = \frac{a_{jk}}{\delta - j + 1} \tag{4.20}$$

it follows, for $i = 1, \ldots, \delta - 1$,

$$\begin{aligned} \left((w'_k)_1 + \sum_{j=1}^{\delta-1} \frac{a_{jk}}{(\delta - j + 1)} \chi_{[-r,-r_j]} \right) (-r_i) &= (w'_k)_1(-r_i) + \sum_{j=1}^{i} \frac{a_{jk}}{(\delta - j + 1)} \\ = \sum_{j=1}^{i} \frac{a_{jk}}{\delta - j + 1} &= \frac{a_{ik}}{\delta - i + 1} + \sum_{j=1}^{i-1} \frac{a_{jk}}{\delta - j + 1} \\ = \lim_{\vartheta \mapsto -r_i^+} (w'_k)_1 (\vartheta) &+ \lim_{\vartheta \mapsto -r_i^+} \sum_{j=1}^{\delta-1} \frac{a_{jk}}{\delta - j + 1} \chi_{[-r,-r_j]}(\vartheta), \end{aligned} \tag{4.21}$$

and so

$$\left((w'_k)_1 + \sum_{j=1}^{\delta-1} \frac{a_{jk}}{\delta - j + 1} \chi_{[-r,-r_j]} \right) \in \boldsymbol{W}^{1,2}, \tag{4.22}$$

(4.17) and (4.22) prove that vectors w'_k, $k = 1, \ldots, N$, belong to $\boldsymbol{D}(\boldsymbol{A}^*)$. □

Remark 4.5. Note that a key idea for the previous theorem is the choice of a type of not uniformly distributed splines.

Remark 4.6. In the case of just one pure delay, vectors generating subspaces Φ_n and Φ'_n become, respectively,

$$v_{jN+k} = \begin{bmatrix} 0 \\ \boldsymbol{\phi}_k spline_j \end{bmatrix}, \quad k = 1, \dots, N, \quad j = 1, 2, \dots, n, \tag{4.23}$$

$$v_k = \begin{bmatrix} \boldsymbol{\phi}_k \\ \boldsymbol{\phi}_k spline_0 \end{bmatrix}, \quad k = 1, \dots, N \tag{4.24}$$

for Φ_n, and

$$w_k^1 = \begin{bmatrix} 0 \\ \boldsymbol{\phi}_k spline_0 \end{bmatrix}, \quad k = 1, \dots, N, \tag{4.25}$$

$$w_{jN+k} = v_{jN+k} \qquad k = 1, \dots, N, \qquad j = 1, 2, \dots, n-1, \tag{4.26}$$

$$w'_k = \begin{bmatrix} \boldsymbol{\phi}_k \\ a_{1k} spline_n \end{bmatrix}, \quad k = 1, \dots, N \tag{4.27}$$

for Φ'_n. As can be seen, a great simplification is obtained with respect to the general case. Vectors v are just the ones in [3], and vectors w differ just for the fact that the nonzero term in $\mathbb{R}^N$ is taken from the first N vectors to the last ones, and the L_2 part of these last N vectors is multiplied for the columns of matrix $\boldsymbol{A}_1^{\mathrm{T}}$. This simplification with respect to the general case is due to the much simpler domain (2.21).

Consider now a test sequence of multi-indexes $\{s_n\}$, and consider the associated sequence of orthoprojection operators $\boldsymbol{\Pi}_{\Phi_n} : \boldsymbol{M}_2 \mapsto \Phi_n$ and $\boldsymbol{\Pi}_{\Phi'_n} : \boldsymbol{M}_2 \mapsto \Phi'_n$. For brevity, from now on the following notation is used:

$$\boldsymbol{\Pi}_n = \boldsymbol{\Pi}_{\Phi_n}, \qquad \boldsymbol{\Pi}'_n = \boldsymbol{\Pi}_{\Phi'_n}. \tag{4.28}$$

Recall that operators $\boldsymbol{\Pi}_n$ and $\boldsymbol{\Pi}'_n$, being orthogonal projectors, have the following properties:

$$\forall \boldsymbol{y} \in \boldsymbol{M}_2, \quad \|\boldsymbol{\Pi}_n \boldsymbol{y} - \boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x} \in \Phi_n, \tag{4.29}$$

$$\forall \boldsymbol{y} \in \boldsymbol{M}_2, \quad \|\boldsymbol{\Pi}'_n \boldsymbol{y} - \boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall \boldsymbol{x} \in \Phi'_n. \tag{4.30}$$

The following results can be given on the convergence of the sequences of projectors $\boldsymbol{\Pi}_n$, $\boldsymbol{\Pi}'_n$, and of the sequence of semigroups generated by $\boldsymbol{\Pi}_n \boldsymbol{A} \boldsymbol{\Pi}_n$ and $\boldsymbol{\Pi}'_n \boldsymbol{A}^* \boldsymbol{\Pi}'_n$.

THEOREM 4.7. *The sequence of orthoprojection operators* $\boldsymbol{\Pi}_n : \boldsymbol{M}_2 \mapsto \Phi_n$ *converges strongly to the identity operator.*

Proof. Let $D = \{\left[\begin{smallmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{smallmatrix}\right] \in \boldsymbol{M}_2 | \boldsymbol{y}_0 = \boldsymbol{y}_1(0), \boldsymbol{y}_1 \in C^2([-r, 0]; \mathbb{R}^N)\}$. Such set D is dense in $\boldsymbol{M}_2$ (see the proof of Lemma 2.2 and Remark 3.2 in [3]). Let $\boldsymbol{x} = \left[\begin{smallmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{smallmatrix}\right] \in D$ and let

$$\begin{aligned} \boldsymbol{x}^n = \begin{bmatrix} \boldsymbol{x}_0^n \\ \boldsymbol{x}_1^n \end{bmatrix} &= \sum_{k=1}^{N} (\boldsymbol{x}_1(-r)^{\mathrm{T}} \phi_k) v_{n_\delta N+k}^{\delta} + \sum_{k=1}^{N} \sum_{i=1}^{\delta-1} (\boldsymbol{x}_1(-r_i)^{\mathrm{T}} \phi_k) v_k^{r_i} \\ &+ \sum_{i=1}^{\delta} \sum_{k=1}^{N} \sum_{j=1}^{n_i-1} (\boldsymbol{x}_1(t_j^i)^{\mathrm{T}} \phi_k) v_{jN+k}^i + \sum_{k=1}^{N} (\boldsymbol{x}_1(0)^{\mathrm{T}} \phi_k) v_k^1. \end{aligned} \tag{4.31}$$

By Theorem 2.5 in [42] it is $\|\boldsymbol{x}_1^n - \boldsymbol{x}_1\| \to 0$, and the thesis follows by

$$\|\boldsymbol{x}^n - \boldsymbol{x}\|_{\boldsymbol{M}_2} = \|\boldsymbol{x}_1^n - \boldsymbol{x}_1\|_{\boldsymbol{L}_2} \tag{4.32}$$

and by property (4.29). □

THEOREM 4.8. *The sequence of semigroups $\boldsymbol{T}_{\Phi_n}$ generated by the operators $\boldsymbol{\Pi}_n \boldsymbol{A} \boldsymbol{\Pi}_n$ converges strongly to the semigroup governing the system* (2.5), (2.6).

Proof. Let D be the set in the proof of the previous theorem. There exists $\lambda > 0$ such that $(\boldsymbol{A} - \lambda \boldsymbol{I})D$ is dense in $\boldsymbol{M}_2$ (see Lemma 2.2 in [3]). There exists α such that $(\boldsymbol{A} - \alpha \boldsymbol{I})$ e $(\boldsymbol{\Pi}_n \boldsymbol{A} \boldsymbol{\Pi}_n - \alpha \boldsymbol{I})$ are dissipative (see Lemma 2.3 and proof of Theorem 3.1 in [3]). Let $\boldsymbol{x} = \left[\begin{smallmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \end{smallmatrix}\right] \in D$. Let $\boldsymbol{\Pi}_n \boldsymbol{x} = \left[\begin{smallmatrix} (\boldsymbol{\Pi}_n \boldsymbol{x})_0 \\ (\boldsymbol{\Pi}_n \boldsymbol{x})_1 \end{smallmatrix}\right]$. From

$$(\boldsymbol{\Pi}_n \boldsymbol{x})_1(-r_i) = (\boldsymbol{\Pi}_n \boldsymbol{x})_1(-r_{i-1}) - \int_{-r_i}^{-r_{i-1}} \frac{d(\boldsymbol{\Pi}_n \boldsymbol{x})_1(\vartheta)}{d\vartheta} d\vartheta \tag{4.33}$$

and as $\|\frac{d(\boldsymbol{x}_1 - (\boldsymbol{\Pi}_n \boldsymbol{x})_1)}{d\vartheta}\| \to 0$, (see Theorem 4.1 in [3], and Theorems 1.5, 2.5 in [42]), it follows that $\|\boldsymbol{A}\boldsymbol{\Pi}_n \boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}\| \to 0$. Take into account that $(\boldsymbol{\Pi}_n \boldsymbol{x})_1(0) = (\boldsymbol{\Pi}_n \boldsymbol{x})_0$ and that $\|(\boldsymbol{\Pi}_n \boldsymbol{x})_0 - \boldsymbol{x}_0\| \to 0$.

Thus the Trotter–Kato theorem hypotheses are satisfied ([38], Lemma 3.1 in [3]). □

As can be seen, the proofs of the above two theorems follow the same lines of the proofs in [3], developed for the case of first order splines uniformly distributed in the interval $[-r, 0]$.

LEMMA 4.9. *The subspace*

$$U = \left\{ \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{bmatrix} \middle| \begin{array}{ll} \boldsymbol{y}_0 \in \mathbb{R}^N, & \boldsymbol{y}_1(0) = \boldsymbol{y}_0, \\ \boldsymbol{y}_1 \in W^{1,2} & \boldsymbol{y}_1(-r_j) = \boldsymbol{A}_j^{\mathrm{T}} \boldsymbol{y}_0,\ j = 1, \dots, \delta \end{array} \right\} \tag{4.34}$$

is dense in $\boldsymbol{M}_2$.

Proof. As usual, let us prove density in $\mathbb{R}^N \times W^{1,2}$. Let $\boldsymbol{y} = \left[\begin{smallmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{smallmatrix}\right] \in \mathbb{R}^N \times W^{1,2}$. Let us define the following sequence of functions $f_n : [-r, 0] \to \mathbb{R}^N$, $n > \sup_{i=1,2,\dots,\delta} \frac{r}{r_i - r_{i-1}}$,

$$\begin{aligned} f_n(\vartheta) = {} & \left(\boldsymbol{y}_0 - \frac{n}{r}\vartheta \left(\boldsymbol{y}_1 \left(\frac{-r}{n} \right) - \boldsymbol{y}_0 \right) \right) \chi_{[\frac{-r}{n}, 0]} \\ & + \left(\boldsymbol{A}_\delta^{\mathrm{T}} \boldsymbol{y}_0 + \frac{n}{r}(\vartheta + r) \left(\boldsymbol{y}_1 \left(-r + \frac{r}{n} \right) - \boldsymbol{A}_\delta^{\mathrm{T}} \boldsymbol{y}_0 \right) \right) \chi_{[-r, -r + \frac{r}{n}]} \\ & + \sum_{i=1}^{\delta} \left(\boldsymbol{A}_i^{\mathrm{T}} \boldsymbol{y}_0 + \frac{n}{r}(\vartheta + r_i) \left(\boldsymbol{y}_1 \left(-r_i + \frac{r}{n} \right) - \boldsymbol{A}_i^{\mathrm{T}} \boldsymbol{y}_0 \right) \right) \chi_{[-r_i, -r_i + \frac{r}{n}]} \\ & \quad + \left(\boldsymbol{A}_i^{\mathrm{T}} \boldsymbol{y}_0 - \frac{n}{r}(\vartheta + r_i) \left(\boldsymbol{y}_1 \left(-r_i - \frac{r}{n} \right) - \boldsymbol{A}_i^{\mathrm{T}} \boldsymbol{y}_0 \right) \right) \chi_{[-r_i - \frac{r}{n}, -r_i)}. \end{aligned} \tag{4.35}$$

Consider the sequence of elements in U,

$$\boldsymbol{y}_n = \begin{bmatrix} \boldsymbol{y}_0 \\ f_n + \sum_{i=1}^{\delta} \boldsymbol{y}_1 \chi_{(-r_i + \frac{r}{n}, -r_{i-1} - \frac{r}{n})} \end{bmatrix}. \tag{4.36}$$

As $\boldsymbol{y}_1$ is bounded, f_n is bounded too, uniformly on n. It follows that

$$\|\boldsymbol{y}_n - \boldsymbol{y}\|^2 \le \left(\sup_{\vartheta, n} \|\boldsymbol{y}_1(\vartheta) - f_n(\vartheta)\|^2 \right) \frac{2\delta r}{n}. \qquad \square \tag{4.37}$$

Remark 4.10. The previous lemma proves that the intersection between the domain of $\boldsymbol{A}$ and the domain of $\boldsymbol{A}^*$ is dense in $\boldsymbol{M}_2$ if the weighted inner product is used. See that the subspace U is contained in both the domains. It is a standard result that such an intersection is in general not dense if the usual inner product is used [11, 14, 24, 27, 43].

THEOREM 4.11. *The sequence of orthoprojection operators* $\boldsymbol{\Pi}'_n : \boldsymbol{M}_2 \mapsto \Phi'_n$ *converges strongly to the identity operator.*

Proof. It is sufficient to prove strong convergence in a dense subspace of $\boldsymbol{M}_2$. Therefore, consider the subspace U in (4.34).

It is shown below that for any $\boldsymbol{y} = \left[\begin{smallmatrix}\boldsymbol{y}_0\\ \boldsymbol{y}_1\end{smallmatrix}\right] \in U$ there exists a sequence of approximations $\boldsymbol{y}^n \in \Phi'_n$ such that $\lim_{n\to\infty} \|\boldsymbol{y}^n - \boldsymbol{y}\|_{\boldsymbol{M}_2} = 0$. Consider the following definition of $\boldsymbol{y}^n \in \Phi'_n$:

$$\begin{aligned}\boldsymbol{y}^n &= \sum_{k=1}^{N}(\boldsymbol{y}_0^{\mathrm{T}}\phi_k)w'_k + \sum_{k=1}^{N}\sum_{i=1}^{\delta-1}(\boldsymbol{y}_1(-r_i)^{\mathrm{T}}\phi_k)w_k^{r_i},\\ &\quad + \sum_{i=1}^{\delta}\sum_{k=1}^{N}\sum_{j=1}^{n_i-1}(\boldsymbol{y}_1(t_j^i)^{\mathrm{T}}\phi_k)w^i_{jN+k} + \sum_{k=1}^{N}(\boldsymbol{y}_1(0)^{\mathrm{T}}\phi_k)w_k^1.\end{aligned} \tag{4.38}$$

It is, by substituting expressions of vectors generating the subspace Φ'_n (4.6), (4.7), (4.8), (4.11),

$$\begin{aligned}\boldsymbol{y}^n = \begin{bmatrix}\boldsymbol{y}_0^n\\ \boldsymbol{y}_1^n\end{bmatrix} &= \begin{bmatrix}\boldsymbol{y}_0\\ \displaystyle\sum_{k=1}^{N}(\boldsymbol{y}_1(-r)^{\mathrm{T}}\phi_k)\phi_k\, spline^{\delta}_{n_\delta}\end{bmatrix}\\ &+ \begin{bmatrix}0\\ \displaystyle\sum_{k=1}^{N}\sum_{j=1}^{\delta-1}\left(\frac{(\boldsymbol{y}_0^{\mathrm{T}}\phi_k)a_{jk}}{\delta-j+1}spline''_{r_j} + (\boldsymbol{y}_1(-r_j)^{\mathrm{T}}\phi_k)\phi_k\, spline'_{r_j}\right)\end{bmatrix}\\ &+ \begin{bmatrix}0\\ \displaystyle\sum_{i=1}^{\delta}\sum_{k=1}^{N}\sum_{j=1}^{n_i-1}(\boldsymbol{y}_1(t_j^i)^{\mathrm{T}}\phi_k)\phi_k\, spline^i_j\end{bmatrix} + \begin{bmatrix}0\\ \displaystyle\sum_{k=1}^{N}(\boldsymbol{y}_1(0)^{\mathrm{T}}\phi_k)\phi_k\, spline^1_0\end{bmatrix}.\end{aligned} \tag{4.39}$$

Moreover, it is readily recognized that

$$\sum_{k=1}^{N}(\boldsymbol{y}_0^{\mathrm{T}}\phi_k)a_{AK} = \boldsymbol{A}_j^{\mathrm{T}}\boldsymbol{y}_0 = \sum_{k=1}^{N}(\boldsymbol{y}_1(-r_j)^{\mathrm{T}}\phi_k)\phi_k, \quad j = 1,\dots,\delta, \tag{4.40}$$

$$\frac{1}{\delta-j+1}spline''_{r_j} + spline'_{r_j} = spline_{r_j}, \quad j = 1,\dots,\delta-1, \tag{4.41}$$

so that

$$\begin{aligned}\|\boldsymbol{y}^n - \boldsymbol{y}\|_{\boldsymbol{M}_2} = \Bigg\| &\sum_{k=1}^{N}(\boldsymbol{y}_1(-r)^{\mathrm{T}}\phi_k)\phi_k\, spline^{\delta}_{n_\delta}\\ &+ \sum_{j=1}^{\delta-1}\boldsymbol{A}_j^{\mathrm{T}}\boldsymbol{y}_0\Big(\frac{1}{\delta-j+1}spline''_{r_j} + spline'_{r_j}\Big)\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{\delta}\sum_{k=1}^{N}\sum_{j=1}^{n_i-1}(\boldsymbol{y}_1(t_j^i)^{\mathrm{T}}\phi_k)\phi_k\, spline_j^i\\
&+\sum_{k=1}^{N}(\boldsymbol{y}_1(0)^{\mathrm{T}}\phi_k)\phi_k\, spline_0^1-\boldsymbol{y}_1\Big\|_{L_2}\\
&=\Big\|\sum_{k=1}^{N}(\boldsymbol{y}_1(-r)^{\mathrm{T}}\phi_k)\phi_k\, spline_{n_\delta}^{\delta}+\sum_{j=1}^{\delta-1}\boldsymbol{y}_1(-r_j)\, spline_{r_j}\\
&+\sum_{i=1}^{\delta}\sum_{k=1}^{N}\sum_{j=1}^{n_i-1}(\boldsymbol{y}_1(t_j^i)^{\mathrm{T}}\phi_k)\phi_k\, spline_j^i\\
&+\sum_{k=1}^{N}(\boldsymbol{y}_1(0)^{\mathrm{T}}\phi_k)\phi_k\, spline_0^1-\boldsymbol{y}_1\Big\|_{L_2}\\
&=\Big\|\sum_{i=1}^{\delta}\sum_{k=1}^{N}\sum_{j=0}^{n_i}(\boldsymbol{y}_1(t_j^i)^{\mathrm{T}}\phi_k)\phi_k\, spline_j^i-\boldsymbol{y}_1\Big\|_{L_2}\\
&=\sum_{i=1}^{\delta}\Big\|\sum_{k=1}^{N}\sum_{j=0}^{n_i}(\boldsymbol{y}_1(t_j^i)^{\mathrm{T}}\phi_k)\phi_k\, spline_j^i-\boldsymbol{y}_1\cdot\chi_{[-r_i,-r_{i-1}]}\Big\|_{L_2},
\end{aligned}
\tag{4.42}
$$

which gives the norm of the error between a function $\boldsymbol{y}_1\in W^{1,2}$ and its approximation with first order splines in which the value at each spline knot (the instants t_j^i) is exactly the value of the function at time t_j^i. It is a standard result that the error tends to zero in L_2 norm for $n\to\infty$ (Theorem 2.4 in [42]) and therefore $\lim_{n\to\infty}\|\boldsymbol{y}^n-\boldsymbol{y}\|_{\boldsymbol{M}_2}=0$. This implies, by property (4.30), the strong convergence to identity of operator $\boldsymbol{\Pi}_n'$. □

LEMMA 4.12. *There exists a real constant α such that the operator $\boldsymbol{A}^*-\alpha\boldsymbol{I}$ and operators $\boldsymbol{\Pi}_n'\boldsymbol{A}^*\boldsymbol{\Pi}_n'-\alpha\boldsymbol{I}$ are dissipative.*

Proof. In [3] it has been proved that there exists α such that operator $\boldsymbol{A}-\alpha\boldsymbol{I}$ is dissipative and therefore generates a semigroup which is a contraction one. This implies that the adjoint semigroup is a contraction one too and therefore its infinitesimal generator $\boldsymbol{A}^*-\alpha\boldsymbol{I}$ is dissipative [1]. Dissipativity of $\boldsymbol{A}^*-\alpha\boldsymbol{I}$ implies that for any n the operator $\boldsymbol{\Pi}_n'\boldsymbol{A}^*\boldsymbol{\Pi}_n'-\alpha\boldsymbol{I}$ is dissipative. This happens because for any $\boldsymbol{x}\in\boldsymbol{M}_2$

$$
\begin{aligned}
\big((\boldsymbol{\Pi}_n'\boldsymbol{A}^*\boldsymbol{\Pi}_n'-\alpha\boldsymbol{I})\boldsymbol{x},\boldsymbol{x}\big)&=(\boldsymbol{A}^*\boldsymbol{\Pi}_n'\boldsymbol{x},\boldsymbol{\Pi}_n'\boldsymbol{x})-\alpha(\boldsymbol{x},\boldsymbol{x})\\
&\le(\boldsymbol{A}^*\boldsymbol{\Pi}_n'\boldsymbol{x},\boldsymbol{\Pi}_n'\boldsymbol{x})-\alpha(\boldsymbol{\Pi}_n'\boldsymbol{x},\boldsymbol{\Pi}_n'\boldsymbol{x})\\
&=\big((\boldsymbol{A}^*-\alpha\boldsymbol{I})\boldsymbol{\Pi}_n'\boldsymbol{x},\boldsymbol{\Pi}_n'\boldsymbol{x}\big)\le 0.
\end{aligned}
\tag{4.43}
$$

□

LEMMA 4.13. *Let $\boldsymbol{D}$ be the dense subspace of $\boldsymbol{M}_2$ defined as*

$$
\boldsymbol{D}=\left\{\begin{bmatrix}\boldsymbol{y}_0\\ \boldsymbol{y}_1\end{bmatrix}\;\middle|\;\begin{array}{c}\boldsymbol{y}_0\in\mathbb{R}^N,\quad \boldsymbol{A}_\delta^{\mathrm{T}}\boldsymbol{y}_0=\boldsymbol{y}_1(-r),\\ \left(\boldsymbol{y}_1-\displaystyle\sum_{j=1}^{\delta-1}\boldsymbol{k}_j(\boldsymbol{y}_0,\boldsymbol{y}_1)\chi_{[-r,-r_j]}\right)\in C^2\end{array}\right\}.
\tag{4.44}
$$

Then there exists $\lambda>0$ such that $(\boldsymbol{A}^*-\lambda\boldsymbol{I})\boldsymbol{D}$ is dense in $\boldsymbol{M}_2$.

Proof. Let us at first assume the following additional property on the term $\boldsymbol{A}_{01}$ in the definition of operator $\boldsymbol{A}^*$:

$$(\mathrm{Hp}_0): \quad \frac{1}{g}\boldsymbol{A}_{01}^{\mathrm{T}} \quad \text{is a matrix of functions in } C^1([-r,0];\mathbb{R}),$$

where g is the weighting function in the inner product (2.4).

Hypothesis (Hp_0) will be removed at the end of the proof.

First, it will be shown that under assumption (Hp_0) there exists a sufficiently large λ and matrices $P_0^j(\lambda) \in \mathbb{R}^{N\times N}$ and $P_1^j(\lambda) \in \mathbb{R}^{N\times \delta N}$, $j = 1, 2, \ldots, \delta - 1$, such that for any $\boldsymbol{z} = \left[\begin{smallmatrix}\boldsymbol{z}_0\\ \boldsymbol{z}_1\end{smallmatrix}\right] \in \mathbb{R}^N \times C^1([-r,0];\mathbb{R}^N)$ there exists $\boldsymbol{y} \in \boldsymbol{D}$ such that

$$(4.45)\qquad \begin{gathered}(\boldsymbol{A}^* - \lambda\boldsymbol{I})\boldsymbol{y} = \begin{bmatrix}\boldsymbol{z}_0\\ \tilde{\boldsymbol{z}}_1(\boldsymbol{z};\lambda)\end{bmatrix},\\ \text{where}\quad \tilde{\boldsymbol{z}}_1(\boldsymbol{z};\lambda) = \boldsymbol{z}_1 + \sum_{j=1}^{\delta-1}\big(P_0^j(\lambda)\boldsymbol{z}_0 + P_1^j(\lambda)F_\lambda(\boldsymbol{z}_1)\big)\chi_{[-r,-r_j]},\end{gathered}$$

in which the linear functional $F_\lambda(\boldsymbol{z}_1) : C^1([-r,0];\mathbb{R}^N) \mapsto \mathbb{R}^{N\delta}$ is defined as follows:

$$(4.46)\qquad F_\lambda(\boldsymbol{z}_1) = \begin{pmatrix}\displaystyle\int_{-r}^{0} e^{\lambda\tau}\boldsymbol{z}_1(\tau)d\tau\\ \displaystyle\int_{-r}^{-r_1} e^{-\lambda(-r_1-\tau)}\boldsymbol{z}_1(\tau)d\tau\\ \vdots\\ \displaystyle\int_{-r}^{-r_{\delta-1}} e^{-\lambda(-r_{\delta-1}-\tau)}\boldsymbol{z}_1(\tau)d\tau\end{pmatrix}.$$

Next, it will be shown that there exists a sufficiently large λ such that for any given $\boldsymbol{x} = \left[\begin{smallmatrix}\boldsymbol{x}_0\\ \boldsymbol{x}_1\end{smallmatrix}\right] \in \boldsymbol{M}_2$ and for any $\varepsilon > 0$ there exists $\boldsymbol{z} = \left[\begin{smallmatrix}\boldsymbol{z}_0\\ \boldsymbol{z}_1\end{smallmatrix}\right] \in \mathbb{R}^N \times C^1([-r,0];\mathbb{R}^N)$ such that

$$(4.47)\qquad \left\| \begin{bmatrix}\boldsymbol{x}_0\\ \boldsymbol{x}_1\end{bmatrix} - \begin{bmatrix}\boldsymbol{z}_0\\ \tilde{\boldsymbol{z}}_1(\boldsymbol{z};\lambda)\end{bmatrix} \right\|_{\boldsymbol{M}_2} \le \varepsilon$$

and therefore, from (4.45),

$$(4.48)\qquad \forall \boldsymbol{x} \in \boldsymbol{M}_2,\ \forall\varepsilon > 0, \quad \exists \boldsymbol{y} \in \boldsymbol{D}: \ \big\|\boldsymbol{x} - (\boldsymbol{A}^* - \lambda\boldsymbol{I})\boldsymbol{y}\big\|_{\boldsymbol{M}_2} \le \varepsilon,$$

that is, the density of $(\boldsymbol{A}^* - \lambda\boldsymbol{I})\boldsymbol{D}$ in $\boldsymbol{M}_2$.

In order to prove (4.45) as a first step it is shown how to find a function $\bar{Y}_1(\boldsymbol{y}_0, \boldsymbol{z}_1)$ such that for any $\boldsymbol{y}_0 \in \mathbb{R}^N$ and $\boldsymbol{z}_1 \in C^1([-r,0];\mathbb{R}^N)$ it is $\left[\begin{smallmatrix}\boldsymbol{y}_0\\ \bar{Y}_1(\boldsymbol{y}_0,\boldsymbol{z}_1)\end{smallmatrix}\right] \in \boldsymbol{D}$. Next, it is shown how to define a function $Y_0(\boldsymbol{z}_0, \boldsymbol{z}_1) : \mathbb{R}^N \times C^1([-r,0];\mathbb{R}^N) \to \mathbb{R}^N$ such that the composed function $Y_1(\boldsymbol{z}_0, \boldsymbol{z}_1) = \bar{Y}_1(Y_0(\boldsymbol{z}_0, \boldsymbol{z}_1), \boldsymbol{z}_1)$ has the property

$$(4.49)\qquad (\boldsymbol{A}^* - \lambda\boldsymbol{I})\begin{bmatrix}Y_0(\boldsymbol{z}_0,\boldsymbol{z}_1)\\ Y_1(\boldsymbol{z}_0,\boldsymbol{z}_1)\end{bmatrix} = \begin{bmatrix}\boldsymbol{z}_0\\ \tilde{\boldsymbol{z}}_1(\boldsymbol{z}_1;\lambda)\end{bmatrix}.$$

For any given pair $\boldsymbol{y}_0 \in \mathbb{R}^N$ and $\boldsymbol{z}_1 \in C^1([-r,0];\mathbb{R}^N)$ let us consider the differential equation in $C^2([-r,0];\mathbb{R}^N)$,

$$(4.50)\qquad \frac{d}{d\vartheta}\boldsymbol{f}(\vartheta) + \lambda\boldsymbol{f}(\vartheta) = -\Big(\boldsymbol{z}_1(\vartheta) - \frac{1}{g(\vartheta)}\boldsymbol{A}_{01}^{\mathrm{T}}(\vartheta)\boldsymbol{y}_0\Big),$$

whose solution is

$$\boldsymbol{f}(\vartheta) = e^{-\lambda(\vartheta+r)}\boldsymbol{f}(-r) - \int_{-r}^{\vartheta} e^{-\lambda(\vartheta-\tau)}\Big(\boldsymbol{z}_1(\tau) - \frac{1}{g(\tau)}\boldsymbol{A}_{01}^{\mathrm{T}}(\tau)\boldsymbol{y}_0\Big)d\tau. \tag{4.51}$$

By Lemma A.1 in the appendix, there exists a unique left-continuous function $\boldsymbol{y}_1$ that satisfies condition (A.1), with $\boldsymbol{f}$ given by (4.51). Such $\boldsymbol{y}_1$ is given by expression (A.6), and its values at the delay instants are such that

$$(\boldsymbol{I}_{(\delta-1)N} - H_{\delta,2})\begin{bmatrix} \boldsymbol{y}_1(-r_1) \\ \vdots \\ \boldsymbol{y}_1(-r_{\delta-1}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}(-r_1) \\ \vdots \\ \boldsymbol{f}(-r_{\delta-1}) \end{bmatrix} - H_{\delta,2}\begin{bmatrix} \boldsymbol{A}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{A}_{\delta-1}^{\mathrm{T}} \end{bmatrix}\boldsymbol{y}_0, \tag{4.52}$$

where $H_{\delta,2}$ is defined in (A.4), Lemma A.1, in the appendix.

In order to guarantee that $\boldsymbol{y} = \left[\begin{smallmatrix}\boldsymbol{y}_0 \\ \boldsymbol{y}_1\end{smallmatrix}\right] \in \boldsymbol{D}$, $\boldsymbol{y}_1$ must satisfy the additional condition

$$\boldsymbol{y}_1(-r) = \boldsymbol{A}_\delta^{\mathrm{T}}\boldsymbol{y}_0. \tag{4.53}$$

By substituting (4.53) in (A.1) one has

$$\boldsymbol{f}(-r) = \boldsymbol{A}_\delta^{\mathrm{T}}\boldsymbol{y}_0 - \sum_{j=1}^{\delta-1}\frac{\boldsymbol{y}_1(-r_j) - \boldsymbol{A}_j^{\mathrm{T}}\boldsymbol{y}_0}{\delta - j + 1}, \tag{4.54}$$

which can be rewritten as

$$\boldsymbol{f}(-r) = h_{\delta,1}\begin{bmatrix} \boldsymbol{y}_1(-r_1) \\ \vdots \\ \boldsymbol{y}_1(-r_{\delta-1}) \end{bmatrix} - h_{\delta,2}\begin{bmatrix} \boldsymbol{A}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{A}_{\delta-1}^{\mathrm{T}} \end{bmatrix}\boldsymbol{y}_0, \tag{4.55}$$

where

$$\begin{aligned} h_{\delta,1} &= \left[\frac{1}{\delta}\boldsymbol{I}_N \quad \cdots \quad \frac{1}{2}\boldsymbol{I}_N \quad \boldsymbol{I}_N\right], \\ h_{\delta,2} &= \left[\frac{1}{\delta}\boldsymbol{I}_N \quad \cdots \quad \frac{1}{2}\boldsymbol{I}_N\right]. \end{aligned} \tag{4.56}$$

By (4.51) the values of function $\boldsymbol{f}$ at the delay instants are as follows:

$$\begin{bmatrix} \boldsymbol{f}(-r_1) \\ \vdots \\ \boldsymbol{f}(-r_{\delta-1}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_N e^{-\lambda(r-r_1)} \\ \vdots \\ \boldsymbol{I}_N e^{-\lambda(r-r_{\delta-1})} \end{bmatrix} \boldsymbol{f}(-r) \quad - \left[0_{(\delta-1)N\times N}\ \boldsymbol{I}_{(\delta-1)N}\right] F_\lambda\left(\boldsymbol{z}_1 - \frac{1}{g}A_{01}^{\mathrm{T}}\boldsymbol{y}_0\right). \tag{4.57}$$

By substituting (4.52) and (4.55) into (4.57) and rearranging we have

$$H_p(\lambda)\begin{bmatrix} \boldsymbol{y}_1(-r_1) \\ \vdots \\ \boldsymbol{y}_1(-r_{\delta-1}) \end{bmatrix} = H_q(\lambda)\begin{bmatrix} \boldsymbol{A}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{A}_{\delta-1}^{\mathrm{T}} \end{bmatrix} \boldsymbol{y}_0 - \left[0_{(\delta-1)N\times N}\ \boldsymbol{I}_{(\delta-1)N}\right]\left(F_\lambda(\boldsymbol{z}_1) - F_\lambda\left(\frac{1}{g}A_{01}^{\mathrm{T}}\right)\boldsymbol{y}_0\right), \tag{4.58}$$

in which matrices H_p and H_q are defined as

$$H_p(\lambda) = \boldsymbol{I}_{(\delta-1)N} - H_{\delta,2} + \begin{bmatrix} \boldsymbol{I}_N e^{-\lambda(r-r_1)} \\ \vdots \\ \boldsymbol{I}_N e^{-\lambda(r-r_{\delta-1})} \end{bmatrix} h_{\delta,2}, \tag{4.59}$$

$$H_q(\lambda) = \begin{bmatrix} \boldsymbol{I}_N e^{-\lambda(r-r_1)} \\ \vdots \\ \boldsymbol{I}_N e^{-\lambda(r-r_{\delta-1})} \end{bmatrix} h_{\delta,1} - \begin{bmatrix} H_{\delta,2} \\ 0_{N\times\delta N} \end{bmatrix}. \tag{4.60}$$

Because H_p is nonsingular (Lemma A.3 in the appendix), by (4.58) it results that

$$\begin{aligned} \begin{bmatrix} \boldsymbol{y}_1(-r_1) \\ \vdots \\ \boldsymbol{y}_1(-r_{\delta-1}) \end{bmatrix} = &\left(H_p^{-1}(\lambda) H_q(\lambda) \begin{bmatrix} \boldsymbol{A}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{A}_{\delta-1}^{\mathrm{T}} \end{bmatrix} + \begin{bmatrix} 0_{(\delta-1)N\times N} & H_p^{-1}(\lambda) \end{bmatrix} F_\lambda \left(\frac{1}{g} \boldsymbol{A}_{01}^{\mathrm{T}} \right) \right) \boldsymbol{y}_0 \\ &- \begin{bmatrix} 0_{(\delta-1)N\times N} & H_p^{-1}(\lambda) \end{bmatrix} F_\lambda(\boldsymbol{z}_1). \end{aligned} \tag{4.61}$$

From (4.61) and (4.53), recalling that $r_\delta = r$, matrices $N_j(\lambda)$ and $M_j(\lambda)$, $j = 1, \dots, \delta$ are defined such that

$$\boldsymbol{y}_1(-r_j) = N_j(\lambda)\boldsymbol{y}_0 + M_j(\lambda) F_\lambda(\boldsymbol{z}_1). \tag{4.62}$$

The left-continuous function $\boldsymbol{y}_1 = \bar{Y}_1(\boldsymbol{y}_0, \boldsymbol{z}_1)$ we were looking for is given by

$$\boldsymbol{y}_1(\vartheta) = \begin{cases} \boldsymbol{y}_1(-r_i), & \vartheta = -r_i, \\ \boldsymbol{f}(\vartheta) + \displaystyle\sum_{j=1}^{\delta-1} \frac{\boldsymbol{y}_1(-r_j) - \boldsymbol{A}_j^{\mathrm{T}} \boldsymbol{y}_0}{\delta - j + 1} \chi_{[-r,-r_j]}(\vartheta), & \vartheta \neq -r_i, \end{cases} \quad i = 1, \dots, \delta - 1, \tag{4.63}$$

in which $\boldsymbol{f}(\vartheta)$ is given by (4.51). This is such that $\left[\begin{smallmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{smallmatrix}\right] \in \boldsymbol{D}$. Let $(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y} = \left[\begin{smallmatrix} [(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_0 \\ [(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_1 \end{smallmatrix}\right]$. It is

$$\begin{aligned} [(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_1 &= \frac{1}{g} \boldsymbol{A}_{01}^{\mathrm{T}} \boldsymbol{y}_0 - \frac{d}{d\vartheta} \left(\boldsymbol{y}_1 - \sum_{j=1}^{\delta-1} \frac{\boldsymbol{y}_1(-r_j) - \boldsymbol{A}_j^{\mathrm{T}} \boldsymbol{y}_0}{\delta - j + 1} \chi_{[-r,-r_j]} \right) - \lambda \boldsymbol{y}_1 \\ &= \frac{1}{g} \boldsymbol{A}_{01}^{\mathrm{T}} \boldsymbol{y}_0 - \frac{d}{d\vartheta} \boldsymbol{f} - \lambda \boldsymbol{f} - \lambda \sum_{j=1}^{\delta-1} \frac{\boldsymbol{y}_1(-r_j) - \boldsymbol{A}_j^{\mathrm{T}} \boldsymbol{y}_0}{\delta - j + 1} \chi_{[-r,-r_j]}. \end{aligned} \tag{4.64}$$

Finally, recalling the definition (4.50) of function $\boldsymbol{f}$, it is

$$[(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_1 = \boldsymbol{z}_1 - \lambda \sum_{j=1}^{\delta-1} \frac{\left(N_j(\lambda) - \boldsymbol{A}_j^{\mathrm{T}}\right)\boldsymbol{y}_0 + M_j(\lambda) F_\lambda(\boldsymbol{z}_1)}{\delta - j + 1} \chi_{[-r,-r_j]}. \tag{4.65}$$

Until now we have showed that, for any $\boldsymbol{y}_0 \in \mathbb{R}^N$ and for any $\boldsymbol{z}_1 \in C^1([-r, 0]; \mathbb{R}^N)$ it is possible to find $\boldsymbol{y}_1$ such that $\boldsymbol{y} = \left[\begin{smallmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{smallmatrix}\right] \in \boldsymbol{D}$ and satisfies (4.65).

Now, using the computed $\boldsymbol{y}_1 = \bar{Y}_1(\boldsymbol{y}_0, \boldsymbol{z}_1)$, we are ready to prove that there exists a positive λ such that for any $\boldsymbol{z}_0 \in \mathbb{R}^N$ and $\boldsymbol{z}_1 \in C^1([-r,0];\mathbb{R}^N)$, a $\boldsymbol{y}_0$ can be found such that $\boldsymbol{y} = \left[\begin{smallmatrix}\boldsymbol{y}_0\\ \boldsymbol{y}_1\end{smallmatrix}\right] \in \boldsymbol{D}$ and satisfies (4.45). The application of operator $(\boldsymbol{A}^* - \lambda \boldsymbol{I})$ gives, for the part in $\mathbb{R}^N$,

$$[(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_0 = \delta\, \boldsymbol{y}_1(0) + \boldsymbol{A}_0^{\mathrm{T}} \boldsymbol{y}_0 - \lambda \boldsymbol{y}_0. \tag{4.66}$$

Note that from (4.63) $\boldsymbol{y}_1(0) = \boldsymbol{f}(0)$ and evaluation of $\boldsymbol{f}(0)$ according to (4.51), in which expression (4.54) of $\boldsymbol{f}(-r)$ is substituted, gives

$$[(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_0 = Q_0(\lambda)\boldsymbol{y}_0 + Q_1(\lambda) F_\lambda(\boldsymbol{z}_1), \tag{4.67}$$

in which

$$\begin{aligned} Q_0(\lambda) &= \delta e^{-\lambda r} \boldsymbol{A}_\delta^{\mathrm{T}} - \delta e^{-\lambda r} \sum_{j=1}^{\delta-1} \frac{N_j(\lambda) - \boldsymbol{A}_j^{\mathrm{T}}}{(\delta - j + 1)} + \delta \int_{-r}^{0} e^{\lambda \tau} \frac{1}{g} \boldsymbol{A}_{01}^{\mathrm{T}} d\tau + \boldsymbol{A}_0^{\mathrm{T}} - \lambda \boldsymbol{I}_N, \\ Q_1(\lambda) &= -\delta \left[\boldsymbol{I}_{N\times N} 0_{N \times N(\delta-1)} \right] - \sum_{j=1}^{\delta-1} \frac{\delta e^{-\lambda r}}{\delta - j + 1} M_j(\lambda). \end{aligned} \tag{4.68}$$

It is clear that there exists a sufficiently large λ such that $Q_0(\lambda)$ is nonsingular, due to the presence of the term $-\lambda \boldsymbol{I}_N$ (the other terms are all bounded functions of λ). Therefore, given $\boldsymbol{z}_0 \in \mathbb{R}^N$ and $\boldsymbol{z}_1 \in C^1([-r,0];\mathbb{R}^N)$, the function $Y_0(\boldsymbol{z}_0, \boldsymbol{z}_1) = \boldsymbol{y}_0 \in \mathbb{R}^N$ such that $[(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_0 = \boldsymbol{z}_0$, thanks to (4.67), is given by

$$\boldsymbol{y}_0 = Y_0(\boldsymbol{z}_0, \boldsymbol{z}_1) = Q_0^{-1}(\lambda)(\boldsymbol{z}_0 - Q_1(\lambda) F_\lambda(\boldsymbol{z}_1)). \tag{4.69}$$

Substitution of (4.69) in the expression (4.65) for $[(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_1$ gives

$$\begin{aligned} [(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}]_1 = \boldsymbol{z}_1 &- \lambda \sum_{j=1}^{\delta-1} \frac{\left(N_j(\lambda) - \boldsymbol{A}_j^{\mathrm{T}}\right) Q_0^{-1}(\lambda) \boldsymbol{z}_0}{(\delta - j + 1)} \chi_{[-r,-r_j]} \\ &- \lambda \sum_{j=1}^{\delta-1} \frac{\left(N_j(\lambda) - \boldsymbol{A}_j^{\mathrm{T}}\right) Q_0^{-1}(\lambda) Q_1(\lambda) F_\lambda(\boldsymbol{z}_1) + M_j(\lambda) F_\lambda(\boldsymbol{z}_1)}{(\delta - j + 1)} \chi_{[-r,-r_j]}. \end{aligned} \tag{4.70}$$

This expression allows one to define the matrices $P_0^j(\lambda)$ and $P_1^j(\lambda)$ used in (4.45) as

$$\begin{aligned} P_0^j &= -\lambda \frac{\left(N_j(\lambda) - \boldsymbol{A}_j^{\mathrm{T}}\right) Q_0^{-1}(\lambda)}{(\delta - j + 1)}, \\ P_1^j &= -\lambda \frac{\left(N_j(\lambda) - \boldsymbol{A}_j^{\mathrm{T}}\right) Q_0^{-1}(\lambda) Q_1(\lambda) + M_j(\lambda)}{(\delta - j + 1)}. \end{aligned} \tag{4.71}$$

Composition of functions $\bar{Y}_1(\boldsymbol{y}_0, \boldsymbol{z}_1)$ and $Y_0(\boldsymbol{z}_0, \boldsymbol{z}_1)$ gives the announced function $Y_1(\boldsymbol{z}_0, \boldsymbol{z}_1)$. This concludes the proof that, for λ sufficiently large, for any $\left[\begin{smallmatrix}\boldsymbol{z}_0\\ \boldsymbol{z}_1\end{smallmatrix}\right] \in \mathbb{R}^N \times C^1([-r,0];\mathbb{R}^N)$ there exists $\left[\begin{smallmatrix}\boldsymbol{y}_0\\ \boldsymbol{y}_1\end{smallmatrix}\right] \in \boldsymbol{D}$ such that (4.45) holds.

Consider now the continuous linear function Φ_λ defined as follows:

$$\begin{aligned} &\Phi^{(\lambda)} : L_2([-r,0];\mathbb{R}^N) \mapsto L_2([-r,0];\mathbb{R}^N), \\ &\Phi^{(\lambda)}(\boldsymbol{g}) = \boldsymbol{g} + \sum_{j=1}^{\delta-1} P_1^j(\lambda) F_\lambda(\boldsymbol{g}) \chi_{[-r,-r_j]}. \end{aligned} \tag{4.72}$$

Let us define the following subspace of $L_2([-r,0];\mathbb{R}^N)$:

$$\mathcal{R} = \Phi^{(\lambda)}(C^1([-r,0];\mathbb{R}^N)). \tag{4.73}$$

The proof of the lemma is obtained if the set $\mathcal{R}$ is proved to be dense in $L_2([-r,0];\mathbb{R}^N)$. This is true because it can be readily shown that density of $\mathcal{R}$ is sufficient to conclude that $\forall \boldsymbol{x} \in \boldsymbol{M}_2$, for any $\varepsilon > 0$, there exists a $\boldsymbol{y} \in \boldsymbol{D}$ such that $\|\boldsymbol{x} - (\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}\|_{\boldsymbol{M}_2} \leq \varepsilon$ (i.e., density of $(\boldsymbol{A}^* - \lambda\boldsymbol{I})\boldsymbol{D}$).

Given a $\boldsymbol{x} = \left[\begin{smallmatrix}\boldsymbol{x}_0\\ \boldsymbol{x}_1\end{smallmatrix}\right] \in \boldsymbol{M}_2$, take $\boldsymbol{y}_A \in \boldsymbol{D}$ as follows: $\boldsymbol{y}_{A,0} = Y_0(\boldsymbol{x}_0, 0) = Q_0^{-1}(\lambda)\boldsymbol{x}_0$ and $\boldsymbol{y}_{A,1} = Y_1(\boldsymbol{x}_0, 0)$. It is, by construction,

$$(\boldsymbol{A}^* - \lambda\boldsymbol{I})\boldsymbol{y}_A = \begin{bmatrix} \boldsymbol{x}_0 \\ \sum_{j=1}^{\delta-1} P_0^j(\lambda)\boldsymbol{x}_0\chi_{[-r,-r_j]} \end{bmatrix}. \tag{4.74}$$

From the density of $\mathcal{R}$, there exists $\boldsymbol{z}_{B,1} \in C^1([-r,0];\mathbb{R}^N)$ such that the function

$$\tilde{\boldsymbol{z}}_{B,1} = \boldsymbol{z}_{B,1} + \sum_{j=1}^{\delta-1} P_1^j(\lambda)F_\lambda(\boldsymbol{z}_{B,1})\chi_{[-r,-r_j]} \tag{4.75}$$

satisfies

$$\left\|\sum_{j=1}^{\delta-1} P_0^j(\lambda)\boldsymbol{x}_0\chi_{[-r,-r_j]} - \tilde{\boldsymbol{z}}_{B,1}\right\|_{L_2} \leq \frac{\varepsilon}{2}, \tag{4.76}$$

and from result (4.45) there exists $\boldsymbol{y}_B \in \boldsymbol{D}$ such that $(\boldsymbol{A}^* - \lambda\boldsymbol{I})\boldsymbol{y}_B = \left[\begin{smallmatrix}0\\ \tilde{\boldsymbol{z}}_{B,1}\end{smallmatrix}\right]$.

Exploiting again the density of $\mathcal{R}$ there exists $\boldsymbol{z}_{C,1} \in C^1([-r,0];\mathbb{R}^N)$ such that the function

$$\tilde{\boldsymbol{z}}_{C,1} = \boldsymbol{z}_{C,1} + \sum_{j=1}^{\delta-1} P_1^j(\lambda)F_\lambda(\boldsymbol{z}_{C,1})\chi_{[-r,-r_j]} \tag{4.77}$$

satisfies

$$\left\|\boldsymbol{x}_1 - \tilde{\boldsymbol{z}}_{C,1}\right\|_{L_2} \leq \frac{\varepsilon}{2}. \tag{4.78}$$

Again, from result (4.45) there exists $\boldsymbol{y}_C \in \boldsymbol{D}$ such that $(\boldsymbol{A}^* - \lambda\boldsymbol{I})\boldsymbol{y}_C = \left[\begin{smallmatrix}0\\ \tilde{\boldsymbol{z}}_{C,1}\end{smallmatrix}\right]$. It is now an easy matter to show that vector $\boldsymbol{y} = \boldsymbol{y}_A - \boldsymbol{y}_B + \boldsymbol{y}_C$ is such that

$$\begin{aligned} \|\boldsymbol{x} - (A^* - \lambda\boldsymbol{I})\boldsymbol{y}\|_{M_2} &= \|\boldsymbol{x} - (A^* - \lambda\boldsymbol{I})(\boldsymbol{y}_A - \boldsymbol{y}_B + \boldsymbol{y}_C)\|_{\boldsymbol{M}_2} \\ &\leq \left\| \begin{bmatrix} 0 \\ \boldsymbol{x}_1 - [(A^* - \lambda\boldsymbol{I})\boldsymbol{y}_A]_1 + \tilde{\boldsymbol{z}}_{B,1} - \tilde{\boldsymbol{z}}_{C,1} \end{bmatrix} \right\|_{\boldsymbol{M}_2} \\ &\leq \left\|\boldsymbol{x}_1 - \tilde{\boldsymbol{z}}_{C,1}\right\|_{L_2} + \left\|\sum_{j=1}^{\delta-1} P_0^j(\lambda)\boldsymbol{x}_0\chi_{[-r,-r_j]} - \tilde{\boldsymbol{z}}_{B,1}\right\|_{L_2} \leq \varepsilon. \end{aligned} \tag{4.79}$$

It remains to prove that $\mathcal{R}$ is dense for sufficiently large λ. We will show that if for any $\boldsymbol{f} \in L_2([-r,0];\mathbb{R}^N)$ there exists a vector $\alpha \in \mathbb{R}^{N\delta}$ such that

$$F_\lambda(\boldsymbol{f}) - F_\lambda\left(\sum_{j=1}^{\delta-1} P_1^j \alpha\chi_{[-r,-r_j]}\right) - \alpha = 0, \tag{4.80}$$

then for any $\boldsymbol{f} \in L_2([-r,0];\mathbb{R}^N)$ a sequence $\{\boldsymbol{f}_k\}$, $\boldsymbol{f}_k \in \mathcal{R}\ \forall k \geq 0$ can be found such that $\|\boldsymbol{f} - \boldsymbol{f}_k\|_{L_2} \to 0$. Existence of α in (4.80) for any $\boldsymbol{f}$ is ensured by the nonsingularity of matrix

$$\Gamma(\lambda) = \boldsymbol{I}_{N\delta\times N\delta} + \sum_{j=1}^{\delta-1} F_\lambda\big(P_1^j(\lambda)\chi_{[-r,-r_j]}\big) \tag{4.81}$$

for sufficiently large λ, and this is a sufficient condition for density of $\mathcal{R}$.

To this purpose consider a $\boldsymbol{f} \in L_2$, let α be the solution of (4.80), and define the function

$$\bar{\boldsymbol{f}} = \boldsymbol{f} - \sum_{j=1}^{\delta-1} P_1^j \alpha \chi_{[-r,-r_j]}. \tag{4.82}$$

It is such that $\Phi^{(\lambda)}(\bar{\boldsymbol{f}}) = \boldsymbol{f}$. Let $\{\boldsymbol{g}_k\}$ be a sequence of functions in $C^1([-r,0];\mathbb{R}^N)$ such that $\|\bar{\boldsymbol{f}} - \boldsymbol{g}_k\|_{L_2} \to 0$. From the continuity of function $\Phi^{(\lambda)}$ it is $\|\Phi^{(\lambda)}(\bar{\boldsymbol{f}}) - \Phi^{(\lambda)}(\boldsymbol{g}_k)\|_{L_2} \to 0$. Defining functions $\boldsymbol{f}_k = \Phi^{(\lambda)}(\boldsymbol{g}_k) \in \mathcal{R}$, the sequence $\{\boldsymbol{f}_k\}$ converges to $\Phi^{(\lambda)}(\bar{\boldsymbol{f}})$, that is, $\boldsymbol{f}$ and density of $\mathcal{R}$, under nonsingularity of $\Gamma(\lambda)$, is proved.

It remains to prove the nonsingularity of the $\delta N \times \delta N$ matrix $\Gamma(\lambda)$ defined in (4.81) for a sufficiently large λ. Such a proof is reported in [39] and is worked out by showing that $\det\big(\Gamma(\lambda)\big)$ is a continuous function of λ and that there exists the limit matrix $\bar{\Gamma} = \lim_{\lambda\to+\infty}\Gamma(\lambda)$. Such a matrix can be easily proved to be nonsingular, because it is block triangular (each block is $N \times N$), in which the diagonal consists of the following nonsingular δ blocks: block 1 is $\boldsymbol{I}_N$, block j, for $j = 2,\ldots,\delta$, is $\boldsymbol{I} + \frac{1}{\delta-j+1}\boldsymbol{I}_N$. It follows that $\lim_{\lambda\to+\infty}\det\big(\Gamma(\lambda)\big) = \det\big(\bar{\Gamma}\big) \neq 0$, and therefore there exists λ_0 such that for every $\lambda > \lambda_0$ matrix $\Gamma(\lambda)$ is nonsingular.

So, chosen λ such that $\Gamma(\lambda)$ and $Q_0(\lambda)$ are both nonsingular, the proof of this lemma is completed in the case of hypothesis Hp_0.

To remove such a hypothesis it is sufficient to consider a sequence $\boldsymbol{A}_{01}^k$ in the space $L_2([-r,0];\mathbb{R}^{N\times N})$, which converges to $\boldsymbol{A}_{01}$ and satisfies hypothesis Hp_0. Let $\boldsymbol{A}_k^*$ be the corresponding sequence of operators. Let $D_k \in L(\boldsymbol{M}_2)$ be defined as

$$D_k \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{g}(\boldsymbol{A}_{01}^k - \boldsymbol{A}_{01})^{\mathrm{T}}\boldsymbol{y}_0 \end{bmatrix}.$$

Thus $\boldsymbol{A}_k^* = \boldsymbol{A}^* + D_k$ and $\|D_k\|_{L(\boldsymbol{M}_2)} \leq \|\boldsymbol{A}_{01}^k - \boldsymbol{A}_{01}\|_{L_2([-r,0];\mathbb{R}^{N\times N})}$. From Proposition 2.3 in [5, page 28], and Theorem 1.1 in [38, page 76], it follows that any λ with $Re(\lambda) > \omega_0 + \sup_k \|D_k\|$ belongs to the resolvent set of $\boldsymbol{A}^*$ and $\boldsymbol{A}_k^*$, for every k, where ω_0 is such that $\|T(t)\| \leq e^{\omega_0 t}$, $\boldsymbol{T}(t)$ being the semigroup generated by $\boldsymbol{A}$ (see Lemma 2.3 in [3]). Let us choose a λ in the resolvent sets of $\boldsymbol{A}^*$ and $\boldsymbol{A}_k^*$, such that $(\boldsymbol{A}_k^* - \lambda\boldsymbol{I})\boldsymbol{D}$ is dense in $\boldsymbol{M}_2$ for every k. It is sufficient that corresponding matrices $\Gamma^k(\lambda)$ in (4.81) and $Q_0^k(\lambda)$ in (4.68) are nonsingular. Thus, given $\boldsymbol{x} \in M_2$, given $\epsilon > 0$, a sequence $\boldsymbol{y}_k$ can be found such that

$$\|(\boldsymbol{A}_k^* - \lambda\boldsymbol{I})\boldsymbol{y}_k - \boldsymbol{x}\| < \frac{\epsilon}{2}. \tag{4.83}$$

From

$$\|(\boldsymbol{A}^* - \lambda\boldsymbol{I})\boldsymbol{y}_k - \boldsymbol{x}\| \leq \|(\boldsymbol{A}_k^* - \lambda\boldsymbol{I})\boldsymbol{y}_k - \boldsymbol{x}\| + \|D_k\|\|\boldsymbol{y}_k\| \tag{4.84}$$

it follows that there exist k_0 such that $\|(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{y}_{k_0} - \boldsymbol{x}\| < \epsilon$, provided $\boldsymbol{y}_k$ is uniformly bounded. It is sufficient that $\|D_{k_0}\| < \frac{\epsilon}{2\sup_k \|\boldsymbol{y}_k\|}$. It remains to prove uniform boundedness of $\boldsymbol{y}_k$. Let $v_k = (\boldsymbol{A}_k^* - \lambda \boldsymbol{I})\boldsymbol{y}_k - \boldsymbol{x}$. From (4.83) it is $\|v_k\| < \frac{\epsilon}{2}$ for every k. From

$$(\boldsymbol{A}_k^* - \lambda \boldsymbol{I})^{-1} = (\boldsymbol{A}^* + D_k - \lambda \boldsymbol{I})^{-1} = [\boldsymbol{I} - (\lambda \boldsymbol{I} - \boldsymbol{A}^*)^{-1} D_k]^{-1} (\boldsymbol{A}^* - \lambda \boldsymbol{I})^{-1} \tag{4.85}$$

it follows that

$$\|(\boldsymbol{A}_k^* - \lambda \boldsymbol{I})^{-1}\| \le \|[\boldsymbol{I} - (\lambda \boldsymbol{I} - \boldsymbol{A}^*)^{-1} D_k]^{-1}\| \|(\boldsymbol{A}^* - \lambda \boldsymbol{I})^{-1}\|. \tag{4.86}$$

If k is sufficiently large such that $\|(\lambda \boldsymbol{I} - \boldsymbol{A}^*)^{-1} D_k\| \le d < 1$, the following inequality holds:

$$\|[\boldsymbol{I} - (\lambda \boldsymbol{I} - \boldsymbol{A}^*)^{-1} D_k]^{-1}\| \le \sum_{l=0}^{\infty} \|(\lambda \boldsymbol{I} - \boldsymbol{A}^*)^{-1} D_k\|^l \le \frac{1}{1-d}, \tag{4.87}$$

which proves the uniform boundedness of $\|(\boldsymbol{A}_k^* - \lambda \boldsymbol{I})^{-1}\|$. The uniform boundedness of $\boldsymbol{y}_k$ follows by

$$\boldsymbol{y}_k = (\boldsymbol{A}_k^* - \lambda \boldsymbol{I})^{-1}(v_k - \boldsymbol{x}). \tag{4.88}$$

Such a device to prove the density of $(\boldsymbol{A}^* - \lambda \boldsymbol{I})\boldsymbol{D}$ in $\boldsymbol{M}_2$ when hypothesis Hp_0 is not satisfied has been introduced in [24, Theorem 7.2] for the one delay case. □

LEMMA 4.14. *The operator* $\boldsymbol{\Pi}'_n \boldsymbol{A}^* \boldsymbol{\Pi}'_n$ *converges strongly to the operator* $\boldsymbol{A}^*$ *in the subspace* $\boldsymbol{D}$ *defined in Lemma* 4.13.

Proof. Since it is

$$\|\boldsymbol{\Pi}'_n \boldsymbol{A}^* \boldsymbol{\Pi}'_n \boldsymbol{x} - \boldsymbol{A}^* \boldsymbol{x}\|_{M_2} \le \|\boldsymbol{\Pi}'_n (\boldsymbol{A}^* \boldsymbol{\Pi}'_n - \boldsymbol{A}^*)\boldsymbol{x}\|_{M_2} + \|\boldsymbol{\Pi}'_n \boldsymbol{A}^* \boldsymbol{x} - \boldsymbol{A}^* \boldsymbol{x}\|_{M_2}, \tag{4.89}$$

and $\lim_{n\to\infty} \|\boldsymbol{\Pi}'_n \boldsymbol{y} - \boldsymbol{y}\|_{M_2} = 0$, $\forall \boldsymbol{y} \in \boldsymbol{M}_2$ (strong convergence), the lemma is proved if for any $\boldsymbol{x} \in \boldsymbol{D}$

$$\|\boldsymbol{A}^* \boldsymbol{\Pi}'_n \boldsymbol{x} - \boldsymbol{A}^* \boldsymbol{x}\| \to 0. \tag{4.90}$$

It is

$$\begin{aligned}
\|\boldsymbol{A}^* \boldsymbol{\Pi}'_n \boldsymbol{x} - \boldsymbol{A}^* \boldsymbol{x}\|^2 &= \|\delta \boldsymbol{x}_1(0) + \boldsymbol{A}_0^{\mathrm{T}} \boldsymbol{x}_0 - \delta(\boldsymbol{\Pi}'_n \boldsymbol{x})_1(0) - \boldsymbol{A}_0^{\mathrm{T}} (\boldsymbol{\Pi}'_n \boldsymbol{x})_0\|^2 \\
&\quad + \left\| \frac{1}{g} \boldsymbol{A}_{01} \boldsymbol{x}_0 - \frac{1}{g} \boldsymbol{A}_{01} (\boldsymbol{\Pi}'_n \boldsymbol{x})_0 - \frac{d}{d\vartheta} \left(\boldsymbol{x}_1 - \sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{x}_0, \boldsymbol{x}_1) \chi_{[-r,-r_j]} \right) \right. \\
&\quad \left. + \frac{d}{d\vartheta} \left((\boldsymbol{\Pi}'_n \boldsymbol{x})_1 - \sum_{j=1}^{\delta-1} \boldsymbol{k}_j((\boldsymbol{\Pi}'_n \boldsymbol{x})_0, (\boldsymbol{\Pi}'_n \boldsymbol{x})_1) \chi_{[-r,-r_j]} \right) \right\|_{L_2}^2 \\
&\le \delta^2 \|\boldsymbol{x}_1(0) - (\boldsymbol{\Pi}'_n \boldsymbol{x})_1(0)\|^2 + \left(\|\boldsymbol{A}_0^{\mathrm{T}}\| + 2 \left\| \frac{1}{g} \boldsymbol{A}_{01}^{\mathrm{T}} \right\|_{L_2} \right) \cdot \left\| \boldsymbol{x}_0 - (\boldsymbol{\Pi}'_n \boldsymbol{x})_0 \right\|^2 + 2 S_n^2(\boldsymbol{x}),
\end{aligned} \tag{4.91}$$

where

$$\begin{aligned}
S_n(\boldsymbol{x}) = \Bigg\| & \frac{d}{d\vartheta} \left(\boldsymbol{x}_1 - \sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{x}_0, \boldsymbol{x}_1) \chi_{[-r,-r_j]} \right) \\
& - \frac{d}{d\vartheta} \left((\boldsymbol{\Pi}'_n \boldsymbol{x})_1 - \sum_{j=1}^{\delta-1} \boldsymbol{k}_j((\boldsymbol{\Pi}'_n \boldsymbol{x})_0, (\boldsymbol{\Pi}'_n \boldsymbol{x})_1) \chi_{[-r,-r_j]} \right) \Bigg\|_{L_2}.
\end{aligned} \tag{4.92}$$

Strong convergence of $\mathbf{\Pi}'_n$ ensures that for $n \to \infty$ the second term in the right-hand side of (4.91) goes to zero.

To prove that the term $S_n(\boldsymbol{x})$ goes to zero too, let

$$\begin{aligned}\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{\boldsymbol{x}}_0 \\ \bar{\boldsymbol{x}}_1 \end{bmatrix} &= \sum_{k=1}^{N} \boldsymbol{x}_0(k) w'_k + \sum_{i=1}^{\delta-1} \sum_{k=1}^{N} (\boldsymbol{x}_1(-r_i)^{\mathrm{T}} \phi_k) w_k^{r_i} \\ &\quad + \sum_{i=1}^{\delta} \sum_{j=1}^{n_i} \sum_{k=1}^{N} (\boldsymbol{x}_1(t_j^i)^{\mathrm{T}} \phi_k) w_{jN+k}^i + \sum_{k=1}^{N} (\boldsymbol{x}_1(0)^{\mathrm{T}} \phi_k) w_k^1.\end{aligned} \tag{4.93}$$

It is such that $\bar{\boldsymbol{x}}_0 = \boldsymbol{x}_0$, so that $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_{\boldsymbol{M}_2} = \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1\|_{L_2}$ and therefore

$$\|\boldsymbol{x} - \mathbf{\Pi}'_n \boldsymbol{x}\|_{\boldsymbol{M}_2} \leq \|\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_1\|_{L_2}. \tag{4.94}$$

Considering that the function $\sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\cdot,\cdot)\chi_{[-r,-r_j]}$ is piecewise constant it is

$$S_n(\boldsymbol{x}) = \left\| \frac{d}{d\vartheta} \boldsymbol{x}_1 - \frac{d}{d\vartheta} (\mathbf{\Pi}'_n \boldsymbol{x})_1 \right\|_{L_2} \tag{4.95}$$

and

$$S_n(\boldsymbol{x}) \leq \left\| \frac{d}{d\vartheta} \boldsymbol{x}_1 - \frac{d}{d\vartheta} \bar{\boldsymbol{x}}_1 \right\|_{L_2} + \left\| \frac{d}{d\vartheta} \bar{\boldsymbol{x}}_1 - \frac{d}{d\vartheta} (\mathbf{\Pi}'_n \boldsymbol{x})_1 \right\|_{L_2}. \tag{4.96}$$

As for the first term at the right-hand side of inequality (4.96), since it is

$$\left\| \frac{d}{d\vartheta} \boldsymbol{x}_1 - \frac{d}{d\vartheta} \bar{\boldsymbol{x}}_1 \right\|_{L_2} = \left(\sum_{i=1}^{\delta} \left\| \frac{d}{d\vartheta} \boldsymbol{x}_1 \chi_{[-r_i,-r_{i-1}]} - \frac{d}{d\vartheta} \bar{\boldsymbol{x}}_1 \chi_{[-r_i,-r_{i-1}]} \right\|_{L_2}^2 \right)^{\frac{1}{2}}, \tag{4.97}$$

by standard results of spline analysis (see Theorem 2.5 in [42]), each term in the summation goes to zero for $n \to \infty$.

As for the second term, from definition of vectors w that generate V'_n, it is, by applying the Schmidt inequality (see Theorem 1.5 in [42]),

$$\begin{aligned}\left\| \frac{d}{d\vartheta} \bar{\boldsymbol{x}}_1 - \frac{d}{d\vartheta} (\mathbf{\Pi}'_n \boldsymbol{x})_1 \right\|_{L_2} &= \left(\sum_{i=1}^{\delta} \left\| \frac{d}{d\vartheta} \bar{\boldsymbol{x}}_1 \chi_{[-r_i,-r_{i-1}]} - \frac{d}{d\vartheta} (\mathbf{\Pi}'_n \boldsymbol{x})_1 \chi_{[-r_i,-r_{i-1}]} \right\|_{L_2}^2 \right)^{\frac{1}{2}} \\ &\leq \left(\sum_{i=1}^{\delta} \left(\sqrt{12} \frac{n_i}{r_i - r_{i-1}} \left\| \left(\bar{\boldsymbol{x}}_1 - (\mathbf{\Pi}'_n \boldsymbol{x})_1 \right) \chi_{[-r_i,-r_{i-1}]} \right\|_{L_2} \right)^2 \right)^{\frac{1}{2}}.\end{aligned}$$

For each term in the summation it is

$$\begin{aligned}\left\| \left(\bar{\boldsymbol{x}}_1 - (\mathbf{\Pi}'_n \boldsymbol{x})_1 \right) \chi_{[-r_i,-r_{i-1}]} \right\|_{L_2} &\leq \left\| \bar{\boldsymbol{x}}_1 - (\mathbf{\Pi}'_n \boldsymbol{x})_1 \right\|_{L_2} \\ &\leq \left\| \bar{\boldsymbol{x}}_1 - \boldsymbol{x}_1 \right\|_{L_2} + \left\| \boldsymbol{x}_1 - (\mathbf{\Pi}'_n \boldsymbol{x})_1 \right\|_{L_2} \leq 2\|\bar{\boldsymbol{x}}_1 - \boldsymbol{x}_1\| \\ &\leq \left(\sum_{i=1}^{\delta} \left(\left\| \left(\bar{\boldsymbol{x}}_1 - \boldsymbol{x}_1 \right) \chi_{[-r_i,-r_{i-1}]} \right\|_{L_2} \right)^2 \right)^{\frac{1}{2}}.\end{aligned} \tag{4.98}$$

Again, by standard results on spline approximation (Theorem 2.5 in [42]), each term in the summation goes to zero for $n \to \infty$. This proves that $S_n(\boldsymbol{x})$ goes to zero for $n \to \infty$.

It remains to prove that the term $\delta\|\boldsymbol{x}_1(0) - (\boldsymbol{\Pi}'_n\boldsymbol{x})_1(0)\|^2$ in the right-hand side of (4.91) goes to zero for $n \to \infty$.

First, note that being $\boldsymbol{x} \in \mathcal{D}(\boldsymbol{A}^*)$ it is such that for $i = 1, \dots, \delta - 1$

$$\boldsymbol{x}_1(-r_i^+) - \sum_{j=1}^{i-1} \boldsymbol{k}_j(\boldsymbol{x}_0, \boldsymbol{x}_1) = \boldsymbol{x}_1(-r_i) - \sum_{j=1}^{i} \boldsymbol{k}_j(\boldsymbol{x}_0, \boldsymbol{x}_1), \tag{4.99}$$

where $\boldsymbol{x}_1(-r_i^+)$ denotes the limit of $\boldsymbol{x}_1(\vartheta)$ for ϑ approaching $-r_i$ from the right (note that in general $\boldsymbol{x}_1(-r_i^+) \neq \boldsymbol{x}_1(-r_i)$). Simple computations, taking into account definition (2.22) of $\boldsymbol{k}_j$, give

$$\boldsymbol{x}_1(-r_i^+) = \frac{\delta - i + 2}{\delta - i + 1}\boldsymbol{x}_1(-r_i) - \frac{1}{\delta - i + 1}\boldsymbol{A}_i^{\mathrm{T}}\boldsymbol{x}_0. \tag{4.100}$$

Since also $\boldsymbol{\Pi}'_n\boldsymbol{x} \in \mathcal{D}(\boldsymbol{A}^*)$, it is such that

$$(\boldsymbol{\Pi}'_n\boldsymbol{x})_1(-r_i^+) = \frac{\delta - i + 2}{\delta - i + 1}(\boldsymbol{\Pi}'_n\boldsymbol{x})_1(-r_i) - \frac{1}{\delta - i + 1}\boldsymbol{A}_i^{\mathrm{T}}(\boldsymbol{\Pi}'_n\boldsymbol{x})_0. \tag{4.101}$$

At point $-r$ it is

$$\boldsymbol{x}_1(-r) = \boldsymbol{A}_\delta^{\mathrm{T}}\boldsymbol{x}_0, \qquad (\boldsymbol{\Pi}'_n\boldsymbol{x})_1(-r) = \boldsymbol{A}_\delta^{\mathrm{T}}(\boldsymbol{\Pi}'_n\boldsymbol{x})_0. \tag{4.102}$$

Since it has been proved that $\lim_{n\to\infty} \|\boldsymbol{x}_0 - (\boldsymbol{\Pi}'_n\boldsymbol{x})_0\| = 0$, from (4.102) it follows

$$\lim_{n\to\infty} \|\boldsymbol{x}_1(-r) - (\boldsymbol{\Pi}'_n\boldsymbol{x})_1(-r)\| = 0. \tag{4.103}$$

Starting from (4.103), the proof that $\|\boldsymbol{x}_1(0) - (\boldsymbol{\Pi}'_n\boldsymbol{x})_1(0)\|$ goes to zero is obtained if we prove the following recursive implication:

$$\lim_{n\to\infty} \|\boldsymbol{x}_1(-r_i) - (\boldsymbol{\Pi}'_n\boldsymbol{x})_1(-r_i)\| = 0, \tag{4.104}$$

$$\Downarrow$$

$$\lim_{n\to\infty} \|\boldsymbol{x}_1(-r_{i-1}) - (\boldsymbol{\Pi}'_n\boldsymbol{x})_1(-r_{i-1})\| = 0. \tag{4.105}$$

First note that if (4.104) is true, then comparing (4.100) and (4.101), recalling that $\|\boldsymbol{x}_0 - (\boldsymbol{\Pi}'_n\boldsymbol{x})_0\| \to 0$, it follows

$$\lim_{n\to\infty} \|\boldsymbol{x}_1(-r_i^+) - (\boldsymbol{\Pi}'_n\boldsymbol{x})_1(-r_i^+)\| = 0. \tag{4.106}$$

From (4.106), since it has been proved that in any interval $[-r_i, -r_{i-1}]$

$$\lim_{n\to\infty} \left\| \left(\frac{d}{d\vartheta}\boldsymbol{x}_1 - \frac{d}{d\vartheta}(\boldsymbol{\Pi}'_n\boldsymbol{x})_1 \right) \chi_{[-r_i, -r_{i-1}]} \right\|_{L_2} = 0, \tag{4.107}$$

implication (4.105) is easily obtained. This completes the proof of the Lemma. □

THEOREM 4.15. *The sequence of semigroups $\boldsymbol{T}^*_{\Phi'_n}$ generated by the operators $\boldsymbol{\Pi}'_n\boldsymbol{A}^*\boldsymbol{\Pi}'_n$ converges strongly to $\boldsymbol{T}^*$, the adjoint of the semigroup generated by $\boldsymbol{A}$.*

Proof. The results in Lemmas 4.12, 4.13, and 4.14 imply that the hypotheses of the Trotter–Kato theorem, as stated in [38] and reported also in Lemma 3.1 in [3], are satisfied, and this proves the convergence of $\boldsymbol{T}^*_{\Phi'_n}$ to $\boldsymbol{T}^*$. □

Now the main result of the paper can be given, that is, the theorem on the convergence of the proposed finite dimensional approximation scheme of the LQG controller for hereditary systems.

THEOREM 4.16. *Let* Φ_n *and* Φ'_n *be the sequences of finite dimension subspaces of* $\boldsymbol{M}_2$ *in Definitions* 4.2, 4.3. *Let* $\boldsymbol{u}_n(t)$ *be the input function obtained by*

$$\boldsymbol{u}_n(t) = -\boldsymbol{B}^* \boldsymbol{R}_n(t_f - t)\boldsymbol{\Pi}'_n \hat{\boldsymbol{x}}_n(t), \tag{4.108}$$

where

$$\begin{aligned} \dot{\hat{\boldsymbol{x}}}_n(t) &= \boldsymbol{\Pi}_n \boldsymbol{A}\boldsymbol{\Pi}_n \hat{\boldsymbol{x}}_n(t) + \boldsymbol{\Pi}_n \boldsymbol{B}\boldsymbol{u}_n(t) + \boldsymbol{P}_n(t)\boldsymbol{\Pi}_n \boldsymbol{C}^*\big(\boldsymbol{y}(t) - \boldsymbol{C}\boldsymbol{\Pi}_n \hat{\boldsymbol{x}}_n(t)\big),\\ \hat{\boldsymbol{x}}_n(0) &= \boldsymbol{\Pi}_n \hat{\boldsymbol{x}}(0) \end{aligned} \tag{4.109}$$

in which $\boldsymbol{P}_n$ *and* $\boldsymbol{R}_n$ *are the finite dimensional solutions of the Riccati equations* (3.15) *and* (3.16) *in which the projectors* $\boldsymbol{\Pi}_n$ *and* $\boldsymbol{\Pi}'_n$ *are considered. Let* $\boldsymbol{u}(t)$ *be the optimal input,* $\hat{\boldsymbol{x}}(t)$ *the optimal estimated state,* $\boldsymbol{x}_n(t)$ *and* $\boldsymbol{x}(t)$ *the actual state evolving when* $\boldsymbol{u}_n(t)$ *and* $\boldsymbol{u}(t)$ *are applied to system* (2.1), (2.2), *respectively.*

Then

$$\lim_{n\to\infty} E\|\boldsymbol{x}_n - \boldsymbol{x}\|^2_{L_2([0,t_f];\boldsymbol{M}_2)} = 0, \tag{4.110}$$

$$\lim_{n\to\infty} E\|\hat{\boldsymbol{x}}_n - \hat{\boldsymbol{x}}\|^2_{L_2([0,t_f];\boldsymbol{M}_2)} = 0, \tag{4.111}$$

$$\lim_{n\to\infty} E\|\boldsymbol{u}_n - \boldsymbol{u}\|^2_{L_2([0,t_f];\mathbb{R}^p)} = 0, \tag{4.112}$$

$$\lim_{n\to\infty} |J_f(\boldsymbol{u}_n) - J_f(\boldsymbol{u})| = 0. \tag{4.113}$$

Proof. The proof comes from Theorem 3.7, whose assumptions (from Hp_1 to Hp_4) are satisfied thanks to Theorems 4.7, 4.8, 4.11, and 4.15. □

5. Implementation of the method.

In this section the numerical implementation of the approximation scheme described in the previous section, and which satisfies all properties listed in the introduction, is reported.

Consider two Hilbert spaces $\mathcal{U}$ and $\mathcal{V}$ and two finite dimensional subspaces $U_n \subset \mathcal{U}$ and $V_m \subset \mathcal{V}$ of dimension n and m, respectively. Let $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_n)$ be a basis of U_n and $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_m)$ a basis of V_m. Consider the nonsingular matrices $\boldsymbol{T}_n \in \mathbb{R}^{n\times n}$ and $Z_m \in \mathbb{R}^{m\times m}$, whose components are defined as

$$\begin{aligned} T_n(i,j) &= (\boldsymbol{u}_i, \boldsymbol{u}_j)_{\mathcal{U}}, \quad i,j = 1,\dots,n,\\ Z_m(h,k) &= (\boldsymbol{v}_i, \boldsymbol{v}_j)_{\mathcal{V}}, \quad i,j = 1,\dots,m. \end{aligned} \tag{5.1}$$

Recall that the orthoprojection operator from $\mathcal{U}$ to U_n performs the following operation on a vector $\boldsymbol{x} \in \mathcal{U}$:

$$\boldsymbol{\Pi}(\boldsymbol{x}; U_n) = \sum_{i=1}^n \alpha_i \boldsymbol{u}_i \quad \text{with} \quad \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = T_n^{-1} \begin{bmatrix} (\boldsymbol{x}, \boldsymbol{u}_1)_{\mathcal{U}} \\ \vdots \\ (\boldsymbol{x}, \boldsymbol{u}_n)_{\mathcal{U}} \end{bmatrix}, \tag{5.2}$$

and the orthoprojection operator from $\mathcal{V}$ to V_m performs the following operation on a vector $\boldsymbol{y} \in H_2$:

$$\boldsymbol{\Pi}(\boldsymbol{y}; V_m) = \sum_{i=1}^m \beta_i \boldsymbol{v}_i \quad \text{with} \quad \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} = Z_m^{-1} \begin{bmatrix} (\boldsymbol{y}, \boldsymbol{v}_1)_{\mathcal{V}} \\ \vdots \\ (\boldsymbol{y}, \boldsymbol{v}_m)_{\mathcal{V}} \end{bmatrix}. \tag{5.3}$$

Let us denote as ξ_n the isomorphism that associates to a vector $\boldsymbol{x} \in U_n$ its coordinate representation

$$\xi_n : \; U_n \mapsto \mathbb{R}^n; \quad \xi_n(\boldsymbol{x}) = T_n^{-1} \begin{bmatrix} (\boldsymbol{x}, \boldsymbol{u}_1)_{\mathcal{U}} \\ \vdots \\ (\boldsymbol{x}, \boldsymbol{u}_n)_{\mathcal{U}} \end{bmatrix} \tag{5.4}$$

and as ξ_m the isomorphism

$$\xi_m : \; V_m \mapsto \mathbb{R}^m; \quad \xi_m(\boldsymbol{y}) = Z_m^{-1} \begin{bmatrix} (\boldsymbol{y}, \boldsymbol{v}_1)_{\mathcal{V}} \\ \vdots \\ (\boldsymbol{y}, \boldsymbol{v}_m)_{\mathcal{V}} \end{bmatrix}. \tag{5.5}$$

Consider now the algebra $\mathcal{S}$ of linear operators from U_n to V_m. It is

$$S \in \mathcal{S}, \quad \xi_m(S(\boldsymbol{u}_i)) = Z_m^{-1} \begin{bmatrix} (S(\boldsymbol{u}_i), \boldsymbol{v}_1)_{\mathcal{V}} \\ \vdots \\ (S(\boldsymbol{u}_i), \boldsymbol{v}_m)_{\mathcal{V}} \end{bmatrix}. \tag{5.6}$$

The following isomorphism η_n^m is induced between $\mathcal{S}$ and the algebra of matrices $m \times n$:

$$S \in \mathcal{S}, \;\; \eta_n^m(S) = Z_m^{-1}\bar{S}, \quad \bar{S} \in \mathbb{R}^{m \times n}, \quad \bar{S}_{i,j} = (S(\boldsymbol{u}_j), \boldsymbol{v}_i)_{\mathcal{V}}, \tag{5.7}$$

that is, such that

$$\xi_m(S(\boldsymbol{x})) = \eta_n^m(S)\, \xi_n(\boldsymbol{x}), \quad \boldsymbol{x} \in U_n. \tag{5.8}$$

Isomorphisms between points of finite dimensional spaces and their coordinate representations and between linear operators on finite dimensional spaces and their matrix representations allow us to write the approximated Riccati equations for control (3.16) and for filtering (3.15) as

$$\begin{aligned} \dot{\widetilde{\boldsymbol{P}}}_n(t) &= \widetilde{\boldsymbol{A}}_n \boldsymbol{W}_n^{-1} \widetilde{\boldsymbol{P}}_n(t) + \widetilde{\boldsymbol{P}}_n(t) \boldsymbol{W}_n^{-1} \widetilde{\boldsymbol{A}}_n^{\mathrm{T}} + \widetilde{\boldsymbol{\Lambda}}_n b - \widetilde{\boldsymbol{P}}_n(t) \boldsymbol{W}_n^{-1} \widetilde{\boldsymbol{\Sigma}}_n \boldsymbol{W}_n^{-1} \widetilde{\boldsymbol{P}}_n(t), \\ \dot{\widetilde{\boldsymbol{R}}}_n(t) &= \widetilde{\boldsymbol{\mathcal{A}}}_n \mathbb{W}_n^{-1} \widetilde{\boldsymbol{R}}_n(t) + \widetilde{\boldsymbol{R}}_n(t) \mathbb{W}_n^{-1} \widetilde{\boldsymbol{\mathcal{A}}}_n^{\mathrm{T}} + \widetilde{\boldsymbol{L}}_n - \widetilde{\boldsymbol{R}}_n(t) \mathbb{W}_n^{-1} \widetilde{\boldsymbol{S}}_n \mathbb{W}^{-1} \widetilde{\boldsymbol{R}}_n(t) \end{aligned} \tag{5.9}$$

and the approximated filter equation (3.46) and control equation (3.47) in the form

$$\begin{aligned} \dot{\hat{\boldsymbol{x}}}_{n,C}(t) &= \boldsymbol{W}_n^{-1}([\widetilde{\boldsymbol{A}}_n - \widetilde{\boldsymbol{P}}_n(t)\widetilde{\boldsymbol{\Sigma}}_n]\hat{\boldsymbol{x}}_{n,C}(t) + \widetilde{\boldsymbol{P}}_n(t)\boldsymbol{W}_n^{-1}\bar{\Gamma}_n \boldsymbol{y}(t) - \widetilde{\boldsymbol{T}}_n(t_f - t)\hat{\boldsymbol{x}}_{n,C}(t)), \\ \boldsymbol{u}_n(t) &= -\begin{bmatrix} 0 & \boldsymbol{B}_0^T \end{bmatrix} \mathbb{W}_n^{-1} \widetilde{\boldsymbol{R}}_n(t_f - t) \mathbb{W}_n^{-1} \widetilde{\boldsymbol{T}}_{n,2} \hat{\boldsymbol{x}}_{n,C}(t), \\ \hat{\boldsymbol{z}}_n(t) &= [\, \boldsymbol{I}_{N \times N} \quad 0_{N \times nN} \,]\, \hat{\boldsymbol{x}}_{n,C}(t). \end{aligned} \tag{5.10}$$

In (5.9) and (5.10) $\hat{\boldsymbol{x}}_{n,C}$ is the coordinate expression of vector $\hat{\boldsymbol{x}}_n$ in the basis of Φ_n, $\hat{\boldsymbol{z}}_n(t)$ is the approximation of the optimal estimate of $\boldsymbol{z}(t)$, $\widetilde{\boldsymbol{P}}_n(t)$ and $\widetilde{\boldsymbol{R}}_n(t)$ are square matrices whose components are defined as $\{\widetilde{\boldsymbol{P}}_n(t)\}_{i,j} = (\boldsymbol{P}_n(t)\boldsymbol{v}_j, \boldsymbol{v}_i)$ and $\{\widetilde{\boldsymbol{R}}_n(t)\}_{i,j} = (\boldsymbol{R}_n(t)\boldsymbol{w}_j, \boldsymbol{w}_i)$. Matrices $\widetilde{\boldsymbol{\Lambda}}_n$, $\widetilde{\boldsymbol{\Sigma}}_n$, $\bar{\boldsymbol{\Gamma}}_n$, $\boldsymbol{W}_n$, $\widetilde{\boldsymbol{A}}_n$, $\widetilde{\boldsymbol{T}}_{n,1}$, $\mathbb{W}_n$, $\widetilde{\boldsymbol{T}}_{n,2}(t)$,

$\widetilde{\mathcal{A}}_n$, $\widetilde{\boldsymbol{L}}_n$, $\widetilde{\boldsymbol{S}}_n$ are numerical matrices computed by simple scalar products of elements in finite dimensional subspaces as follows:

$$
\begin{aligned}
\widetilde{\boldsymbol{\Lambda}}_n(i,j) &= (\boldsymbol{F}\boldsymbol{F}^*\boldsymbol{v}_j, \boldsymbol{v}_i)\,,\\
\widetilde{\boldsymbol{\Sigma}}_n(i,j) &= (\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{v}_j, \boldsymbol{v}_i)\,,\\
\bar{\boldsymbol{\Gamma}}_n(i,j) &= (\boldsymbol{C}^*\phi_j, \boldsymbol{v}_i)\,,\\
\boldsymbol{W}_n(i,j) &= (\boldsymbol{v}_i, \boldsymbol{v}_j),\\
\widetilde{\boldsymbol{A}}_n(i,j) &= (\boldsymbol{A}\boldsymbol{v}_j, \boldsymbol{v}_i),\\
\widetilde{\boldsymbol{T}}_{n,1}(i,j) &= (\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{w}_j, \boldsymbol{v}_i)\,,\\
\mathbb{W}_n(i,j) &= (\boldsymbol{w}_i, \boldsymbol{w}_j),\\
\widetilde{\boldsymbol{T}}_{n,2}(i,j) &= (\boldsymbol{v}_j, \boldsymbol{w}_i),\\
\widetilde{\mathcal{A}}_n(i,j) &= (\boldsymbol{A}^*\boldsymbol{w}_j, \boldsymbol{w}_i)\,,\\
\widetilde{\boldsymbol{L}}_n(i,j) &= (\boldsymbol{Q}\boldsymbol{w}_j, \boldsymbol{w}_i),\\
\widetilde{\boldsymbol{S}}_n(i,j) &= (\boldsymbol{B}\boldsymbol{B}^*\boldsymbol{w}_j, \boldsymbol{w}_i).
\end{aligned}
\tag{5.11}
$$

Finally, it is $\widetilde{\boldsymbol{T}}_n(t) = \widetilde{\boldsymbol{T}}_{n,1}\mathbb{W}_n^{-1}\widetilde{\boldsymbol{R}}_n(t)\mathbb{W}_n^{-1}\widetilde{\boldsymbol{T}}_{n,2}$.

Thus, denoting by

$$
\begin{aligned}
S_c(n,t) &= \boldsymbol{W}_n^{-1}(\widetilde{\boldsymbol{A}}_n - \widetilde{\boldsymbol{P}}_n(t)\boldsymbol{W}_n^{-1}\widetilde{\boldsymbol{\Sigma}}_n - \widetilde{\boldsymbol{T}}_{n,1}\mathbb{W}_n^{-1}\widetilde{\boldsymbol{R}}_n(t_f - t)\mathbb{W}_n^{-1}\widetilde{\boldsymbol{T}}_{n,2}),\\
P_c(n,t) &= \boldsymbol{W}_n^{-1}\widetilde{\boldsymbol{P}}_n(t)\boldsymbol{W}_n^{-1}\bar{\boldsymbol{\Gamma}}_n,\\
Q_c(n,t) &= -\,[\,0 \quad \boldsymbol{B}_0^{\mathrm{T}}\,]\,\mathbb{W}_n^{-1}\widetilde{\boldsymbol{R}}_n(t_f - t)\mathbb{W}_n^{-1}\widetilde{\boldsymbol{T}}_{n,2}
\end{aligned}
\tag{5.12}
$$

with $\widetilde{\boldsymbol{P}}_n(t)$ and $\widetilde{\boldsymbol{R}}_n(t)$ solutions of the matrix differential equations (Riccati) in (5.9), the approximate LQG controller can be written as follows:

$$
\begin{aligned}
\dot{\hat{\boldsymbol{x}}}_{n,C}(t) &= S_c(n,t)\hat{\boldsymbol{x}}_{n,C}(t) + P_c(n,t)\boldsymbol{y}(t),\\
\boldsymbol{u}_n(t) &= Q_c(n,t)\hat{\boldsymbol{x}}_{n,C}(t).
\end{aligned}
\tag{5.13}
$$

The vector $\hat{\boldsymbol{x}}_{n,C}(t) \in \mathbb{R}^{(n_1+1+\sum_{i=2}^{\delta} n_i)N}$.

Remark 5.1. It is important to stress the fact that matrices in (5.11) have a fixed structure and, in the case of hereditary systems without distributed delay, such matrices depend only on the multiindex s_n and on the matrices $\boldsymbol{A}_j$ $(j = 0, 1, \dots, \delta)$, $\boldsymbol{B}_0$, $\boldsymbol{C}_0$, $\boldsymbol{F}_0$, $\boldsymbol{G}$ that describe the system and on the weight matrix Q_0 that defines the cost functional. This property follows from the fact that splines are not uniformly distributed over the interval $[-r, 0]$: each interval $[-r_i, -r_{i-1}]$ has an independent spline distribution.

The numerical computation of matrices (5.11) is a straightforward function of the multi-index s_n and of the system matrices. As an example, the expressions of matrices in (5.11) are reported for systems with two pure delay terms (multi-index $s_n = (n_1, n_2)$ with $n = \inf(n_1, n_2)$) and no distributed delay.

$$
\boldsymbol{W}_n = \begin{bmatrix} \boldsymbol{W}_{n,a} & \boldsymbol{W}_{n,b} \\ 0_{n_2N\times n_1N} & \boldsymbol{W}_{n,c} \end{bmatrix},
\tag{5.14}
$$

$$\boldsymbol{W}_{n,a} = \left(\frac{r_1}{n_1}\right) \begin{bmatrix} \frac{n_1}{r_1} + 2/3 & 1/3 & 0 & \dots & 0 \\ 1/3 & 4/3 & 1/3 & \ddots & \vdots \\ 0 & 1/3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 4/3 & 1/3 \\ 0 & \cdot & 0 & 1/3 & 2/3 + \frac{n_1}{r_1}\frac{r-r_1}{3n_2} \end{bmatrix}_{(n_1+1)\times(n_1+1)} \otimes \boldsymbol{I}_{N\times N},$$

$$\boldsymbol{W}_{n,b} = \begin{bmatrix} 0 & \dots & \dots & 0 \\ \frac{r-r_1}{6n_2}\boldsymbol{I}_{N\times N} & 0 & \dots & 0 \end{bmatrix}_{(n_1+1)N\times n_2 N},$$

$$\boldsymbol{W}_{n,c} = \left(\frac{r-r_1}{n_2}\right) \begin{bmatrix} 1/6 & 2/3 & 1/6 & 0 & \dots \\ 0 & 1/6 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 2/3 & 1/6 \\ 0 & \cdot & 0 & 1/6 & 1/3 \end{bmatrix}_{n_2\times(n_2+1)} \otimes \boldsymbol{I}_{N\times N}.$$

$$\widetilde{\boldsymbol{A}}_n = \widetilde{\boldsymbol{A}}_{n,1} + \widetilde{\boldsymbol{A}}_{n,2}, \tag{5.15}$$

$$\widetilde{\boldsymbol{A}}_{n,1} = \begin{bmatrix} \boldsymbol{A}_0 & 0 & \dots & 0 & \boldsymbol{A}_1 & 0 & \dots & 0 & \boldsymbol{A}_2 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \end{bmatrix},$$

$$\widetilde{\boldsymbol{A}}_{n,2} = \begin{bmatrix} \widetilde{\boldsymbol{A}}_{n,2,a} & \widetilde{\boldsymbol{A}}_{n,2,b} \\ 0_{n_2 N\times n_1 N} & \widetilde{\boldsymbol{A}}_{n,2,c} \end{bmatrix},$$

$$\widetilde{\boldsymbol{A}}_{n,2,a} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & -1 \\ 0 & \vdots & 0 & 1 & -1/2 \end{bmatrix}_{n_1+1\times(n_1+1)} \otimes \boldsymbol{I}_{N\times N},$$

$$\widetilde{\boldsymbol{A}}_{n,2,b} = \begin{bmatrix} 0_{n_1 N\times n_2 N} \\ [-1/2\boldsymbol{I}_{N\times N} \quad 0 \quad \dots \quad 0] \end{bmatrix},$$

$$\widetilde{\boldsymbol{A}}_{n,2,c} = \begin{bmatrix} 1/2 & 0 & -1/2 & 0 & \dots \\ 0 & 1/2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & -1/2 \\ 0 & \vdots & 0 & 1/2 & -1/2 \end{bmatrix}_{n_2\times(n_2+1)} \otimes \boldsymbol{I}_{N\times N}.$$

$$\mathbb{W}_n = \begin{bmatrix} \mathbb{W}_{n,a} & \mathbb{W}_{n,b} \\ \mathbb{W}_{n,b}^{\mathrm{T}} & \mathbb{W}_{n,c} \end{bmatrix} \tag{5.16}$$

with

$$\mathbb{W}_{n,a} = \frac{r_1}{n_1}\begin{bmatrix} 2/3 & 1/3 & 0 & \dots & 0 \\ 1/3 & 4/3 & 1/3 & \ddots & \vdots \\ 0 & 1/3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 4/3 & 1/3 \\ 0 & \cdot & 0 & 1/3 & 4/3 \end{bmatrix}_{n_1\times n_1} \otimes \boldsymbol{I}_{N\times N},$$

$$\mathbb{W}_{n,b} = \begin{bmatrix} \boldsymbol{0}_{(n_1-1)N\times(n_2+1)N} \\ [\,\frac{r_1}{6n_1}\boldsymbol{I}_{N\times N} \quad 0 \quad \dots \quad 0 \quad \frac{r_1}{6n_1}\boldsymbol{A}_1^{\mathrm{T}}\,] \end{bmatrix}_{n_1N\times(n2+1)N},$$

$$\mathbb{W}_{n,c} = \mathbb{W}_{n,c,1} + \mathbb{W}_{n,c,2},$$

$$\mathbb{W}_{n,c,1} = \left(\frac{r-r_1}{n_2}\right)\begin{bmatrix} 1/3 & 1/6 & 0 & \dots & 0 \\ 1/6 & 2/3 & 1/6 & \ddots & \vdots \\ 0 & 1/6 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 2/3 & 1/6\boldsymbol{A}_2^{\mathrm{T}} \\ 0 & \cdot & 0 & 1/6\boldsymbol{A}_2 & 0 \end{bmatrix}_{(n_2+1)\times(n_2+1)} \otimes \boldsymbol{I}_{N\times N},$$

$$\mathbb{W}_{n,c,2} = \begin{bmatrix} \frac{r_1}{6n_1}\boldsymbol{I}_{N\times N} & 0 & \dots & 0 & \frac{r_1}{6n_1}\boldsymbol{A}_1^{\mathrm{T}} \\ 0 & \dots & \dots & \dots & 0 \\ \frac{r_1}{6n_1}\boldsymbol{A}_1 & 0 & \dots & 0 & \boldsymbol{I} + \frac{r_1}{6n_1}\boldsymbol{A}_1\boldsymbol{A}_1^{\mathrm{T}} + \frac{r-r_1}{3n_2}\boldsymbol{A}_2\boldsymbol{A}_2^{\mathrm{T}} \end{bmatrix}.$$

$$\widetilde{\boldsymbol{T}}_{n,2} = \begin{bmatrix} \widetilde{\boldsymbol{T}}_{n,2,a} & \widetilde{\boldsymbol{T}}_{n,2,b} \\ \widetilde{\boldsymbol{T}}_{n,2,c} & \widetilde{\boldsymbol{T}}_{n,2,d} \\ \widetilde{\boldsymbol{T}}_{n,2,e} & \widetilde{\boldsymbol{T}}_{n,2,f} \\ \widetilde{\boldsymbol{T}}_{n,2,g} & \widetilde{\boldsymbol{T}}_{n,2,h} \end{bmatrix}_{(n_1+n_2+1)N\times(n_1+n_2+1)N}, \tag{5.17}$$

$$\widetilde{\boldsymbol{T}}_{n,2,a} = \left(\frac{r_1}{n_1}\right)\begin{bmatrix} \begin{bmatrix} 2/3 & 1/3 & 0 & \dots & 0 \\ 1/3 & 4/3 & 1/3 & \ddots & \vdots \\ 0 & 1/3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 4/3 & 1/3 \\ 0 & \cdot & 0 & 1/3 & 4/3 \end{bmatrix}_{n_1\times n_1} \otimes \boldsymbol{I}_{N\times N} & \begin{bmatrix} \boldsymbol{0} \\ \frac{1}{3}\boldsymbol{I}_{N\times N} \end{bmatrix} \end{bmatrix}_{n_1N\times(n_1+1)N},$$

$$\widetilde{\boldsymbol{T}}_{n,2,b} = 0_{n_1N\times n_2N},$$

$$\widetilde{\boldsymbol{T}}_{n,2,c} = [\,0 \quad \dots \quad 0 \quad \tfrac{r_1}{6n_1}\boldsymbol{I}_{N\times N} \quad \tfrac{r-r_1}{3n_2}\boldsymbol{I}_{N\times N} + \tfrac{r_1}{3n_1}\boldsymbol{I}_{N\times N}\,]_{N\times(n_1+1)N},$$

$$\widetilde{\boldsymbol{T}}_{n,2,d} = [\,\tfrac{r-r_1}{6n_2}\boldsymbol{I}_{N\times N}0 \quad \dots \quad 0]_{N\times n_2N},$$

$$\widetilde{\boldsymbol{T}}_{n,2,e} = \begin{bmatrix} 0 & \dots & 0 & \frac{r-r_1}{6n_2}\boldsymbol{I}_{N\times N} \\ 0 & \dots & \dots & 0 \end{bmatrix}_{(n_2-1)N\times(n_1+1)N},$$

$$\widetilde{\boldsymbol{T}}_{n,2,f} = \left(\frac{r-r_1}{n_2}\right)\left[\begin{bmatrix} 2/3 & 1/6 & 0 & \dots \\ 1/6 & \ddots & \ddots & 0 \\ 0 & \ddots & 2/3 & 1/6 \\ \dots & 0 & 1/6 & 2/3 \end{bmatrix}_{(n_2-1)\times n_2} \otimes \boldsymbol{I}_{N\times N} \quad \begin{bmatrix} \boldsymbol{0} \\ \frac{1}{6}\boldsymbol{I}_{N\times N} \end{bmatrix}\right]_{(n_2-1)N\times(n_2-1)N},$$

$$\widetilde{\boldsymbol{T}}_{n,2,g} = [\,\boldsymbol{I}_{N\times N} \quad \boldsymbol{0} \quad \tfrac{r_1}{6n_1}\boldsymbol{A}_1 \quad \tfrac{r_1}{3n_1}\boldsymbol{A}_1\,]_{N\times(n_1+1)N},$$

$$\widetilde{\boldsymbol{T}}_{n,2,h} = [\,\boldsymbol{0} \quad \dots \quad 0 \quad \tfrac{r-r_1}{6n_2}\boldsymbol{A}_2 \quad \tfrac{r-r_1}{3n_2}\boldsymbol{A}_2\,]_{N\times n_2 N}.$$

$$\widetilde{\mathcal{A}}_n = \widetilde{\mathcal{A}}_{n,1} + \widetilde{\mathcal{A}}_{n,2}, \tag{5.18}$$

$$\widetilde{\mathcal{A}}_{n,1} = \begin{bmatrix} 0 & \dots & \dots & \dots & 0 \\ 2\boldsymbol{I}_{N\times N} & 0 & \dots & 0 & \boldsymbol{A}_0^{\mathrm{T}} \end{bmatrix}_{(n_1+n_2+1)N\times(n_1+n_2+1)N},$$

$$\widetilde{\mathcal{A}}_{n,2} = \begin{bmatrix} \widetilde{\mathcal{A}}_{n,2,a} & \widetilde{\mathcal{A}}_{n,2,b} \\ \widetilde{\mathcal{A}}_{n,2,c} & \widetilde{\mathcal{A}}_{n,2,d} \end{bmatrix}_{(n_1+n_2+1)N\times(n_1+n_2+1)N},$$

$$\widetilde{\mathcal{A}}_{n,2,a} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \ddots & \vdots \\ 0 & -1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \vdots & 0 & -1 & 0 \end{bmatrix}_{n_1\times n_1} \otimes \boldsymbol{I}_{N\times N},$$

$$\widetilde{\mathcal{A}}_{n,2,b} = \begin{bmatrix} 0 & \dots & \dots & \dots & 0 \\ 1/2\boldsymbol{I}_{N\times N} & 0 & \dots & 0 & 1/2\boldsymbol{A}_1^{\mathrm{T}} \end{bmatrix}_{n_1 N\times(n_2+1)N},$$

$$\widetilde{\mathcal{A}}_{n,2,c} = \begin{bmatrix} 0 & \dots & 0 & -1/2\boldsymbol{I}_{N\times N} \\ \vdots & \vdots & \vdots & 0 \\ 0 & \dots & 0 & -1/2\boldsymbol{A}_1 \end{bmatrix}_{(n_2+1)N\times n_1 N},$$

$$\widetilde{\mathcal{A}}_{n,2,d} = \left[\begin{array}{c|c} \begin{bmatrix} -1/4 & 1/2 & 0 & \dots & 0 \\ -1/2 & 0 & 1/2 & \ddots & \vdots \\ 0 & -1/2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1/2 \\ 0 & \vdots & 0 & -1/2 & 0 \end{bmatrix}_{n_2\times n_2} \otimes \boldsymbol{I}_{N\times N} & \begin{matrix} 1/4\boldsymbol{A}_1^{\mathrm{T}} \\ 0 \\ \vdots \\ 0 \\ \frac{1}{2}\boldsymbol{A}_2^{\mathrm{T}} \end{matrix} \\ \hline 1/4\boldsymbol{A}_1 \quad 0 \quad \dots \quad 0 \quad -\frac{1}{2}\boldsymbol{A}_2 & \frac{1}{4}\boldsymbol{A}_1\boldsymbol{A}_1^{\mathrm{T}} + 1/2\boldsymbol{A}_2\boldsymbol{A}_2^{\mathrm{T}} \end{array}\right].$$

Remark 5.2. In the case of just one pure delay, matrices in (5.11) are much simpler, due to the fact that vectors v and w are much simpler. Matrices which involve v vectors have been computed in [3]. Here, just to have an idea of such a simplification, matrix $\widetilde{\mathcal{A}}_n$ in (5.18) is reported in the case of one pure delay term.

$$\widetilde{\mathcal{A}}_n = \widetilde{\mathcal{A}}_{n,1} + \widetilde{\mathcal{A}}_{n,2}, \tag{5.19}$$

$$\widetilde{\mathcal{A}}_{n,1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{I}_{N\times N} & \mathbf{0} & \boldsymbol{A}_0^{\mathrm{T}} \end{bmatrix},$$

$$\widetilde{\mathcal{A}}_{n,2} = \left[\begin{array}{c|c} \begin{bmatrix} -1/2 & 1/2 & 0 & \dots & 0 \\ -1/2 & 0 & 1/2 & \ddots & \vdots \\ 0 & -1/2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1/2 \\ 0 & \vdots & 0 & -1/2 & 0 \end{bmatrix}_{n\times n} \otimes \boldsymbol{I}_{N\times N} & \begin{array}{c} \mathbf{0} \\ \frac{1}{2}\boldsymbol{A}_1^{\mathrm{T}} \end{array} \\ \hline \mathbf{0} \quad -\frac{1}{2}\boldsymbol{A}_1 & \frac{1}{2}\boldsymbol{A}_1\boldsymbol{A}_1^{\mathrm{T}} \end{array}\right].$$

If there is the distributed delay too, then the following matrix must be added in the right-hand side of (5.19):

$$\widetilde{\mathcal{A}}_{n,3} = \begin{bmatrix} \mathbf{0} & (\boldsymbol{D}_0'^{n})^{\mathrm{T}} \\ \vdots & \vdots \\ \mathbf{0} & (\boldsymbol{D}_{n-1}'^{n})^{\mathrm{T}} \\ \mathbf{0} & \boldsymbol{A}_1(\boldsymbol{D}_n'^{n})^{\mathrm{T}} \end{bmatrix},$$

where

$$\boldsymbol{D}_j'^{n} = \int_{-r}^{0} \boldsymbol{A}_{01}(s)\, spline_j(s) ds \qquad j = 0, 1, \dots, n.$$

6. Remarks on the infinite horizon case. The methodology here presented for LQG control of hereditary systems over a finite time-horizon can be applied also for LQG control over infinite time-horizon. The basis is the paper [14] in which, under suitable conditions, the convergence of the solution of an approximate Riccati differential equation, evaluated in a sufficiently large time, to the solution of the corresponding infinite dimensional algebraic Riccati equation is proved. The hypotheses required in [14] for such a convergence are satisfied by hereditary systems and by the approximation scheme here presented. Such hypotheses are the Hilbert–Schmidt property of operators $\boldsymbol{Q}$ and $\boldsymbol{F}\boldsymbol{F}^*$, the convergence of the sequence of projection operators involved, and the convergence of the semigroups approximating the semigroup generated by the operator which, in the algebraic Riccati equation, multiplies on the left the unknown Riccati operator. Structural properties are requested in paper [14] of approximate controllability of pairs $(\boldsymbol{A}, \boldsymbol{F})$ and $(\boldsymbol{A}^*, \boldsymbol{Q})$. In that paper the approximate solution of an algebraic Riccati equation is found by exploiting the approximability of the corresponding dynamical Riccati equation and its time convergence toward the steady state, and finding a large enough time-horizon T to approximate the steady-state solution. Such a solving method, which requires only convergence of one approximating semigroup, does not allow for a uniform convergence of the approximate solution toward the actual one (see Theorem 3.2 of [14], relationship among ϵ, T, and n). On the other hand, such a solving method does not require the uniform exponential stability of the approximating semigroups nor the convergence of the adjoint approximate semigroups.

Using the approximate solutions of the Riccati algebraic equations, by using the above paper, the infinite horizon LQG controller can be built. The problem of guaranteeing the convergence of the approximation schemes in infinite horizon case continues to be worthy of attention.

Nevertheless, when the state is fully available, the approximation scheme here presented has the nice property to guarantee convergence also in the infinite horizon case, as stated in the following theorem.

Theorem 6.1. *Consider system* (2.5)*, with fully available state, that is,*

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) \tag{6.1}$$

and the following cost functional

$$J_I(\boldsymbol{u}) = \lim_{t_f \to \infty} \frac{1}{t_f} \int_0^{t_f} E[(\boldsymbol{Q}\boldsymbol{x}(t), \boldsymbol{x}(t)) + \boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{u}(t)]dt, \tag{6.2}$$

with $\boldsymbol{Q} : \boldsymbol{M}_2 \mapsto \boldsymbol{M}_2$ *as in* (3.2)*. Let the pair* $(\boldsymbol{A}, \boldsymbol{B})$ *be stabilizable and the pair* $(\boldsymbol{A}, \boldsymbol{Q})$ *be detectable. Let*

$$\boldsymbol{u}_n(t) = -\boldsymbol{B}^{\star}\boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n\boldsymbol{x}_n(t), \tag{6.3}$$

where $\boldsymbol{R}_n(T)$ *is the approximate solution of the algebraic Riccati equation for control*

$$\boldsymbol{A}^{\star}\boldsymbol{R} + \boldsymbol{R}\boldsymbol{A} - \boldsymbol{R}\boldsymbol{B}\boldsymbol{B}^{\star}\boldsymbol{R} + \boldsymbol{Q} = 0 \tag{6.4}$$

obtained [14] *by evaluating the approximate dynamic Riccati equation* (3.16) *in a suitable time* T*, and* $\boldsymbol{x}_n(t)$ *is the corresponding evolving state. Let* $\boldsymbol{x}(t)$ *be the state evolving when the optimal infinite horizon LQG control law is applied to the system. Then, for every* $\epsilon > 0$*, there exists a* T_ϵ*, such that for every* $T > T_\epsilon$ *there exists an* n_T*, such that for every* $n > n_T$ *the semigroup which governs the closed loop system, that is, the one generated by* $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{B}^{\star}\boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n$*, is exponentially stable and, moreover,*

$$E\|\boldsymbol{x}_n(t) - \boldsymbol{x}(t)\| < \epsilon \qquad \forall t \in [0, \infty). \tag{6.5}$$

Proof. First let us prove that $E\|x(t)\|$ is uniformly bounded. Let $S(t)$ be the semigroup generated by the optimal closed loop infinitesimal generator $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{B}^{\star}\boldsymbol{R}$, with $\boldsymbol{R}$ the solution of the algebraic Riccati control equation. There exist positive constants M and σ such that $\|S(t)\| \le Me^{-\sigma t}$. It is

$$\begin{aligned} E\|\boldsymbol{x}(t)\| &\le E\|S(t)\boldsymbol{x}(0)\| + E\|\int_0^t S(t-\tau)\boldsymbol{F}\boldsymbol{\omega}(\tau)d\tau\| \\ &\le Me^{-\sigma t}\sqrt{E(\|\boldsymbol{x}(0)\|)^2} + \left(\int_0^t M^2 e^{-2\sigma(t-\tau)}\|\boldsymbol{F}\|^2_{H.S.}\right)^{\frac{1}{2}} \\ &\le M\sqrt{\mathrm{Tr}(\boldsymbol{P}_0)} + \frac{M}{\sqrt{2\sigma}}\|\boldsymbol{F}\|_{H.S.} \end{aligned} \tag{6.6}$$

Consider now the equation

$$\dot{\xi}_{n,T}(t) = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{B}^\star \boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n)\xi_{n,T}(t). \tag{6.7}$$

It is

$$\xi_{n,T}(t) = S(t)\xi_{n,T}(0) + \int_0^t S(t-\tau)\boldsymbol{B}\boldsymbol{B}^\star(\boldsymbol{R} - \boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n)\xi_{n,T}(\tau)d\tau \tag{6.8}$$

by which it follows that

$$\|\xi_{n,T}(t)\| \le Me^{-\sigma t}\|\xi_{n,T}(0)\| \ + \int_0^t Me^{-\sigma(t-\tau)}\|\boldsymbol{B}\boldsymbol{B}^\star(\boldsymbol{R} - \boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n)\|_{H.S.}\xi_{n,T}(\tau)d\tau \tag{6.9}$$

and by the Gronwall inequality

$$\|\xi_{n,T}(t)\| \le Me^{(-\sigma + M\|\boldsymbol{B}\boldsymbol{B}^\star(\boldsymbol{R}-\boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n)\|_{H.S.}t)}\xi_{n,T}(0). \tag{6.10}$$

Now let $\epsilon > 0$. By Theorem 3.2 in [14] and by the inequality

$$\|\boldsymbol{R} - \boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n\|_{H.S.} \le \|\boldsymbol{\Pi}'_n\boldsymbol{R}\boldsymbol{\Pi}'_n - \boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n\|_{H.S.} + \|\boldsymbol{R} - \boldsymbol{\Pi}'_n\boldsymbol{R}\boldsymbol{\Pi}'_n\|_{H.S.}, \tag{6.11}$$

it follows that there exists T_ϵ such that for every $T > T_\epsilon$ there exists n_T such that for every $n > n_T$

$$\|\boldsymbol{R} - \boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n\|_{H.S.} < \min\left\{\frac{\sigma}{2M\|\boldsymbol{B}\boldsymbol{B}^\star\|}, \frac{\epsilon\sigma}{2M\|\boldsymbol{B}\boldsymbol{B}^\star\| \sup_{\tau\in[-r,\infty)} E\|\boldsymbol{x}(\tau)\|}\right\} \tag{6.12}$$

and so

$$\|\xi_{n,T}(t)\| \le Me^{\frac{-\sigma}{2}t}\xi_{n,T}(0), \tag{6.13}$$

which implies the exponential stability of the closed loop semigroup.

As far as the second part of the thesis is concerned, let

$$e_n(t) = \boldsymbol{x}(t) - \boldsymbol{x}_n(t).$$

It is

$$\dot{e}_n(t) = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{B}^\star\boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n)e_n(t) + \boldsymbol{B}\boldsymbol{B}^\star(\boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n - \boldsymbol{R})\boldsymbol{x}(t) \tag{6.14}$$

by which, taking into account (6.12),

$$\begin{aligned} E\|e_n(t)\| \le \int_0^t & Me^{(-\sigma+M\|\boldsymbol{B}\boldsymbol{B}^\star(\boldsymbol{R}-\boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n)\|_{H.S.})(t-\tau)} \\ & \cdot\|\boldsymbol{B}\boldsymbol{B}^\star(\boldsymbol{R} - \boldsymbol{R}_n(T)\boldsymbol{\Pi}'_n)\|_{H.S.}E\|\boldsymbol{x}(\tau)\|d\tau < \epsilon. \qquad \square \end{aligned} \tag{6.15}$$

7. Examples. Simulations reported in this section have been performed by MATLAB on a PC using the 3rd order Runge–Kutta integration algorithm.[1]

[1] Simulation programs are available upon request.

Example 1. Consider the following unstable hereditary system:

$$\begin{aligned}\frac{d^2z(t)}{dt^2} &= \frac{dz(t)}{dt} + z(t) + \frac{dz(t-r_1)}{dt} + z(t-r_1)\\ &\quad + \frac{dz(t-r_2)}{dt} + z(t-r_2) + u(t) + \omega_1(t),\\ y(t) &= z(t) + \omega_2(t),\end{aligned} \tag{7.1}$$

where $z(t), u(t), y(t) \in \mathbb{R}$, $\omega_1(t), \omega_2(t) \in \mathbb{R}$ are independent white Gaussian standard noises.

By denoting $Z(t) = \left[\begin{smallmatrix} z(t) \\ \frac{dz(t)}{dt}\end{smallmatrix}\right]$, and $\omega(t) = \left[\begin{smallmatrix}\omega_1(t)\\ \omega_2(t)\end{smallmatrix}\right]$, the system (7.1) can be rewritten as follows:

$$\begin{aligned}\dot{Z}(t) &= \begin{bmatrix} 0 & 1\\ 1 & 1\end{bmatrix} Z(t) + \begin{bmatrix} 0 & 0\\ 1 & 1\end{bmatrix} Z(t-r_1)\\ &\quad + \begin{bmatrix} 0 & 0\\ 1 & 1\end{bmatrix} Z(t-r_2) + \begin{bmatrix} 0\\ 1\end{bmatrix} u(t) + \begin{bmatrix} 0 & 0\\ 1 & 0\end{bmatrix} \omega(t)\\ y(t) &= [1 \quad 0]\, Z(t) + [0 \quad 1]\, \omega(t).\end{aligned} \tag{7.2}$$

The weight matrix in the functional (3.1) has been chosen as

$$\boldsymbol{Q}_0 = \begin{bmatrix} 1000 & 0\\ 0 & 0\end{bmatrix}. \tag{7.3}$$

The time t_f has been chosen equal to 10, and the delays have been chosen as $r_1 = 1.2$ and $r_2 = 2.5$. The initial value of $Z(\vartheta), \vartheta \in [-r_2, 0]$, has been chosen as follows:

$$Z(\vartheta) = \begin{bmatrix} e^{-\vartheta}\\ \cos(10\vartheta)\end{bmatrix},$$

while the initial estimate $\hat{Z}(\vartheta)$ has been set to 0 in the same interval. The covariance operator $\boldsymbol{P}_0$ of the initial state in M_2 has been chosen as follows:

$$\boldsymbol{P}_0 \boldsymbol{x} = (\boldsymbol{x}, \phi)\phi,$$

where $\boldsymbol{x}, \phi \in M_2$, $\phi = \left[\begin{smallmatrix}\phi_0\\ \phi_1\end{smallmatrix}\right]$, $\phi_0 = \phi_1(\vartheta) = \left[\begin{smallmatrix}1\\ 1\end{smallmatrix}\right], \vartheta \in [-r_2, 0]$.

The integration step has been chosen equal to 0.0025.

Figures 2–5 report the first component of the actual $Z(t)$ evolving when the approximated input is applied to the system and of the estimated $\hat{Z}(t)$ for different values of n_1 and n_2. Figures 6–9 report the second component of $Z(t)$ and $\hat{Z}(t)$. In Figures 10–13 the first component and the second one of $Z(t)$ and $\hat{Z}(t)$, the approximated control input and the noisy output, are reported for $n_1 = n_2 = 6$.

Example 2. Consider now the well-known National Transonic Facility [4, 27, 40], the liquid nitrogen wind tunnel at NASA Langley Research Center in Hampton, VA. Here only one of the state variables is measured, the guide vane angle, while no measurement of the mach number nor of the guide vane angle derivative is available. Moreover, we suppose an additive Gaussian noise corrupts the dynamics of the system and the above measure. A simplified model of such a system is given by (see [4] for the

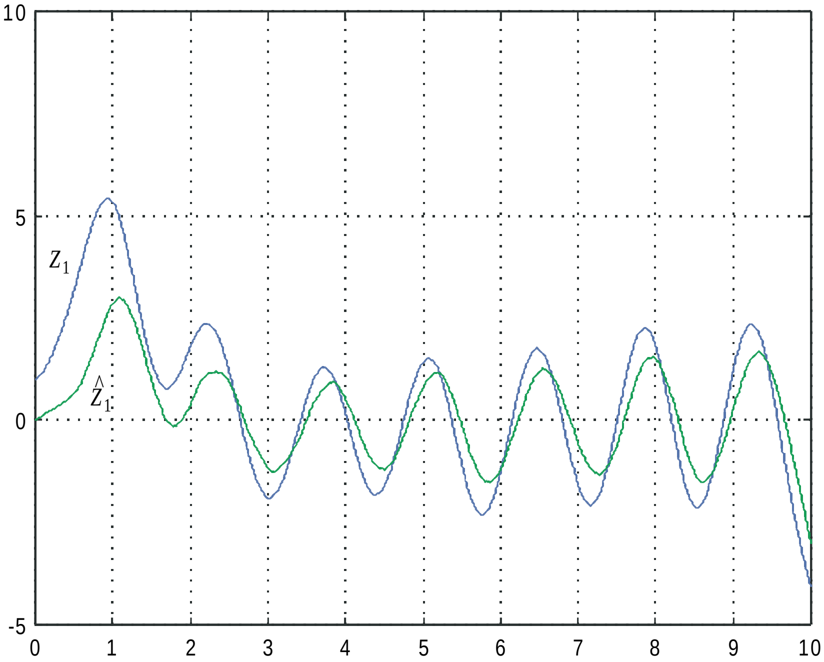


FIG. 2. *The case of* $n_1 = n_2 = 2$: *true and estimated* $Z_1(t)$.

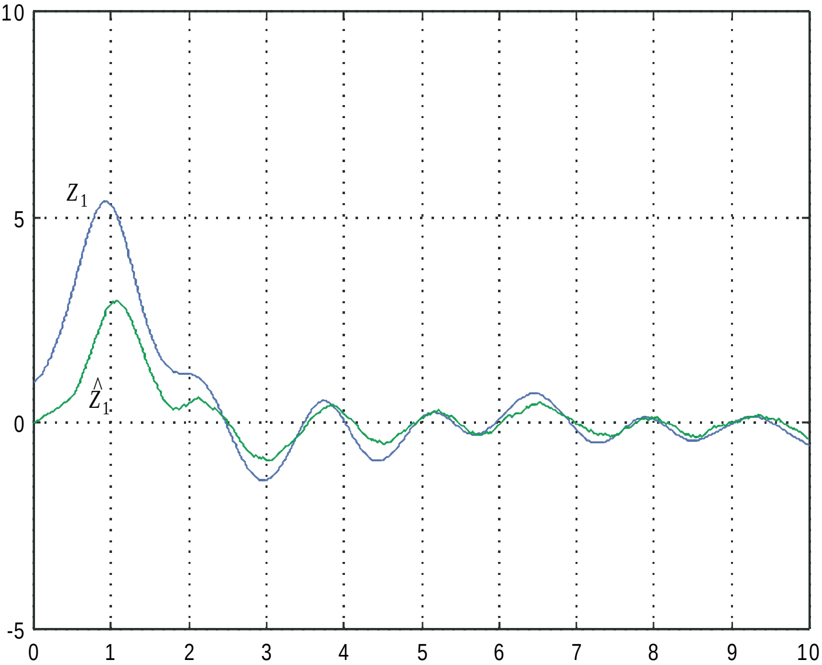


FIG. 3. *The case of* $n_1 = 3$, $n_2 = 2$: *true and estimated* $Z_1(t)$.

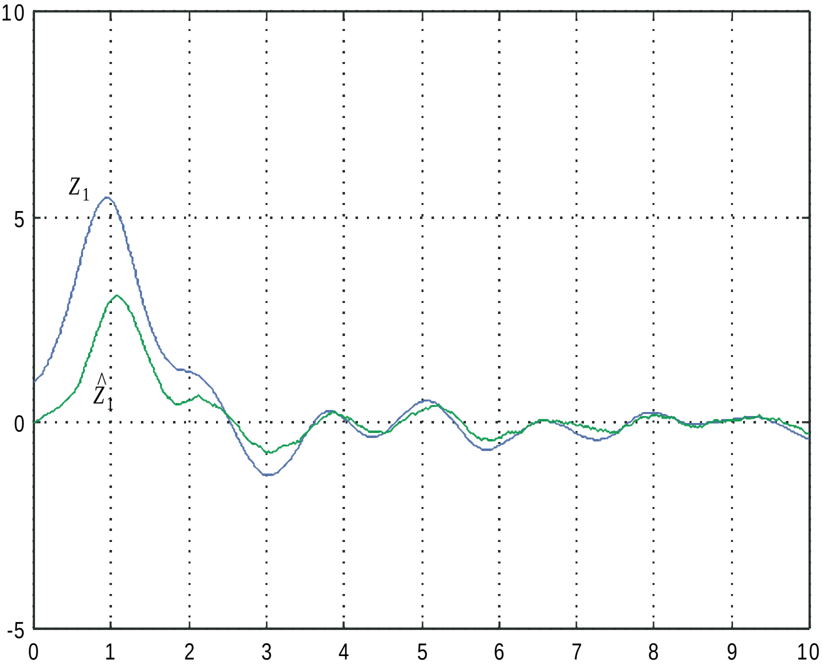


FIG. 4. *The case of* $n_1 = 3$, $n_2 = 3$: *true and estimated* $Z_1(t)$.

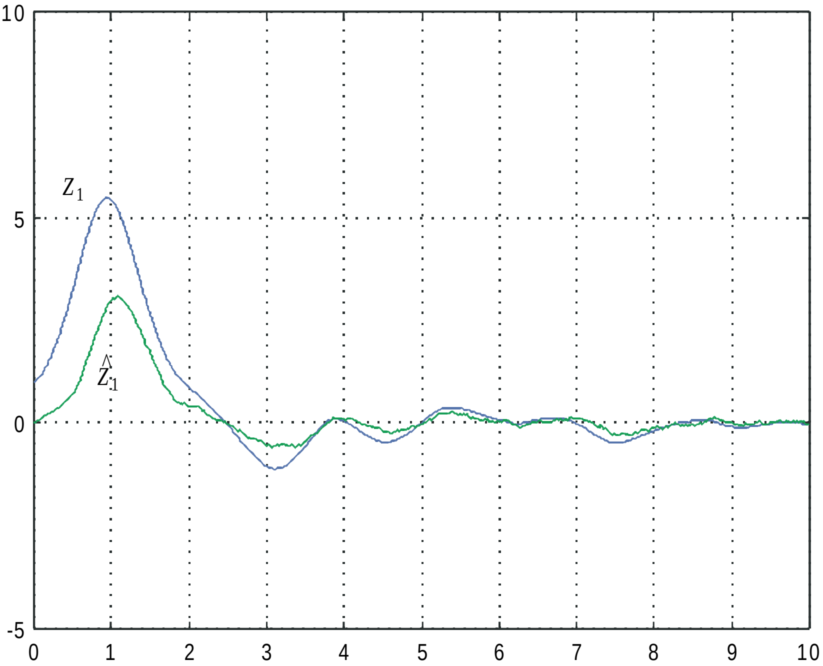


FIG. 5. *The case of* $n_1 = 4$, $n_2 = 3$: *true and estimated* $Z_1(t)$.

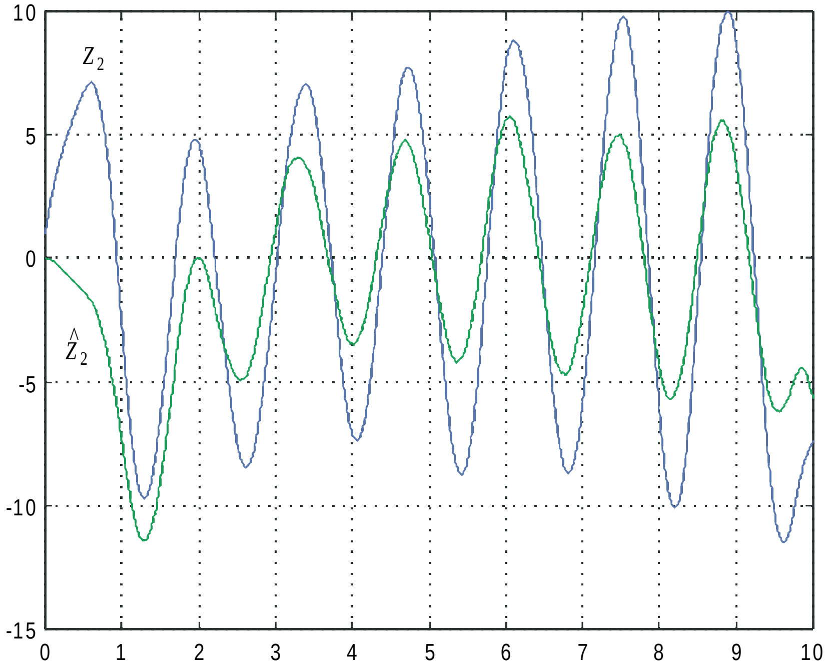


FIG. 6. *The case of* $n_1 = n_2 = 2$: *true and estimated* $Z_2(t)$.

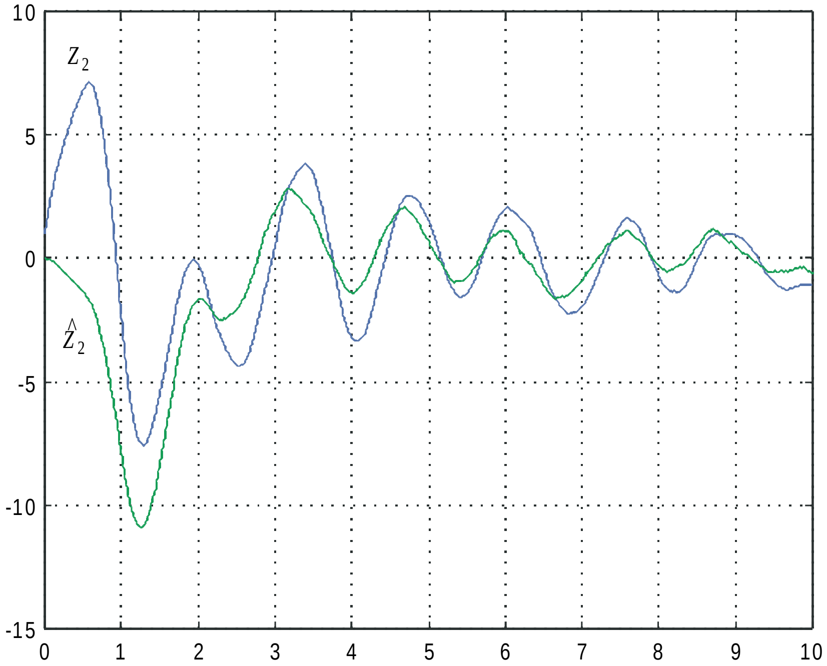


FIG. 7. *The case of* $n_1 = 3$, $n_2 = 2$: *true and estimated* $Z_2(t)$.

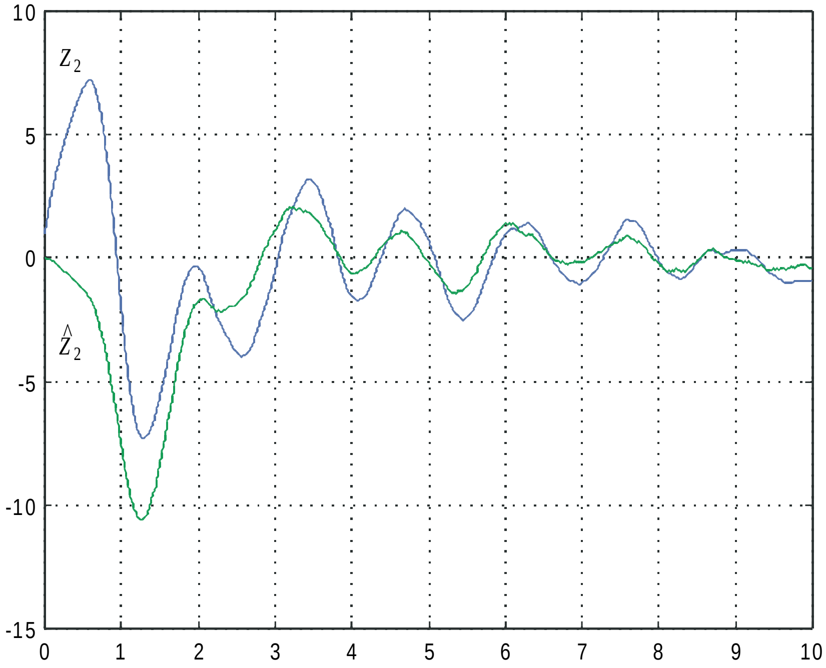


FIG. 8. *The case of* $n_1 = 3$, $n_2 = 3$*: true and estimated* $Z_2(t)$*.*

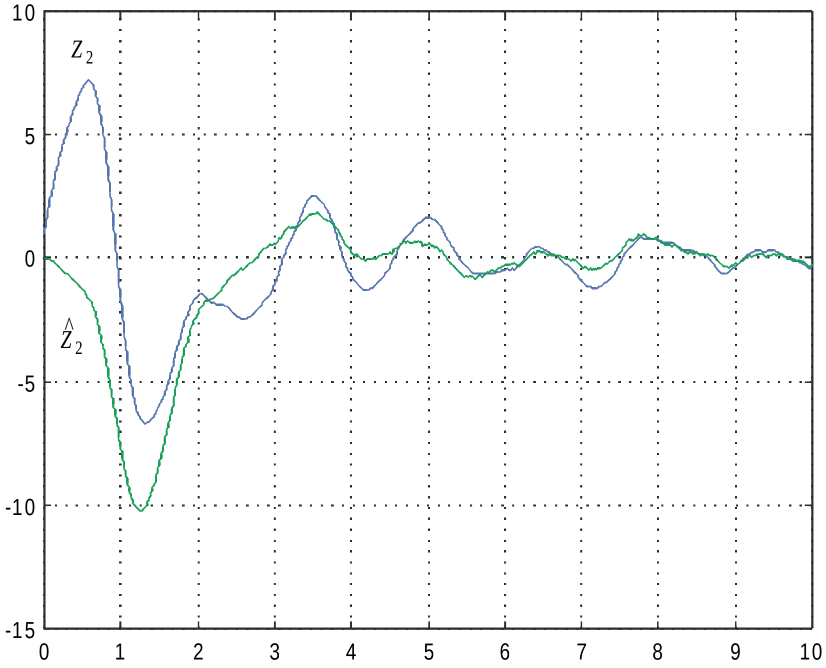


FIG. 9. *The case of* $n_1 = 4$, $n_2 = 3$*: true and estimated* $Z_2(t)$*.*

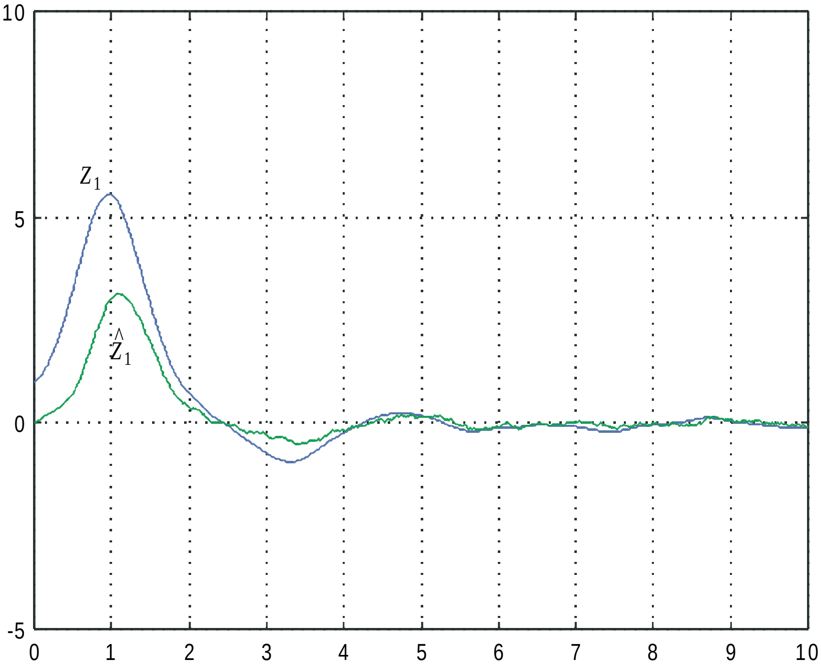


FIG. 10. *The case of* $n_1 = n_2 = 6$*: true and estimated* $Z_1(t)$*.*

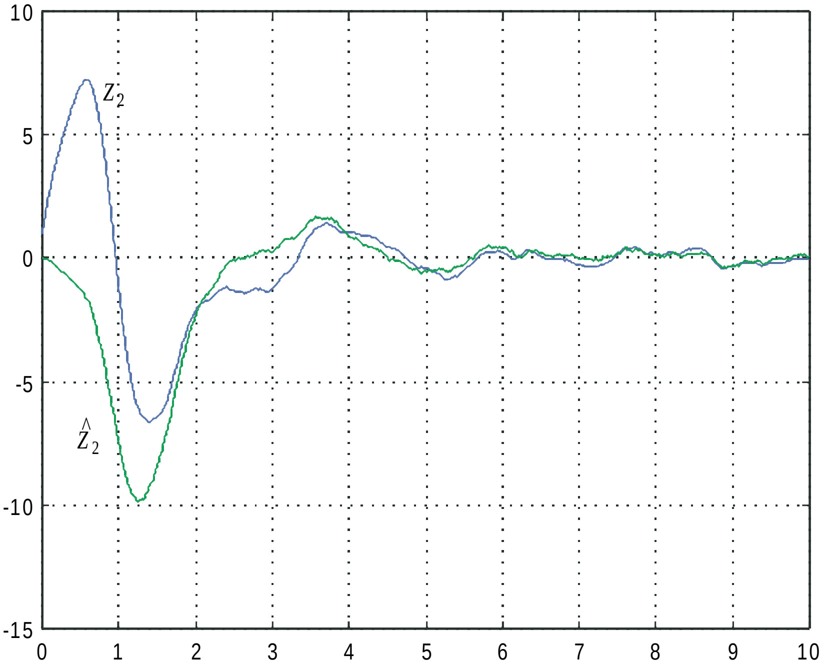


FIG. 11. *The case of* $n_1 = n_2 = 6$*: true and estimated* $Z_2(t)$*.*

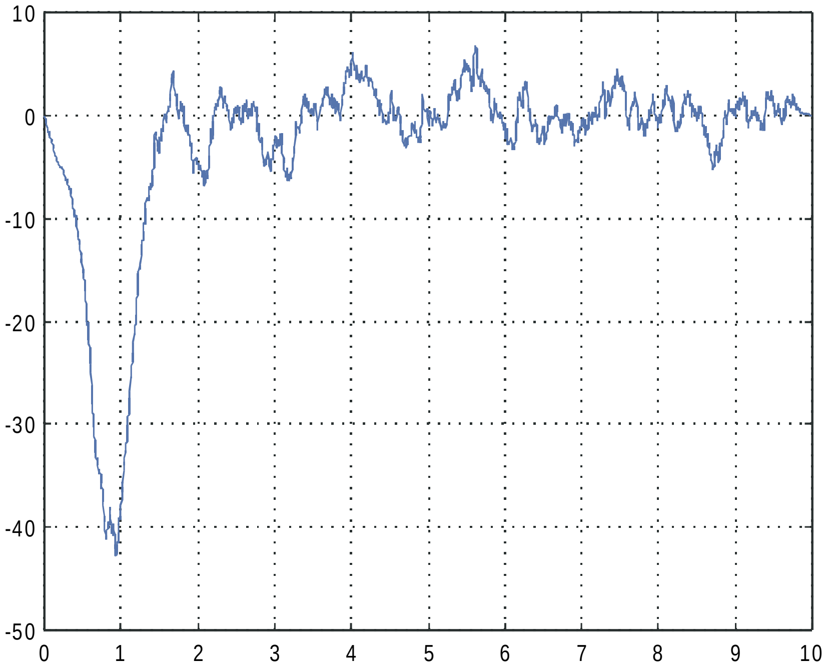


FIG. 12. *The case of $n_1 = n_2 = 6$: the input $u(t)$.*

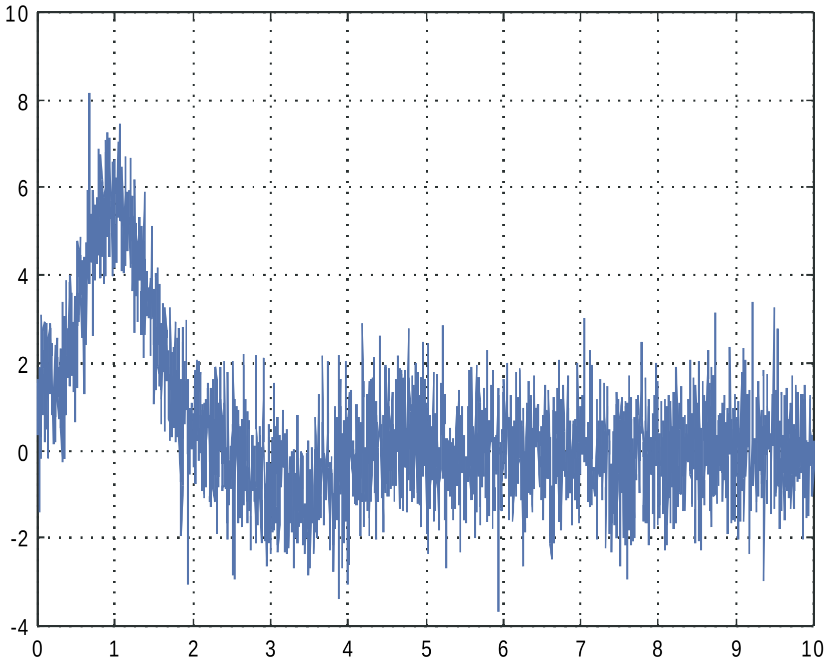


FIG. 13. *The case of $n_1 = n_2 = 6$:: the noisy output $y(t)$.*

deterministic model)

$$\begin{aligned}\dot{z}(t) &= \begin{bmatrix} -a & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -\bar{\omega}^2 & -2\xi\bar{\omega} \end{bmatrix} z(t) + \begin{bmatrix} 0 & ka & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} z(t-0.33) \\ &\quad + \begin{bmatrix} 0 \\ 0 \\ -\bar{\omega}^2 \end{bmatrix} \boldsymbol{u}(t) + F_0\omega_1(t) \\ \boldsymbol{y}(t) &= [\,0 \quad 1 \quad 0)\,]\, z(t) + G_0\omega_2(t)\end{aligned} \tag{7.4}$$

with $(1/a) = 1.964$, $k = -0.0117$, $\xi = 0.8$, $\bar{\omega} = 6.0$, and $\omega_1(t), \omega_2(t) \in \mathbb{R}$ independent white Gaussian standard noises. As in the LQ problem developed in [4, 27, 40], the matrix $\boldsymbol{Q}_0$ in the functional (3.1) has been chosen as follows:

$$\boldsymbol{Q}_0 = \begin{bmatrix} 10000 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{7.5}$$

In simulations we have supposed to know exactly the initial state

$$z(\tau) = \begin{bmatrix} -0.1 \\ 8.547 \\ 0 \end{bmatrix}, \qquad \tau \in [-0.33, 0],$$

and we have used

$$\boldsymbol{F}_0 = \begin{bmatrix} 0 \\ 0 \\ 10 \end{bmatrix}, \qquad \boldsymbol{G} = 1.$$

Figures 14–17 show the three components of the state and the input for $n = 2$. Computed values of the functional

$$J_{10} = \int_0^{10} [\boldsymbol{z}^{\mathrm{T}}(t)\boldsymbol{Q}_0\boldsymbol{z}(t) + \boldsymbol{u}^2(t)]dt \tag{7.6}$$

for different n and the same noise realization are reported in Table 1. The integration step has been chosen equal to $dT = 0.001$, the integral J_{10} has been computed as $dT\sum_{k=0}^{10/dT} \boldsymbol{z}^{\mathrm{T}}(kdT)\boldsymbol{Q}_0\boldsymbol{z}(kdT) + u^2(kdT)$.

We have considered also the infinite horizon LQ problem: this means that we have considered only the Riccati equation for control and evaluated the dynamic approximated Riccati operator in a sufficiently large time. We have stopped integration when the norm of the difference between the Riccati matrix operators evaluated in time kdT and $(k+1)dT$ was less than 10^{-10}.

Tables 2–5 report the values of matrices Π_0^n and of matrices of functions $\Pi_1^n(\vartheta)$ [4, 27, 40] of the approximated, not yet implementable, LQ control law

$$u_n(t) = \Pi_0^n z(t) + \int_{-r}^{0} \Pi_1^n(\theta)z(t+\theta)d\theta. \tag{7.7}$$

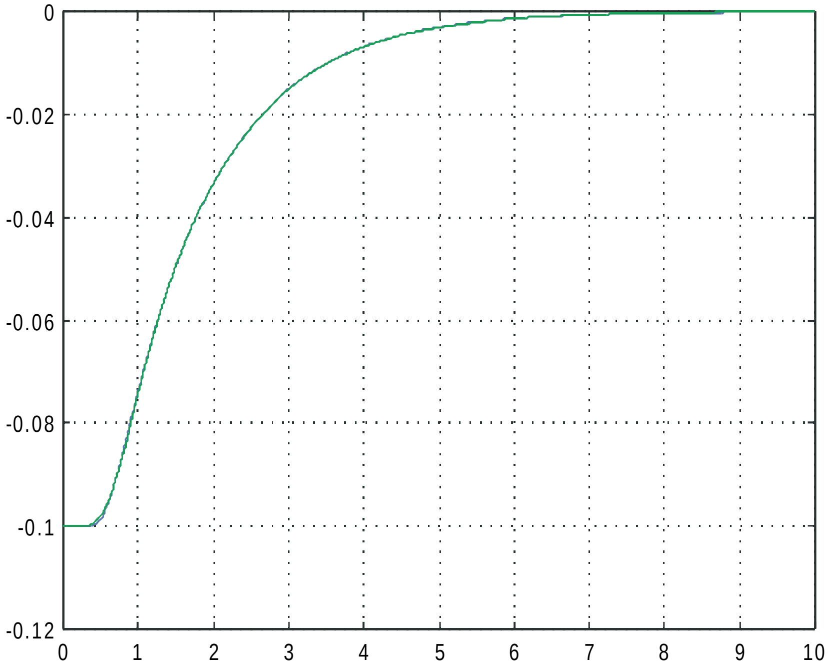


FIG. 14. *Finite horizon LQG for the wind tunnel. The case of $n = 2$: true and estimated $z_1(t)$ (almost coincident).*

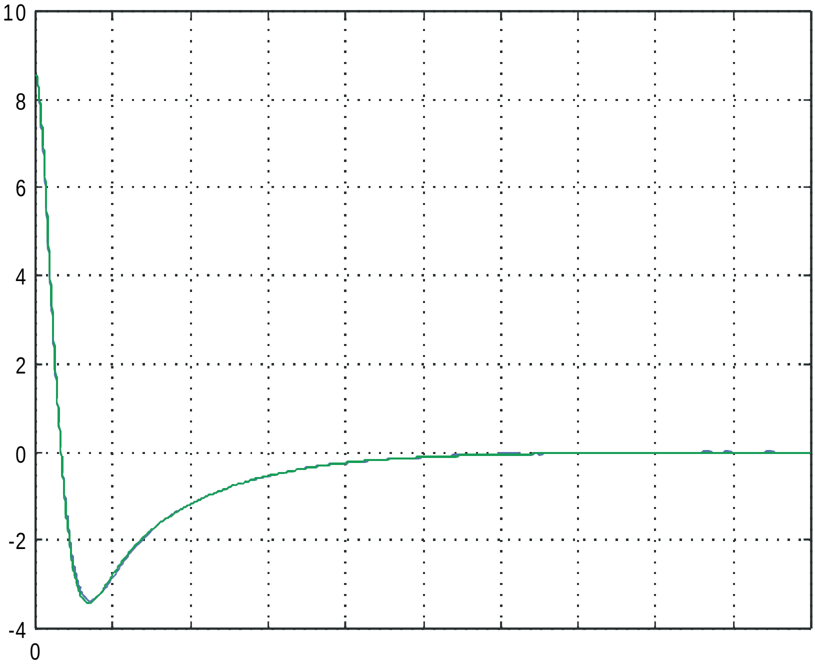


FIG. 15. *Finite horizon LQG for the wind tunnel. The case of $n = 2$: true and estimated $z_2(t)$ (almost coincident).*

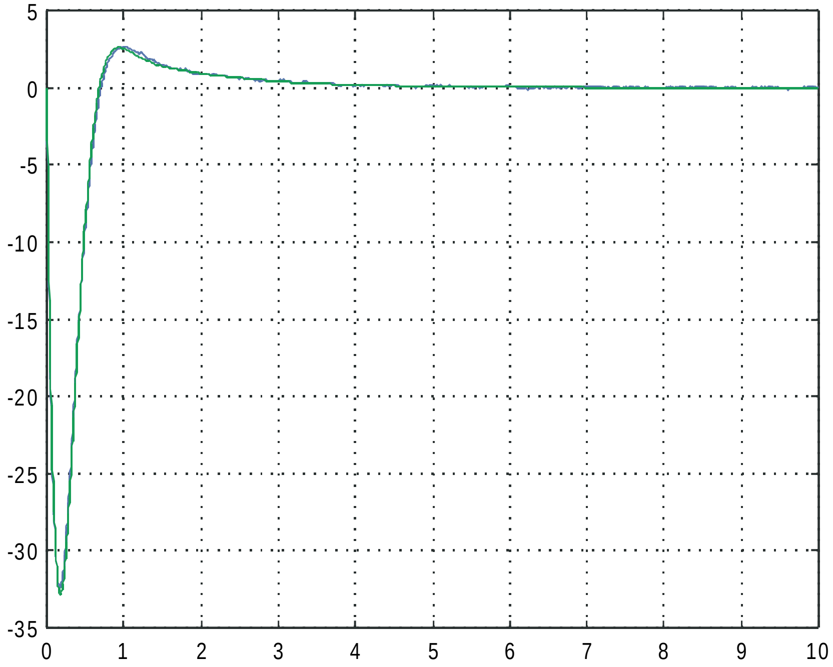


FIG. 16. *Finite horizon LQG for the wind tunnel. The case of $n = 2$: true and estimated $z_3(t)$ (almost coincident).*

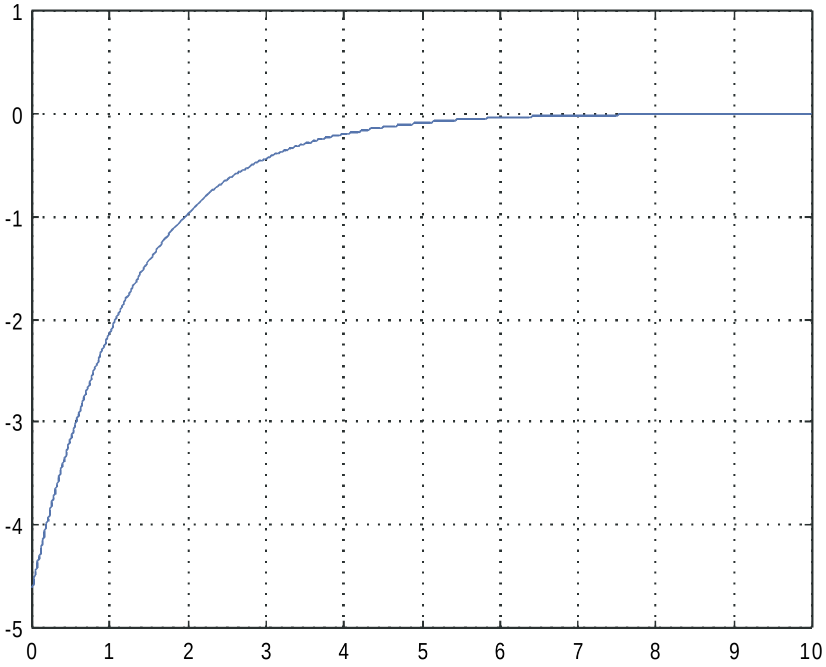


FIG. 17. *Finite horizon LQG for the wind tunnel. The case of $n = 2$: the input $u(t)$.*

Table 1
Values of J_{10} computed for different values of n.

n	J_{10}
2	136.41324
4	136.41311
8	136.41310

Table 2
Values of matrix Π_0^n for different values of n.

$n = 2$	$\begin{bmatrix} 8676.5662 & -9.8145 & -0.9479 \\ -9.8145 & 0.0182 & 0.0018 \\ -0.9479 & 0.0018 & 0.0002 \end{bmatrix}$
$n = 4$	$\begin{bmatrix} 8676.9112 & -9.8149 & -0.9477 \\ -9.8149 & 0.0184 & 0.0018 \\ -0.9477 & 0.0018 & 0.0002 \end{bmatrix}$
$n = 8$	$\begin{bmatrix} 8676.9959 & -9.8150 & -0.9477 \\ -9.8150 & 0.0185 & 0.0018 \\ -0.9477 & 0.0018 & 0.0002 \end{bmatrix}$
$n = 16$	$\begin{bmatrix} 8677.0170 & -9.8150 & -0.9477 \\ -9.8150 & 0.0185 & 0.0018 \\ -0.9477 & 0.0018 & 0.0002 \end{bmatrix}$

Table 3
Values of $\Pi_1^n(1, 2)$ for different values of n.

j	$\Pi_1^2(-jr/16)$	$\Pi_1^4(-jr/16)$	$\Pi_1^8(-jr/16)$	$\Pi_1^{16}(-jr/16)$
0	−41.3798	−41.3931	−41.3962	−41.3970
1	—	—	—	−42.0024
2	—	—	−42.6103	−42.6140
3	—	—	—	−43.2288
4	—	−43.8343	−43.8491	−43.8499
5	—	—	—	−44.4742
6	—	—	−45.1014	−45.1051
7	—	—	—	−45.7393
8	−46.3182	−46.3773	−46.3796	−46.3802
9	—	—	—	−47.0246
10	—	—	−47.6721	−47.6757
11	—	—	—	−48.3306
12	—	−48.9777	−48.9920	−48.9923
13	—	—	—	−49.6579
14	—	—	−50.3271	−50.3306
15	—	—	—	−51.0071
16	−51.6883	−51.6904	−51.6909	−51.6910

Table 4
Values of $\Pi_1^n(2,2)$ *for different values of* n.

j	$\Pi_1^2(-jr/16)$	$\Pi_1^4(-jr/16)$	$\Pi_1^8(-jr/16)$	$\Pi_1^{16}(-jr/16)$
0	0.0684	0.0690	0.0691	0.0692
1	—	—	—	0.0685
2	—	—	0.0678	0.0677
3	—	—	—	0.0670
4	—	0.0665	0.0663	0.0663
5	—	—	—	0.0656
6	—	—	0.0650	0.0649
7	—	—	—	0.0643
8	0.0646	0.0635	0.0636	0.0636
9	—	—	—	0.0629
10	—	—	0.0623	0.0623
11	—	—	—	0.0616
12	—	0.0613	0.0610	0.0610
13	—	—	—	0.0604
14	—	—	0.0598	0.0597
15	—	—	—	0.0591
16	0.0585	0.0585	0.0585	0.0585

Table 5
Values of $\Pi_1^n(3,2)$ *for different values of* n.

j	$\Pi_1^2(-jr/16)$	$\Pi_1^4(-jr/16)$	$\Pi_1^8(-jr/16)$	$\Pi_1^{16}(-jr/16)$
0	0.0067	0.0067	0.0067	0.0067
1	—	—	—	0.0066
2	—	—	0.0065	0.0065
3	—	—	—	0.0065
4	—	0.0064	0.0064	0.0064
5	—	—	—	0.0063
6	—	—	0.0063	0.0063
7	—	—	—	0.0062
8	0.0062	0.0061	0.0061	0.0061
9	—	—	—	0.0061
10	—	—	0.0060	0.0060
11	—	—	—	0.0060
12	—	0.0059	0.0059	0.0059
13	—	—	—	0.0058
14	—	—	0.0058	0.0058
15	—	—	—	0.0057
16	0.0056	0.0056	0.0056	0.0056

In Tables 3–5 the values of the second column, the only one not zero, of matrices Π_1^n of continuous functions are reported, just in instants $-jr/n$, $j = 0, 1, \ldots, n$ (between such points the continuous function in consideration is a one degree polynomial).

In the wind tunnel example, and in all other examples we have simulated, no oscillations appear for $\Pi_1^n(\vartheta)$, which was an important problem arising while consid-

Table 6
Values of Π_0^8 for different approximation schemes.

$[2,24]$	$\begin{bmatrix} 8671.3161 & -9.8336 & -0.9500 \\ -9.8336 & 0.0179 & 0.0018 \\ -0.9500 & 0.0018 & 0.0002 \end{bmatrix}$
[4]	$\begin{bmatrix} 8676.9829 & -9.8154 & -0.9477 \\ -9.8154 & 0.0185 & 0.0019 \\ -0.9477 & 0.0019 & 0.0002 \end{bmatrix}$
[27]	$\begin{bmatrix} 8677.02698 & -9.81505 & -0.94768 \\ -9.81505 & 0.01851 & 0.00186 \\ -0.94768 & 0.00186 & 0.00019 \end{bmatrix}$
[40]	$\begin{bmatrix} 8677.02502 & -9.81503 & -0.94768 \\ -9.81503 & 0.01851 & 0.00186 \\ -0.94768 & 0.00186 & 0.00019 \end{bmatrix}$
[this paper]	$\begin{bmatrix} 8676.99592 & -9.81502 & -0.94769 \\ -9.81502 & 0.01850 & 0.00186 \\ -0.94769 & 0.00186 & 0.00019 \end{bmatrix}$
Π_0	$\begin{bmatrix} 8677.02405 & -9.81505 & -0.94768 \\ -9.81505 & 0.01851 & 0.00186 \\ -0.94768 & 0.00186 & 0.00019 \end{bmatrix}$

Table 7
Values of $J(u_n)$ for different schemes and $n = 8$.

$[2,24]$	[4]	[27]	[40]	[This paper]	Theor. value [27]
136.7361	136.7354	136.40094	136.4490	136.4131	136.40490

ering the approximation scheme [4]. In that approximation scheme Π_1^n is increasingly oscillatory with increasing n, while in our scheme, as in [27, 40], each function in Π_1^n is monotone. This property becomes very important if one wants to implement the approximating feedback law in a real system [27].

In Table 6 the approximations of order $n = 8$ of matrix Π_0, denoted Π_0^8, are reported, computed with approximation schemes in [2, 24, 4, 27, 40] and with the method presented in this paper. Also the exact optimal Π_0 is reported, as computed in [27].

Table 7 reports the values of the functional

$$\int_0^\infty [\boldsymbol{z}^{\mathrm{T}}(t)\boldsymbol{Q}_0\boldsymbol{z}(t) + \boldsymbol{u}^2(t)]dt \tag{7.8}$$

computed using the same approximation schemes for $n = 8$.

The value computed with the method proposed in this paper is obtained by numerical integration of (7.6). The value computed with $t_f = 20$ is quite the same (for $n = 2$ it is $J_{20} = 136.4133$).

TABLE 8
Numerical values of J_{10} computed for different values of n.

n	J_{10}
2	136.41325
4	136.41312
8	136.41311

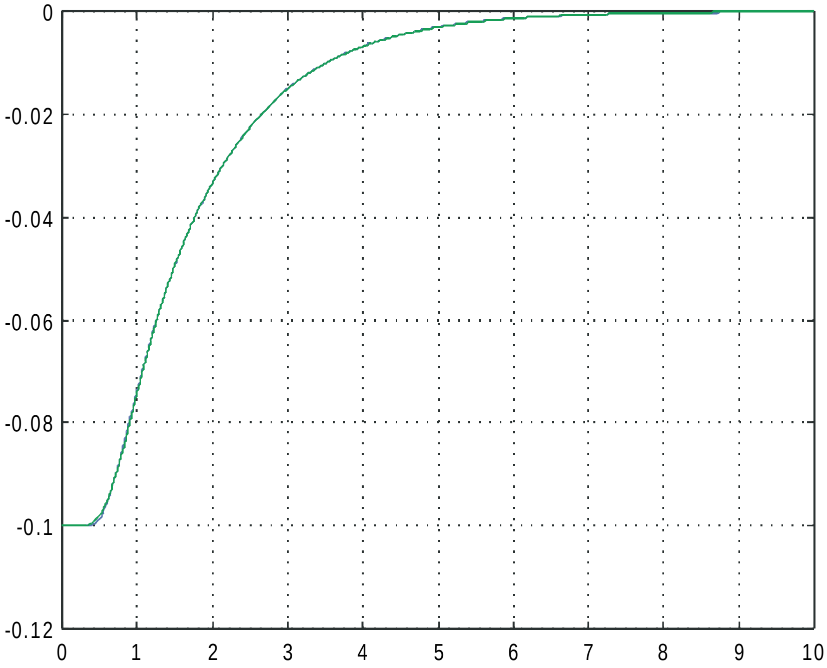


FIG. 18. *Infinite horizon LQG for the wind tunnel. The case of $n = 2$: true and estimated $z_1(t)$ (almost coincident).*

To conclude, the infinite horizon LQG control is considered: the solutions of the approximate algebraic Riccati equations for control and filtering are used in the control scheme. The resulting controller is a dynamic finite dimensional stationary system driven by the noisy output. The values of the index $\int_0^{10}[\boldsymbol{z}^{\mathrm{T}}(t)\boldsymbol{Q}_0\boldsymbol{z}(t)+\boldsymbol{u}^2(t)]dt$, computed for different n and for the same noise realization, are reported in Table 8.

All approximation schemes give, within numerical errors, practically the same value of the cost functional. It is indeed remarkable that the proposed approximation scheme is able to reach such value of the functional starting from noisy output measurements and not, as the other schemes do, starting from noiseless full state information (in a delay interval).

In Figures 18–20 the three components of the state are reported in the infinite horizon case, for $n = 2$. The plots of the input and of the output are reported in Figures 21 and 22.

Comparison with the methods presented in [25, 26] cannot be reported because such papers do not contain numerical results.

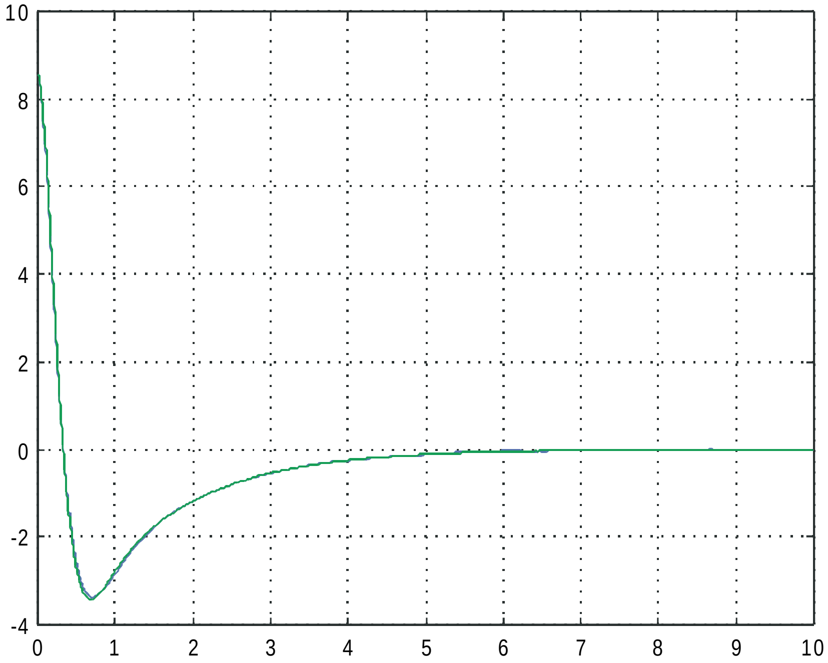


FIG. 19. *Infinite horizon LQG for the wind tunnel. The case of $n = 2$: true and estimated $z_2(t)$ (almost coincident).*

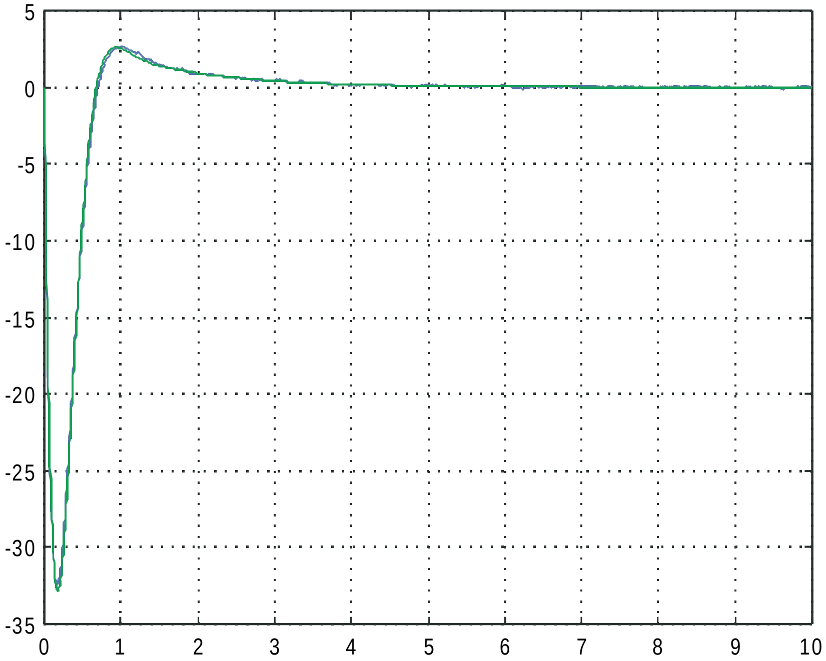


FIG. 20. *Infinite horizon LQG for the wind tunnel. The case of $n = 2$: true and estimated $z_3(t)$ (almost coincident).*

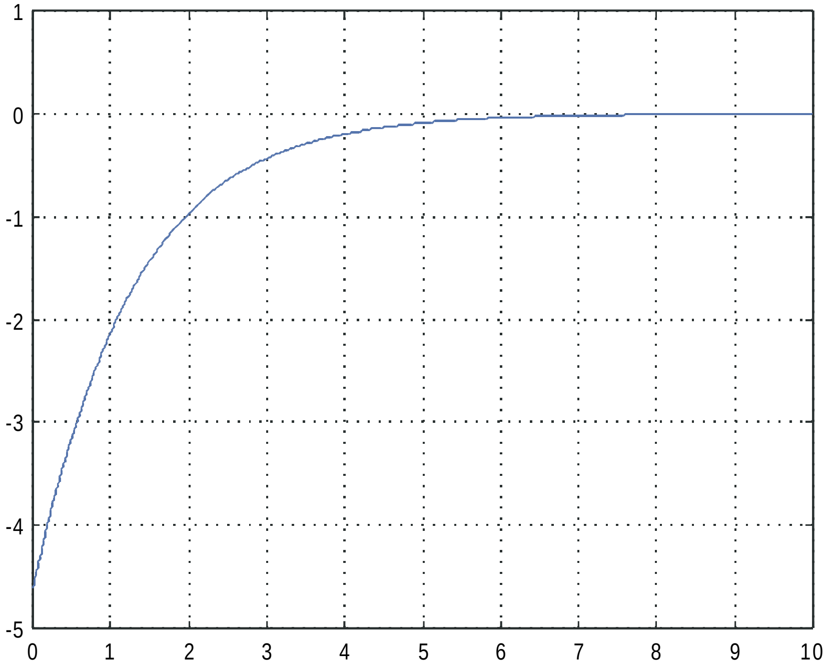


FIG. 21. *Infinite horizon LQG for the wind tunnel. The case of $n = 2$: the input $u(t)$.*

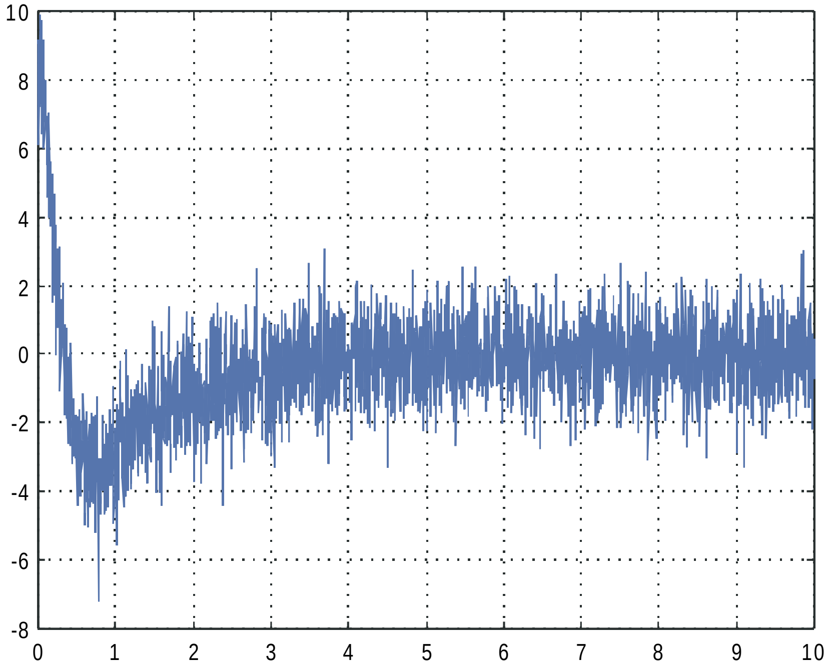


FIG. 22. *Infinite horizon LQG for the wind tunnel. The case of $n = 2$: the noisy output $y(t)$.*

8. Conclusions. In this paper a new spline approximation scheme has been developed for the finite horizon LQG control of general hereditary systems. The approximated LQG controller is a finite dimensional linear system driven by the noisy output. It has been proved that the approximated implementable feedback input and the corresponding state converge to the optimal ones. The approximation scheme makes use of first order splines, introduced by [3] for hereditary systems, suitably adapted to the LQG problem which involves three differential equations, that is, the filter equation and the two Riccati equations required for the computation of the optimal stochastic control. Generally in the literature one of these equations is considered, that is the Riccati equation for the deterministic state feedback optimal control [4, 24, 26, 27, 40]. A methodology with two approximating subspaces has been necessary to apply such spline functions to obtain convergence of the overall LQG problem.

The main feature of the proposed scheme is from a numerical point of view. Indeed our proposal of an implementable LQG controller gives practically the same results of the well-known LQ controller, with a complete knowledge of the infinite dimensional state in a deterministic setting, with reference to an important widely studied case as the NASA National Transonic Facility. The choice of spline environment instead of averaging one is motivated in [4], where its numerical advantages are stressed.

Moreover, the proposed method for choosing splines has the important degree of freedom regarding the possibility of approximating separately the semigroup governing the system and its adjoint. This allows us to use splines of any order [3]. This is very promising for obtaining very good performances in the future.

Future work will involve the infinite horizon LQG problem, which in this paper has been only sketched. For such a problem, the approximation scheme developed here can be used, and convergence of the type in paper [14] can be obtained. As a final remark, we would like to stress that the methodology presented in this paper can involve more than one type of approximation for the three equations governing the LQG stochastic control in order to get the best combination of theoretical and numerical convergences of approximation schemes developed until now [2, 3, 24, 26, 27, 40].

Appendix.

LEMMA A.1. *For any $\boldsymbol{y}_0 \in \mathbb{R}^N$ and for any function $\boldsymbol{f} \in C^0([-r,0];\mathbb{R}^N)$, there exists a unique left-continuous function $\boldsymbol{y}_1 : [-r,0] \mapsto \mathbb{R}^N$ such that*

$$\boldsymbol{y}_1 - \sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{y}_0,\boldsymbol{y}_1)\chi_{[-r,-r_j]} = \boldsymbol{f}, \tag{A.1}$$

where $\boldsymbol{k}_j$ are functions defined in (2.22).

Proof. In the case $\delta = 1$ the summation vanishes and the lemma is trivially true. In the case $\delta > 1$, consider (A.1) in time instants $-r_k$

$$\boldsymbol{y}_1(-r_k) - \sum_{j=1}^{k} \frac{\boldsymbol{y}_1(-r_j) - \boldsymbol{A}_j^{\mathrm{T}}\boldsymbol{y}_0}{\delta - j + 1} = \boldsymbol{f}(-r_k), \quad k = 1,\dots,\delta-1, \tag{A.2}$$

which can be put in matrix form as

$$\begin{bmatrix} \boldsymbol{y}_1(-r_1) \\ \vdots \\ \boldsymbol{y}_1(-r_{\delta-1}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}(-r_1) \\ \vdots \\ \boldsymbol{f}(-r_{\delta-1}) \end{bmatrix} + H_{\delta,2} \begin{bmatrix} \boldsymbol{y}_1(-r_1) \\ \vdots \\ \boldsymbol{y}_1(-r_{\delta-1}) \end{bmatrix} - H_{\delta,2} \begin{bmatrix} \boldsymbol{A}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{A}_{\delta-1}^{\mathrm{T}} \end{bmatrix} \boldsymbol{y}_0, \tag{A.3}$$

where matrix $H_{\delta,2}$ is defined as follows ($\boldsymbol{I}_N$ is the $N \times N$ identity matrix):

$$\text{(A.4)} \qquad H_{\delta,2} = \begin{bmatrix} \frac{1}{\delta}\boldsymbol{I}_N & 0 & \cdots & 0 & 0 \\ \frac{1}{\delta}\boldsymbol{I}_N & \frac{1}{\delta-1}\boldsymbol{I}_N & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{\delta}\boldsymbol{I}_N & \frac{1}{\delta-1}\boldsymbol{I}_N & \cdots & \frac{1}{3}\boldsymbol{I}_N & 0 \\ \frac{1}{\delta}\boldsymbol{I}_N & \frac{1}{\delta-1}\boldsymbol{I}_N & \cdots & \frac{1}{3}\boldsymbol{I}_N & \frac{1}{2}\boldsymbol{I}_N \end{bmatrix}.$$

Now, let us define a vector $\boldsymbol{\eta} \in \mathbb{R}^{(\delta-1)N}$ as

$$\text{(A.5)} \qquad \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_{\delta-1} \end{bmatrix} = (\boldsymbol{I}_{(\delta-1)N} - H_{\delta,2})^{-1} \left(\begin{bmatrix} \boldsymbol{f}(-r_1) \\ \vdots \\ \boldsymbol{f}(-r_{\delta-1}) \end{bmatrix} - H_{\delta,2} \begin{bmatrix} \boldsymbol{A}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{A}_{\delta-1}^{\mathrm{T}} \end{bmatrix} \boldsymbol{y}_0 \right)$$

and the left-continuous function

$$\text{(A.6)} \qquad \bar{\boldsymbol{y}}_1(\vartheta) = \begin{cases} \boldsymbol{\eta}_i, & \vartheta = -r_i, \\ \boldsymbol{f}(\vartheta) + \displaystyle\sum_{j=1}^{\delta-1} \frac{\boldsymbol{\eta}_j - \boldsymbol{A}_j^{\mathrm{T}}\boldsymbol{y}_0}{\delta - j + 1} \chi_{[-r,-r_j]}(\vartheta), & \vartheta \neq -r_i, \end{cases}$$

in which $i = 1, \dots, \delta - 1$. It is readily verified that $\bar{\boldsymbol{y}}_1$ satisfies (A.1).

Uniqueness is proved by recognizing that any other function $\tilde{\boldsymbol{y}}_1$ satisfying (A.1) verifies also (A.3), and therefore $\tilde{\boldsymbol{y}}_1(-r_k) = \bar{\boldsymbol{y}}_1(-r_k)$, $k = 1, \dots, \delta - 1$. The difference between expression (A.1) with $\boldsymbol{y}_1 = \bar{\boldsymbol{y}}_1$ and the same expression in which $\boldsymbol{y}_1 = \tilde{\boldsymbol{y}}_1$ is used gives $\bar{\boldsymbol{y}}_1 - \tilde{\boldsymbol{y}}_1 = 0$. This concludes the proof of uniqueness. □

Proof of Proposition 2.2. Only (2.20) and (2.21) require a little mathematics. The case $\delta = 1$ (summations in (2.20) and (2.21) vanish) is a standard result [24, 43]. For the case $\delta > 1$, let $L : \mathcal{D}(L) \mapsto \boldsymbol{M}_2$ be the operator defined as (see (2.20), (2.21))

$$L : \mathcal{D}(L) \mapsto \boldsymbol{M}_2,$$

$$L \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{bmatrix} = \begin{bmatrix} \delta\, \boldsymbol{y}_1(0) + \boldsymbol{A}_0^{\mathrm{T}}\boldsymbol{y}_0 \\ \dfrac{1}{g}\boldsymbol{A}_{01}^{\mathrm{T}}\boldsymbol{y}_0 - \dfrac{d}{d\vartheta}\left(\boldsymbol{y}_1 - \displaystyle\sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_1)\chi_{[-r,-r_j]} \right) \end{bmatrix},$$

$$\mathcal{D}(L) = \left\{ \begin{bmatrix} \boldsymbol{y}_0 \\ \\ \boldsymbol{y}_1 \end{bmatrix} \,\middle|\, \begin{matrix} \boldsymbol{y}_0 \in \mathbb{R}^N, \quad \boldsymbol{A}_\delta^{\mathrm{T}}\boldsymbol{y}_0 = \boldsymbol{y}_1(-r), \\ \left(\boldsymbol{y}_1 - \displaystyle\sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_1)\chi_{[-r,-r_j]} \right) \in W^{1,2} \end{matrix} \right\}.$$

We will show that

(a) for every $\boldsymbol{x} \in \mathcal{D}(\boldsymbol{A})$ and every $y \in \mathcal{D}(L)$, it is $(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}) - (L\boldsymbol{y}, \boldsymbol{x}) = 0$;

(b) the set $\mathcal{D}(L)$ is dense in M_2.

These two items together state that L is the adjoint of $\boldsymbol{A}$, that is, $\boldsymbol{A}^*$ as defined in (2.20), (2.21).

Let us prove item (a). Take $\boldsymbol{x} \in \mathcal{D}(\boldsymbol{A})$ and $\boldsymbol{y} \in \mathcal{D}(L)$, let us show that $(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}) - (L\boldsymbol{y}, \boldsymbol{x}) = 0$:

$$\begin{aligned}(\boldsymbol{y},\boldsymbol{A}\boldsymbol{x}) - (L\boldsymbol{y},\boldsymbol{x}) &= \boldsymbol{y}_0^{\mathrm{T}}\textstyle\sum_{j=1}^{\delta}\boldsymbol{A}_j\boldsymbol{x}_1(-r_j) - \delta\,\boldsymbol{y}_1^{\mathrm{T}}(0)\boldsymbol{x}_0\\ &\quad+ \textstyle\int_{-r}^{0} g(\vartheta)\left[\boldsymbol{y}_1^{\mathrm{T}}(\vartheta)\frac{d}{d\vartheta}\boldsymbol{x}_1(\vartheta) + \frac{d}{d\vartheta}\left(\boldsymbol{y}_1 - \sum_{j=1}^{\delta-1}\boldsymbol{k}_j(\boldsymbol{y}_0,\boldsymbol{y}_1)\chi_{[-r,-r_j]}\right)^{\mathrm{T}}\boldsymbol{x}_1(\vartheta)\right]d\vartheta.\end{aligned}\tag{A.7}$$

$\boldsymbol{y}_1 - \sum_{j=1}^{\delta-1}\boldsymbol{k}_j(\boldsymbol{y}_0,\boldsymbol{y}_1)\chi_{[-r,-r_j]}$ being absolutely continuous, the integral term in (A.7) can be rewritten as

$$\begin{aligned}&\sum_{i=1}^{\delta}\int_{-r_i}^{-r_{i-1}}(\delta-i+1)\frac{d}{d\vartheta}\left[\left(\boldsymbol{y}_1 - \sum_{j=1}^{\delta-1}\boldsymbol{k}_j(\boldsymbol{y}_0,\boldsymbol{y}_1)\chi_{[-r,-r_j]}\right)^{\mathrm{T}}\boldsymbol{x}_1(\vartheta)\right]d\vartheta\\ &\quad+\sum_{i=1}^{\delta}\int_{-r_i}^{-r_{i-1}}(\delta-i+1)\left(\sum_{j=1}^{\delta-1}\boldsymbol{k}_j(\boldsymbol{y}_0,\boldsymbol{y}_1)\chi_{[-r,-r_j]}\right)^{\mathrm{T}}\frac{d}{d\vartheta}\boldsymbol{x}_1(\vartheta)d\vartheta\end{aligned}\tag{A.8}$$

and after a simple computation

$$\begin{aligned}&\sum_{i=1}^{\delta}(\delta-i+1)\left[\left(\boldsymbol{y}_1 - \sum_{j=1}^{\delta-1}\boldsymbol{k}_j(\boldsymbol{y}_0,\boldsymbol{y}_1)\chi_{[-r,-r_j]}\right)^{\mathrm{T}}\boldsymbol{x}_1\right]_{-r_i}^{-r_{i-1}}\\ &\quad+\sum_{i=2}^{\delta}(\delta-i+1)\left(\sum_{h=1}^{i-1}\boldsymbol{k}_h(\boldsymbol{y}_0,\boldsymbol{y}_1)\right)^{\mathrm{T}}[\boldsymbol{x}_1(-r_{i-1}) - \boldsymbol{x}_1(-r_i)]\\ &= \delta\boldsymbol{y}_1^{\mathrm{T}}(0)\,\boldsymbol{x}_1(0) - \boldsymbol{y}_1^{\mathrm{T}}(-r)\,\boldsymbol{x}_1(-r) - \sum_{i=1}^{\delta-1}\boldsymbol{y}_0^{\mathrm{T}}\boldsymbol{A}_i\boldsymbol{x}_1(-r_i).\end{aligned}\tag{A.9}$$

Now, replacing the integral term in (A.7) with the above expression and taking into account that $\boldsymbol{x}_1(0) = \boldsymbol{x}_0$ and $\boldsymbol{y}_1(-r) = \boldsymbol{A}_\delta^{\mathrm{T}}\boldsymbol{y}_0$, it follows that $(\boldsymbol{y},\boldsymbol{A}\boldsymbol{x}) - (L\boldsymbol{y},\boldsymbol{x}) = 0$.

Let us now prove item (b). It is sufficient to prove that $\mathcal{D}(L)$ is dense in $\mathbb{R}^N \times W^{1,2}$. Let $\boldsymbol{y}_0 \in \mathbb{R}^N$, $\boldsymbol{y}_1 \in W^{1,2}$. Consider the following sequence of functions $\boldsymbol{y}_{1,k} \in W^{1,2}$, defined for integers $k > \frac{1}{r-r_{\delta-1}}$:

$$\boldsymbol{y}_{1,k}(\vartheta) = \begin{cases}\boldsymbol{y}_1(\vartheta), & \vartheta \in [-r+\frac{1}{k}, 0],\\ (1-k(\vartheta+r))\boldsymbol{A}_\delta^{\mathrm{T}}\boldsymbol{y}_0 + k(\vartheta+r)\boldsymbol{y}_1(-r+\frac{1}{k}), & \vartheta \in [-r, -r+\frac{1}{k}).\end{cases}\tag{A.10}$$

Being that $-r + 1/k < -r_{\delta-1}$, it is $\boldsymbol{y}_{1,k}(-r_j) = \boldsymbol{y}_1(-r_j)$ for $j = 1,\dots,\delta-1$. As $\boldsymbol{y}_1$ is uniformly bounded in $[-r, 0]$, given any positive ϵ there exists k_ϵ such that

$$\|\boldsymbol{y}_1 - \boldsymbol{y}_{1,k}\|_{L_2} < \frac{\epsilon}{2} \quad \text{for} \quad k > k_\epsilon.\tag{A.11}$$

Note that, since $\boldsymbol{y}_{1,k}(-r) = \boldsymbol{A}_d^{\mathrm{T}}\boldsymbol{y}_0$, if $\delta = 1$, then $\left[\begin{smallmatrix}\boldsymbol{y}_0\\ \boldsymbol{y}_{1,k}\end{smallmatrix}\right] \in \mathcal{D}(\boldsymbol{A}^*)$, and the density of $\mathcal{D}(\boldsymbol{A}^*)$ in $\mathbb{R}^N \times W^{1,2}$ and hence in $\boldsymbol{M}_2$ is proved.

If $\delta > 1$ in general $\left[\begin{smallmatrix}\boldsymbol{y}_0\\ \boldsymbol{y}_{1,k}\end{smallmatrix}\right] \notin \mathcal{D}(\boldsymbol{A}^*)$. For any integer $n > \sup_{j=1,2,\dots,d}\frac{1}{r_j - r_{j-1}}$ it is convenient to define δ functions in $W^{1,2}$ as follows:

$$\chi_j^n(\vartheta) = \begin{cases}\chi_{[-r,-r_j]}(\vartheta), & \vartheta \notin (-r_j, -r_j+\frac{1}{n}),\\ -n(\vartheta + r_j - \frac{1}{n}), & \vartheta \in (-r_j, -r_j+\frac{1}{n}),\end{cases} \quad \text{for} \quad j = 1,\dots,\delta-1,\tag{A.12}$$

are such that

$$\|\chi_j^n - \chi_{[-r,-r_j]}\|_{L_2} = \frac{1}{\sqrt{3n}}, \quad \text{for } j = 1, \dots, \delta - 1. \tag{A.13}$$

By Lemma A.1, for any $n > \sup_{j=1,2,\dots,d} \frac{1}{r_j - r_{j-1}}$, there exists a function $\boldsymbol{y}_{1,k,n}$ such that

$$\boldsymbol{y}_{1,k,n} - \sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_{1,k,n})\chi_{[-r,-r_j]} = \boldsymbol{y}_{1,k} - \sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_{1,k})\chi_j^n \tag{A.14}$$

(note that the right-hand side term is in C^0 and therefore Lemma A.1 can be applied).

It can be shown that $\boldsymbol{y}_{1,k,n}(-r_j) = \boldsymbol{y}_{1,k}(-r_j)$, $j = 1, \dots, \delta$, so that $\left[\begin{smallmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_{1,k,n} \end{smallmatrix}\right] \in \mathcal{D}(\boldsymbol{A}^*)$.

Moreover,

$$\begin{aligned} \left\| \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_{1,k} \end{bmatrix} - \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_{1,k,n} \end{bmatrix} \right\|_{\boldsymbol{M}_2} &= \left\| \sum_{j=1}^{\delta-1} \boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_1)(\chi_j^n - \chi_{[-r,-r_j]}) \right\|_{L_2} \\ &\le \sum_{j=1}^{\delta-1} \left\| \boldsymbol{k}_j(\boldsymbol{y}_0, \boldsymbol{y}_1) \right\| \frac{1}{\sqrt{3n}}, \end{aligned} \tag{A.15}$$

where formula (A.13) is used.

Thus, there exists an integer n_ϵ such that for $n > n_\epsilon$ it is

$$\left\| \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_{1,k} \end{bmatrix} - \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_{1,k,n} \end{bmatrix} \right\|_{\boldsymbol{M}_2} < \frac{\epsilon}{2}. \tag{A.16}$$

Finally, for any pair k, n such that $k > k_\epsilon$ and $n > n_\epsilon$ it is

$$\left\| \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \end{bmatrix} - \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_{1,k,n} \end{bmatrix} \right\|_{\boldsymbol{M}_2} < \epsilon, \tag{A.17}$$

which proves the density of $\mathcal{D}(L)$ in $\boldsymbol{M}_2$. □

Remark A.2 The proof of this proposition concerning the adjoint operator $\boldsymbol{A}^\star$ can also be done by methodology shown in [15]. Using standard Lax–Milgram-type representation theorems, a relationship follows between equivalent inner products, so that an adjoint operator in a given inner product can be found by another one obtained in an equivalent inner product (see [15] and references therein).

LEMMA A.3. *For any nonnegative λ the matrix $H_p(\lambda)$ defined in* (4.59) *is nonsingular.*

Proof. The expression of $H_p(\lambda)$ is here reported for the reader's convenience:

$$H_p(\lambda) = \boldsymbol{I}_{(\delta-1)N} - H_{\delta,2} + \begin{bmatrix} \boldsymbol{I}_N e^{-\lambda(r-r_1)} \\ \vdots \\ \boldsymbol{I}_N e^{-\lambda(r-r_{\delta-1})} \end{bmatrix} h_{\delta,2}. \tag{A.18}$$

As a first step, nonsingularity of matrix $\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2}$ is proved. By the definition of $H_{\delta,2}$ in (A.4) it is

$$\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2} = \begin{bmatrix} \boldsymbol{I}_N - \frac{1}{\delta}\boldsymbol{I}_N & 0 & \cdots & 0 & 0 \\ -\frac{1}{\delta}\boldsymbol{I}_N & \boldsymbol{I}_N - \frac{1}{\delta-1}\boldsymbol{I}_N & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{\delta}\boldsymbol{I}_N & -\frac{1}{\delta-1}\boldsymbol{I}_N & \cdots & \boldsymbol{I}_N - \frac{1}{3}\boldsymbol{I}_N & 0 \\ -\frac{1}{\delta}\boldsymbol{I}_N & -\frac{1}{\delta-1}\boldsymbol{I}_N & \cdots & -\frac{1}{3}\boldsymbol{I}_N & \boldsymbol{I}_N - \frac{1}{2}\boldsymbol{I}_N \end{bmatrix}. \tag{A.19}$$

A direct computation shows that the inverse of $\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2}$ is

$$(\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2})^{-1} = \boldsymbol{I}_{N(\delta-1)} + \overline{H}_{\delta,2}, \tag{A.20}$$

where matrix $\overline{H}_{\delta,2}$ is defined as

$$\overline{H}_{\delta,2} = \begin{bmatrix} \frac{1}{\delta-1}\boldsymbol{I}_N & 0 & \cdots & 0 & 0 \\ \frac{1}{\delta-2}\boldsymbol{I}_N & \frac{1}{\delta-2}\boldsymbol{I}_N & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2}\boldsymbol{I}_N & \frac{1}{2}\boldsymbol{I}_N & \cdots & \frac{1}{2}\boldsymbol{I}_N & 0 \\ \boldsymbol{I}_N & \boldsymbol{I}_N & \cdots & \boldsymbol{I}_N & \boldsymbol{I}_N \end{bmatrix}. \tag{A.21}$$

The verification can be made writing the following expression for the kth column block of matrix $\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2}$ as

$$\begin{bmatrix} 0 \\ \vdots \\ \boldsymbol{I}_N \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ \vdots \\ \frac{1}{\delta-(k-1)}\boldsymbol{I}_N \\ \vdots \\ \frac{1}{\delta-(k-1)}\boldsymbol{I}_N \end{bmatrix} \tag{A.22}$$

(the first $k-1$ blocks are zero) and the following expression for the jth row block of matrix $\left(\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2}\right)^{-1}$ as

$$[0 \quad \cdots \quad 0 \quad \boldsymbol{I}_N \quad 0 \quad \cdots 0] + \left[\tfrac{1}{\delta-j}\boldsymbol{I}_N \quad \ldots \quad \tfrac{1}{\delta-j}\boldsymbol{I}_N \quad 0 \quad \cdots \quad 0\right] \tag{A.23}$$

(the first j blocks are nonzero). The product when $j < k$ is a sum of zeroes and is trivially zero. It can also be verified that when $j > k$, the product gives zero, and when $j = k$, the product gives $\boldsymbol{I}_N$. This verifies the expression (A.20) for $(\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2})^{-1}$.

Given the invertible matrix $\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2}$, the determinant of $H_p(\lambda)$ can be written as follows:

(A.24)

$$\det(\mathbf{I}_{N(\delta-1)} - H_{\delta,2}) \cdot \det\left(\mathbf{I}_{(\delta-1)N} + (\mathbf{I}_{N(\delta-1)} - H_{\delta,2})^{-1} \begin{bmatrix} \mathbf{I}_N e^{-\lambda(r-r_1)} \\ \vdots \\ \mathbf{I}_N e^{-\lambda(r-r_{\delta-1})} \end{bmatrix} h_{\delta,2}\right).$$

Since for any pair of matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$

$$\det(\boldsymbol{I}_n + A \cdot B) = \det(\boldsymbol{I}_m + B \cdot A), \tag{A.25}$$

the determinant of $H_p(\lambda)$ can also be written as

(A.26)

$$\det(\mathbf{I}_{N(\delta-1)} - H_{\delta,2}) \det\left(\mathbf{I}_{(\delta-1)B} + h_{\delta,2}(\mathbf{I}_{N(\delta-1)} - H_{\delta,2})^{-1} \begin{bmatrix} \mathbf{I}_N e^{-\lambda(r-r_1)} \\ \vdots \\ \mathbf{I}_N e^{-\lambda(r-r_{\delta-1})} \end{bmatrix}\right).$$

Recalling the expression of $h_{\delta,2}$ defined in (4.56) it follows that

$$h_{\delta,2}\left(\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2}\right)^{-1} \begin{bmatrix} \boldsymbol{I}_N e^{-\lambda(r-r_1)} \\ \vdots \\ \boldsymbol{I}_N e^{-\lambda(r-r_{\delta-1})} \end{bmatrix} = c(\lambda)\boldsymbol{I}_N, \tag{A.27}$$

where $c(\lambda)$ is the sum of positive terms that are functions of λ. Therefore it is

$$\det(H_p(\lambda)) = \det(\boldsymbol{I}_{N(\delta-1)} - H_{\delta,2}) \det((1 + c(\lambda))\boldsymbol{I}_N) = \frac{1}{\delta^N}(1 + c(\lambda))^N, \tag{A.28}$$

and this proves the nonsingularity of $H_p(\lambda)$ for any nonnegative λ.

REFERENCES

[1] A. V. Balakrishnan, *Applied Functional Analysis*, Springer-Verlag, New York, 1981.

[2] H. T. Banks and J. A. Burns, *Hereditary control problems: Numerical methods based on averaging approximations*, SIAM J. Control Optimization, 16 (1978), pp. 169–208.

[3] H. T. Banks and F. Kappel, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.

[4] H. T. Banks, G. I. Rosen, and K. Ito, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830–855.

[5] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Birkhaüser, Boston, 1992.

[6] J. Burns, K. Ito, and G. Propst, *On nonconvergence of adjoint semigroups for control systems with delays*, SIAM J. Control Optim., 26 (1988), pp. 1441–1454.

[7] F. Colonius, *Stable and regular reachability for relaxed hereditary differential systems*, SIAM J. Control Optim., 20 (1982), pp. 675–694.

[8] F. Colonius, *Addendum: Stable and regular reachability of relaxed hereditary differential systems*, SIAM J. Control Optim., 23 (1985), pp. 803–807.

[9] F. Colonius, *The maximum principle for relaxed hereditary differential systems with function space end condition*, SIAM J. Control Optim., 5 (1982), pp. 695–712.

[10] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, 1978.

[11] M. C. Delfour, *The linear quadratic optimal control problem for hereditary differential systems: Theory and numerical solution*, Appl. Math. Optim., 3 (1977), pp. 101–162.

[12] M. C. Delfour, *The linear-quadratic optimal control problem with delays in state and control variables: A state space approach*, SIAM J. Control Optim., 24 (1986), pp. 835–883.

[13] M. C. Delfour and S. K. Mitter, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, SIAM J. Control, 10 (1972), pp. 298–328.

[14] A. De Santis, A. Germani, and L. Jetto, *Approximation of the algebraic Riccati equation in the Hilbert space of Hilbert-Schmidt operators*, SIAM J. Control Optim., 4 (1993), pp. 847–874.

[15] R. H. Fabiano, *Stability preserving spline approximations for scalar functional differential equations*, Computers Math. Appl., 8 (1995), pp. 87–94.

[16] R. H. Fabiano and K. Ito, *Semigroup theory and numerical approximation for equations in linear viscoelasticity*, SIAM J. Math. Anal., 21 (1990), pp. 374–393.

[17] A. Feliachi and A. Thowsen, *Memoryless stabilization of linear delay differential systems*, IEEE Trans. Automat. Control, 2 (1981), pp. 586–587.

[18] A. Germani and L. Jetto, *A twofold spline scheme approximation method for the LQG control problem on Hilbert spaces*, in Proceedings of the 12th IFAC Triennal World Congress, Sydney, Australia, 1993, Vol. 2, pp. 401–404.

[19] A. Germani, L. Jetto, C. Manes, and P. Pepe, *The LQG control problem for a class of hereditary systems: A method for computing its approximate solution*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Orlando, FL, 1993, Vol. 2, IEEE, New York, 1993, pp. 1362–1367.

[20] A. Germani, L. Jetto, and M. Piccioni, *Galerkin approximation for optimal filtering of infinite-dimensional linear systems*, SIAM J. Control Optim., 26 (1988), pp. 1287–1305.

[21] A. Germani, C. Manes, and P. Pepe, *Numerical solution for optimal regulation of stochastic hereditary systems with multiple discrete delays*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, Vol. 2, IEEE, New York, pp. 1497–1502.

[22] A. Germani, C. Manes, and P. Pepe, *Implementation of an LQG control scheme for linear systems with delayed feedback action*, in Proceedings of the 3rd European Control Conference, Roma, Italy, 1995, Vol. 4, EUCA, Rome, 1995, pp. 2886–2891.

[23] J. S. Gibson, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.

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

[25] K. Ito, *Finite-dimensional compensators for infinite-dimensional systems via Galerkin-type approximation*, SIAM J. Control Optim., 28 (1990), pp. 1251–1269.

[26] K. Ito and F. Kappel, *A uniformly differentiable approximation scheme for delay systems using splines*, Appl. Math. Optim., 23 (1991), pp. 217–262.

[27] F. Kappel and D. Salamon, *Spline approximation for retarded systems and the Riccati equation*, SIAM J. Control Optim., 25 (1987), pp. 1082–1117.

[28] F. Kappel and D. Salamon, *An approximation theorem for the algebraic Riccati equation*, SIAM J. Control Optim., 28 (1990), pp. 1136–1147.

[29] F. Kappel and D. Salamon, *On the stability properties of spline approximations for retarded systems*, SIAM J. Control Optim., 27 (1989), pp. 407–431.

[30] K. Kunisch, *Approximation schemes for the linear-quadratic optimal control problem associated with delay equations*, SIAM J. Control Optim., 20 (1982), pp. 506–540.

[31] R. H. Kwong, *A stability theory for the linear-quadratic-Gaussian problem for systems with delays in the state, control, and observations*, SIAM J. Control Optim., 18 (1980), pp. 49–75.

[32] I. Lasiecka and A. Manitius, *Differentiability and convergence rates of approximating semigroups for retarded functional differential equations*, SIAM J. Numer. Anal, 25 (1988), pp. 883–907.

[33] A. Manitius and H. T. Tran, *Numerical approximations for hereditary systems with input and output delays: Convergence results and convergence rates*, SIAM J. Control Optim., 32 (1994), pp. 1332–1363.

[34] T. Mori, *Criteria for asymptotic stability of linear time delay systems*, IEEE Trans. Automat. Control, 2 (1985), pp. 158–161.

[35] K. A. Morris, *Convergence of controllers designed using state-space techniques*, IEEE Trans. Automat. Control, 10 (1994), pp. 2100–2104.

[36] K. A. Morris, *Design of finite-dimensional controllers for infinite-dimensional systems by approximation*, J. Math. Systems Estim. Control, 4 (1994), pp. 1–30.

[37] L. Pandolfi, *The standard regulator problem for systems with input delays. An approach through singular control theory*, Appl. Math. Optim., 31 (1995), pp. 119–136.

[38] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[39] P. Pepe, *Il Controllo LQG dei Sistemi con Ritardo*, Ph. D. Thesis, Department of Electrical Engineering, Università degli studi dell'Aquila, Italy, 1996.

[40] G. Propst, *Piecewise linear approximation for hereditary control problems*, SIAM J. Control Optim., 28 (1990), pp. 70–96.

[41] D. Salamon, *Structure and stability of finite-dimensional approximations for functional-differential equations*, SIAM J. Control Optim., 23 (1985), pp. 928–951.

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

[43] R. B. Vinter, *On the evolution of the state of linear differential delay equations in M^2: Properties of the generator*, J. Inst. Math. Appl., 21 (1978), pp. 13–23.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 4, pp. 1296–1322

DIFFERENTIAL GAMES AND NONLINEAR $\mathcal{H}_\infty$ CONTROL IN INFINITE DIMENSIONS*

MACIEJ KOCAN† AND PIERPAOLO SORAVIA‡

Abstract. This paper studies the $\mathcal{H}_\infty$ control problem for a nonlinear, unbounded, infinite dimensional system with state constraints. We characterize the solvability of the problem by means of a Hamilton–Jacobi–Isaacs (HJI) equation, proving that the $\mathcal{H}_\infty$ problem can be solved if and only if the HJI equation has a positive definite viscosity supersolution, vanishing and continuous at the origin. In order to do so, the standard definition of the $\mathcal{H}_\infty$ problem has to be relaxed by using the theory of differential games. We apply our results to the one phase Stefan problem.

Key words. viscosity solutions, nonlinear semigroups, accretive operators, dynamic programming, nonlinear $\mathcal{H}_\infty$ control, optimality principles

AMS subject classifications. 93B36, 49J20, 49L25, 90D25

PII. S0363012998343562

1. Introduction.

We study the infinite dimensional, controlled dynamical system subject to unknown deterministic disturbances with full state information

$$\left\{\begin{array}{ll} y'(t) + Ay(t) \ni f(y(t)) + Bu(t) + Cw(t) & \text{for } t \geq 0, \\ y(0) = x \in \overline{D(A)}. \end{array}\right. \tag{1.1}$$

The state space H, the control set U, and the disturbance set W are real Hilbert spaces; $A : D(A) \subset H \to H$ is a maximal monotone operator in H, possibly nonlinear and multivalued, with $0 \in \overline{D(A)}$. We will assume that

$$B\colon\ U \to H, \quad C\colon\ W \to H \quad \text{are bounded and linear operators,} \tag{1.2}$$

$$f\colon\ \overline{D(A)} \to H \quad \text{is Lipschitz continuous.} \tag{1.3}$$

For $x \in \overline{D(A)}$, disturbance $w \in L^2_{\text{loc}}(0,+\infty;W)$, and control $u \in L^2_{\text{loc}}(0,+\infty;U)$, it is well known that with the previous assumptions system (1.1) admits a unique trajectory solution $y(\cdot) \equiv y(\cdot,x,u,w) \in C([0,\infty);H)$ in the *mild* sense. For the general theory of nonlinear semigroups and its applications to partial differential equations the reader can consult, for instance, the books by Brézis [8], Barbu [2], [3], Deimling [11], Vrabie [32], and Evans [14]. Since the operators B, C are linear and bounded, our state equation does not apply to systems modeling boundary control problems. However, we put ourselves under assumptions that include large classes of distributed parameter systems, such as, for instance, parabolic variational inequalities describing free boundary parabolic problems. Such systems, in order to be modeled as abstract evolution equations, require the use of a nonlinear, unbounded, and multivalued A.

*Received by the editors August 17, 1998; accepted for publication (in revised form) August 17, 2000; published electronically December 20, 2000.
http://www.siam.org/journals/sicon/39-4/34356.html

†Mathematics Institute, University of Cologne, 50923 Cologne, Germany. The work of this author was supported by the Australian Research Council and an Alexander von Humboldt Fellowship.

‡Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, 35131 Padova, Italy (soravia@math.unipd.it). The work of this author was partially supported by the EEC-TMR network project "Viscosity Solutions and Their Applications" and by the MURST-Cofin project "Analisi e controllo di equazioni di evoluzione deterministiche e stocastiche."

For explicit examples, the reader can check Barbu [3], Vrabie [32], and [17]. Moreover, in section 2 we recall the abstract setting for the one phase Stefan problem, the prototype of free boundary parabolic problems and well-known model for the melting of ice, and apply our results to that example. We do not seek the most general assumptions on the right-hand side of the differential equation in (1.1) under which our techniques work, in order to reduce some technicalities that might obscure some of the delicate steps of the arguments. For instance, every result could be formulated in the generality of [17].

Our goal is to extend to nonlinear, infinite dimensional systems the equivalence between solvability of the $\mathcal{H}_\infty$ suboptimal control problem for (1.1) and the existence of a nonnegative (positive definite) (super)solution of the corresponding Hamilton–Jacobi–Isaacs (HJI) equation. More precisely, we define the problem as follows. Given $\gamma > 0$ (the disturbance attenuation level) and

$$g\colon \overline{D(A)} \to [0, +\infty] \quad \text{lower semicontinuous}, \quad g(0) = 0, \tag{1.4}$$

we are concerned with existence of a family of *strategies* for the controller $\{\alpha_x[w]\}_{x\in\overline{D(A)}}$, i.e., causal functionals of the disturbance w, which guarantee that the undisturbed system ($w \equiv 0$ in (1.1)) is locally (or locally asymptotically) Lyapunov stable at the origin and such that the following L^2-gain condition is satisfied:

$$\begin{aligned}\int_0^T \left(g(y(t)) + \|\alpha_x[w](t)\|^2\right) dt \quad &\leq \gamma^2 \int_0^T \|w(t)\|^2 dt + K(x) \\ &\text{for all } T > 0,\ w \in L^2_{\mathrm{loc}}(0, \infty; W).\end{aligned} \tag{1.5}$$

In (1.5) we want $K\colon \overline{D(A)} \to [0, +\infty]$, $\mathrm{dom}(K) \equiv \{x\colon\ K(x) \text{ is finite}\} = \overline{\mathrm{dom}(g)} =: \mathcal{K}$, and we require that $K(0) = 0$. The terminology strategy comes from the theory of differential games, in particular from Elliott–Kalton [13] and the references therein. We will denote by Δ the set of strategies, that is, functionals $\alpha\colon L^2_{\mathrm{loc}}(0, +\infty; W) \to L^2_{\mathrm{loc}}(0, +\infty; U)$ characterized by causality, i.e., such that

$$\begin{aligned}w_1 = w_2 \ \text{ almost everywhere (a.e.) in } [0, T],\ T > 0,\ \text{implies} \\ \alpha[w_1] = \alpha[w_2] \ \text{ a.e. in } [0, T].\end{aligned}$$

We note that one can give an equivalent formulation of (1.5) which involves only $w \in L^2(0, \infty; W)$. This is an easy exercise that we leave to the reader. A discontinuous running cost g is often useful in practice, and allowing g to be extended real valued is motivated by state constraints, as we discuss at the beginning of the next section.

The reader will recognize that our definition of the $\mathcal{H}_\infty$ suboptimal problem is nonstandard. The classical one requires that the family of strategies $\{\alpha_x\}$ can be constructed with a feedback control. We need to relax the definition in order to pursue the equivalence between the solvability of the problem and of the HJI equation which is

$$\langle Ax, DV_\gamma\rangle + H(x, DV_\gamma) = g(x) \quad \text{in } \overline{D(A)} \tag{1.6}$$

with Hamiltonian (note that the so-called Isaacs condition holds)

$$\begin{aligned}H(x,p) \ &= \inf_{w\in W} \sup_{u\in U} \left\{-\langle f(x) + Bu + Cw, p\rangle - \|u\|^2 + \gamma^2\|w\|^2\right\} \\ &= \sup_{u\in U} \inf_{w\in W} \left\{-\langle f(x) + Bu + Cw, p\rangle - \|u\|^2 + \gamma^2\|w\|^2\right\} \\ &= -\langle f(x), p\rangle + \frac{1}{4}\|B^*p\|^2 - \frac{1}{4\gamma^2}\|C^*p\|^2.\end{aligned} \tag{1.7}$$

In general, equations like (1.6) do not have smooth (super)solutions, and this is the main problem that arises when studying nonlinear systems. One therefore first needs to interpret solutions of (1.6) in an appropriate weak sense, and we will use the Crandall–Lions theory of viscosity solutions (see [9]). Moreover, nonsmooth solutions make it very difficult to implement the standard *recipe* for the feedback solution of the $\mathcal{H}_\infty$ problem for affine systems, in this case, $u(x) = -\frac{1}{2}B^*DV(x)$. In fact, a continuous feedback solution may not exist in general for a nonlinear system, even though (1.6) has a (nonsmooth) viscosity solution and the $\mathcal{H}_\infty$ problem could be solvable in our weaker sense. This fact gives us the reason for relaxing the definition of the $\mathcal{H}_\infty$ problem. Note that solving the problem in our sense is a necessary condition in order to be able to design a feedback solution; see the statement of Theorem 2.5 below. How far we can go in improving the solution that we characterize in this paper in the quest for a feedback solution of the problem is still an open and intriguing question; see, however, Theorem 2.5 and Corollary 2.6 for possible results in this direction.

In order to achieve our goal, we proceed as follows. We show that checking the gain condition (1.5) is equivalent to checking whether the value function of a differential game

$$(1.8) \qquad V_\gamma(x) = \inf_{\alpha\in\Delta} \sup_{w\in L^2_{\mathrm{loc}}(0,+\infty;W)} \sup_{T\geq 0} \int_0^T \left(g(y(t)) + \|\alpha[w](t)\|^2 - \gamma^2\|w(t)\|^2\right) dt$$

is finite for all $x \in \overline{\mathrm{dom}(g)}$ and vanishes at the origin. Note that in our assumptions V_γ will be discontinuous in general. Note also that since $g \geq 0$, choosing $w(\cdot) \equiv 0 \in L^2(0,\infty;W)$ shows that $V_\gamma \geq 0$; thus $V_\gamma\colon \overline{D(A)} \to [0,+\infty]$. Checking when V_γ is finite is the crucial step of the problem. When the gain condition is obtained, the existence of a solution of the $\mathcal{H}_\infty$ problem is a rather easy consequence of the continuity of V_γ at the origin. Following the ideas of the dynamic programming approach for differential games, the value function V_γ is expected to solve the HJI equation (1.6). (The use of strategies here is essential to implement this approach without smoothness of V_γ.) This fact allows us to characterize the finiteness of the value function V_γ by means of a partial differential equation. We will show that indeed V_γ is a viscosity solution of the HJI equation and it is, in fact, the minimal nonnegative lower semicontinuous supersolution. These facts bear the following consequences: (1) (Theorem 2.3) the L^2-gain condition (1.5) is equivalent to the existence of a finite nonnegative (super)solution, vanishing at the origin, of the HJI equation (1.6); (2) (Theorem 2.4) the existence of a solution of the $\mathcal{H}_\infty$ problem is equivalent to the existence of a positive definite (super)solution of (1.6), vanishing and *continuous* at the origin.

Our results generalize well-known statements holding in the finite dimensional case and for linear systems although, as far as we know, this is the first paper introducing state constraints in the problem; see the beginning of section 2. We extend the current literature even for a linear A in that we never suppose the term g to be continuous. To give a flavor of the difficulties involved, note that solutions of (1.6) are never unique, so classical ideas of viscosity solutions do not apply directly. In general, (1.6) might even have multiple solutions vanishing at the origin. We will use instead versions of the so-called *optimality principles* for solutions of HJI equations without uniqueness that we develop here in an infinite dimensional differential games framework for the first time. Our proofs follow the lines of those in Soravia [23] and [25], where analogous results were proved in the finite dimensional case, but with several additional difficulties, due to the infinite dimensional setting and the unbounded term in the dynamics. Some ideas to deal with state constraints in the infinite dimensional setting were developed in [17] for control systems. In [18] we already studied dissipative systems and the way they are related to the $\mathcal{H}_\infty$ problem. That case turns out to be technically much easier and corresponds to checking the L^2-gain condition (1.5) with a given predetermined Lipschitz continuous feedback $u = u(x)$ in (1.1). Those results can apply here only when (1.6) has a smooth supersolution, and, as a result, the differential games framework is never needed. In general, one should first determine whether the problem is solvable in our sense by solving (1.6) in the viscosity sense, and next try to improve the result and build a (possibly discontinuous) feedback solution.

We recall that the infinite dimensional case of the $\mathcal{H}_\infty$-control problem was previously studied for linear systems with distributed parameters by Van Keulen–Peters–Curtain [29] and Van Keulen [30], in the full information and partial information settings, respectively. Linear abstract systems under assumptions containing boundary control problems were considered by Barbu [5]. The case of nonlinear systems, but with a linear unbounded operator A, was studied by Barbu [4], [6]. The $\mathcal{H}_\infty$ problem in finite dimensions traces back to Zames [33] for linear systems; see also [15], [22], and [12]. For nonlinear-affine systems and smooth solutions of the HJI equation the problem was studied by van der Schaft [28] and Ball–Helton–Walker [1]. For general nonlinear systems and the approach using differential games and viscosity solutions, we refer the reader to the papers by one of the authors [24] and [23]. A general reference to the differential games approach is the book by Basar–Bernhard [7]. The theory of viscosity solutions for HJI equations with nonlinear unbounded terms, see section 2 for the definition, has been developed in works by Tataru [26], [27], Crandall–Lions [10], and Kocan–Soravia [17]. We do emphasize that this theory is quite recent, and to us it seems interesting by itself that the details of a technically complicated problem, such as the $\mathcal{H}_\infty$ suboptimal control problem, can be carried through.

The plan of the paper is as follows. In section 2 we present the main statements and a motivating example. The rest of the paper is devoted to the proofs. In section 3 we study the regularity of the value function V_γ and its relation with the HJI equation (1.6). Section 4 contains technical steps needed in the other proofs: we describe a change of variables for (1.6) which is implemented through an auxiliary problem. Eventually, in section 5 we complete the proofs of the main results.

In this paper $(S(\cdot))_{t\geq 0}$ indicates the nonlinear semigroup generated by $-A$. We denote by $Lip(\Omega)$ the space of all Lipschitz continuous functions on Ω, and, for $\psi \in Lip(\Omega)$, $L(\psi)$ will denote its best Lipschitz constant. If L is a Lipschitz constant for ψ, we say that ψ is L-Lipschitz continuous. We also indicate with P the projection of H onto $\overline{D(A)}$ and, for any normed space X, $B_r^X(x)$ stands for the closed ball in X of radius r centered at x. $USC(X)$ and $LSC(X)$ will stand for the upper and lower semicontinuous extended real-valued functions on X, respectively.

2. Main results, viscosity solutions. We start this section discussing the role of a lower semicontinuous and extended real-valued cost function g. This generality allows us the possibility of introducing state constraints to the system. Note for this purpose that when the left-hand side of (1.5) is finite, then necessarily $g(y(t))$ is finite for a.e. $t \geq 0$, and then

$$y(t) \in \mathcal{K} = \overline{\mathrm{dom}(g)} \quad \text{for all } t \geq 0.$$

Therefore, $\overline{\mathrm{dom}(g)}$ acts as a state constraint on the system. This fact is particularly helpful in the applications, especially in the infinite dimensional setting, where it is even useful to allow $\mathcal{K}$ to have an empty interior relative to $\overline{D(A)}$; see the examples in the paper by the authors [17].

Now we discuss the main assumption on the system. We will impose the following rather natural stability assumption on (1.1).

(2.1)

$$\begin{array}{l}\text{If } u_n \rightharpoonup u \text{ weakly in } L^2(0,T;U) \text{ for some } T>0, \text{ then for every } x \in \overline{D(A)},\\ t \in [0,T] \text{ and } w \in L^2(0,t;W),\ y(t,x,u_n,w) \to y(t,x,u,w) \text{ in } H.\end{array}$$

Remark 1. Condition (2.1), which is the strongest, although natural, assumption that we make on the system, is discussed in detail in Kocan–Soravia [17], and we refer the reader to that paper for additional references. For instance, it turns out that if $-A$ generates a compact semigroup and (1.2) and (1.3) hold, then (2.1) is satisfied; see Proposition 2.7 in [17]. All the statements we prove can also be shown by similar arguments if we assume the following condition instead of (2.1).

$$\left\{\begin{array}{l}\text{If } u_n \rightharpoonup u \text{ weakly in } L^2(0,T;U) \text{ for some } T>0, \text{ then for all } x, x_n \in \overline{D(A)},\\ x_n \rightharpoonup x,\ t \in [0,T] \text{ and } w \in L^2(0,t;W),\ y(t,x_n,u_n,w) \rightharpoonup y(t,x,u,w) \text{ in } H.\end{array}\right.$$

In this case the function g should be also assumed to be weakly lower semicontinuous, and then the weak lower semicontinuity of the value function can be achieved in the statements below. If $-A$ is the linear generator of a semigroup and $f \equiv 0$, then it is well known that the system (1.1) satisfies such a condition. We will not explicitly include this part in the statements for better clarity of the exposition.

We now recall the concept of stability that we require on our system.

DEFINITION 2.1. *We say that the undisturbed system*

$$y' + Ay \ni f(y) + Bu$$

is locally Lyapunov stable at 0, *if for every open neighborhood* $\mathcal{U}$ *of* 0 *in* H *we can find* $\delta > 0$ *such that for* $\|x\| \le \delta$, $x \in \overline{D(A)}$, *there is* $u \in L^2_{\mathrm{loc}}(0,+\infty;U)$ *such that* $y(t,x,u,0) \in \overline{D(A)} \cap \mathcal{U}$ *for all* $t \ge 0$. *If, in addition,*

$$\lim_{t\to+\infty} y(t,x,u,0) = 0,$$

then we say that the system is locally asymptotically stable at 0.

We will also use the following definition below.

DEFINITION 2.2. *We say that a function* $U\colon \overline{D(A)} \to [0,+\infty]$ *is locally positive definite at* 0 *if* $U(0)=0$ *and there is* $\sigma > 0$ *and a continuous, nondecreasing function* $\omega\colon [0,+\infty) \to [0,+\infty)$ *such that* $U(x) \ge \omega(\|x\|)$ *for* $x \in \overline{D(A)}$, $\|x\| \le \sigma$, *and* $\omega(r) = 0$ *if and only if* $r = 0$.

We now state our main results. The precise notion of viscosity solution that we use will be explained at the end of the section. Note that we allow solutions to be extended real valued.

THEOREM 2.3 (characterization of L^2-gain). *Assume that* (1.2), (1.3), (1.4) *and* (2.1) *hold. Then there is a family of strategies* $\{\alpha_x\colon\ x \in \overline{D(A)}\}$ *and a function* $K\colon \overline{D(A)} \to [0,+\infty]$ *satisfying* (1.5), $K(0)=0$, *and* $\overline{\mathrm{dom}(g)} \subset \mathrm{dom}(K)$ *if and only if the HJI equation* (1.6), *where* H *is given in* (1.7), *has a viscosity supersolution* $U\colon \overline{D(A)} \to [0,+\infty]$ *satisfying* $\mathrm{dom}(U) \supset \overline{\mathrm{dom}(g)}$ *and* $U(0)=0$. *In this case, a* *function with these properties is* V_γ, *defined in* (1.8), *and the choice* $K = V_\gamma$ *also gives the best possible estimate in* (1.5) *when choosing a family of optimal strategies for* V_γ.

If, moreover, there is $M > 0$ *such that*

$$V_\gamma(x),\ g(x) \leq M(1 + \|x\|^2), \quad x \in \overline{D(A)},$$

then ${V_\gamma}^*$ *is also a viscosity subsolution of the HJI equation* (1.6) *in* $\{x \in \overline{D(A)}\colon\ {V_\gamma}^*(x) > 0\}$.

The notation ${V_\gamma}^*$ in the statement above refers to the upper semicontinuous envelope of V_γ which is recalled below in (2.8). We recall that ${V_\gamma}^* = V_\gamma$ when the latter is upper semicontinuous.

Theorem 2.4 (characterization of the $\mathcal{H}_\infty$ suboptimal problem). *Assume that* (1.2), (1.3), (1.4), *and* (2.1) *hold. If there is a nonnegative viscosity supersolution* $U\colon \overline{D(A)} \to [0, +\infty]$ *of* (1.6), *locally positive definite at* 0 *and continuous at* 0, *then there exist a family of strategies* $\{\alpha_x\}$ *and a function* $K\colon \overline{D(A)} \to [0, +\infty]$ *continuous at* 0, $K(0) = 0$, *satisfying* (1.5) *and such that the family of controls* $\{\alpha_x[0]\}$ *provides local Lyapunov stability of the undisturbed system at the origin. The system is, moreover, locally asymptotically stable at* 0 *if* g *is locally positive definite at* 0.

If V_γ *is locally positive definite at* 0, *the previous condition is also necessary.*

Sometimes one can really construct smooth supersolutions of the HJI equation (1.6). Then the sufficiency part of Theorem 2.4 can be used to construct in the usual way a feedback solution of the $\mathcal{H}_\infty$ suboptimal control problem. Below we give a specific example where it can be applied. Suppose that $U\colon \overline{D(A)} \to \mathbb{R}$ is a C^1 function, meaning that U is continuously Fréchet differentiable on some open set $\Omega \subset H$ containing $\overline{D(A)}$. As usual, we think of $DU(x)$—the Fréchet differential of U at x—as an element of H, thus identifying H with its dual. We will say that such U is a *classical* supersolution of (1.6) if for every $(x, y) \in A$, i.e., $y \in Ax$, we have

$$\langle y, DU(x)\rangle + H(x, DU(x)) \geq g(x).$$

Classical supersolutions are viscosity supersolutions; see Proposition 6.3 in [19]. Define the feedback control $u(x)\colon = -\frac{1}{2}B^* DU(x)$, and consider the closed loop system

$$y'(t) + Ay(t) \ni f(y(t)) - \frac{1}{2}BB^* DU(y(t)) + Cw(t), \quad y(0) = x. \tag{2.2}$$

Below we will denote by $y(\cdot) = y(\cdot, x, w)$ its mild solution.

Theorem 2.5 (feedback solution). *Assume that* (1.2), (1.3), *and* (1.4) *hold. Suppose that* $U\colon \overline{D(A)} \to \mathbb{R}$ *is a* C^1, *nonnegative, and locally positive definite at* 0 *classical supersolution of* (1.6) *and that* $DU(\cdot)$ *is Lipschitz continuous. Then* $u = u(x)$, *as above, is a feedback solution of the* $\mathcal{H}_\infty$ *suboptimal problem and the family of strategies defined by the position*

$$\alpha_x[w](t) = u(y(t, x, w)), \quad t \geq 0, \tag{2.3}$$

solves the $\mathcal{H}_\infty$ *suboptimal problem in the sense of the definition in the introduction.*

The proof of Theorem 2.5 can be obtained in a straightforward way by observing that U in the statement is a classical supersolution of

$$\begin{aligned}\langle Ax, DU(x)\rangle + \inf\nolimits_{w \in W} \left\{-\langle f(x) + Cw - Bu(x), DU(x)\rangle + \gamma^2 \|w\|^2\right\}\\ \geq \|u(x)\|^2 + g(x) \quad \text{in } \overline{D(A)},\end{aligned}$$

and by using Theorem 2.4 in the case where the control set U is a singleton. We will leave the easy details to the reader.

We pause a little to discuss an example where our framework applies.

Example 1. We consider a simplified version of the one phase Stefan problem, the well-known model for melting ice, see, e.g., Barbu [3]. Let $\Omega \subset \mathbb{R}^N$ be a bounded, open, and smooth set, $N \geq 2$. We want to study the $\mathcal{H}_\infty$ suboptimal problem for the following system, $(t,x) \in (0,+\infty) \times \Omega$,

$$\left\{ \begin{array}{ll} y_t(t,x) - \Delta y(t,x) & \geq f(y(t,x)) + Bu(t) + Cw(t), \\ y_t(t,x) - \Delta y(t,x) & = f(y(t,x)) + Bu(t) + Cw(t) \text{ if } \ y(t,x) > 0, \\ y(t,x) & \geq 0, \quad (t,x) \in (0,+\infty) \times \Omega, \\ \alpha y(t,x) + \beta \frac{\partial y}{\partial n}(t,x) & = 0, \quad (t,x) \in (0,+\infty) \times \partial\Omega, \\ y(0,x) & = y_0(x), \quad x \in \Omega, \end{array} \right. \tag{2.4}$$

where $\alpha,\ \beta \geq 0$ and $\alpha + \beta > 0$. Here we choose as state space $H = L^2(\Omega)$. We can model $-\Delta$ and the boundary conditions with a linear, bounded operator $\tilde{A} \in \mathcal{L}(V, V')$, where $V = H^1(\Omega)$ (or $V = H^1_0(\Omega)$ if $\beta = 0$) which, at least in the case $\beta \neq 0$, is

$$\langle \tilde{A}y, z\rangle = \int_\Omega Dy \cdot Dz \, dx + \frac{\alpha}{\beta} \int_{\partial\Omega} y \cdot z \, d\sigma.$$

To model the variational inequality

$$\min\{y_t - \Delta y - f(y) - Bu - Cw, y\} = 0,$$

which is equivalent to (2.4) except for the boundary conditions, let $\varphi\colon\ H \to \mathbb{R} \cup \{+\infty\}$ be the indicator function of the set $\{y \in V\colon\ \ y \geq 0\}$, i.e.,

$$\varphi(y) = \left\{ \begin{array}{ll} 0, & y \in V,\ y \geq 0, \\ +\infty, & \text{elsewhere}. \end{array} \right.$$

Then define $A = (\tilde{A} + \partial\varphi) \cap (H \times H)$, where $\partial\varphi$ is the subgradient of the convex function φ in the sense of convex analysis; see Rockafellar [31]. It is well known (see Barbu [3]) that then $A = \partial\Phi$, where

$$\Phi(y) = \left\{ \begin{array}{ll} \frac{1}{2}\langle \tilde{A}y, y\rangle, & y \in V,\ y \geq 0, \\ +\infty, & \text{elsewhere}. \end{array} \right.$$

Note that Φ is convex with nontrivial domain, and therefore A is maximal monotone, nonlinear, and multivalued. Moreover, it can be shown that Φ is of compact type, and hence $-A$ generates a compact semigroup. In particular, the assumption (2.1) is satisfied by our results in [17]. System (2.4) is now written in the form (1.1).

For the system modeled in the previous example, let us now consider the running cost g with at most quadratic growth, i.e., $g(x) \leq M\|x\|^2$ for $x \in H$ and some $M > 0$. (We do not add extra state constraints and will only apply Theorem 2.5 for simplicity.) We will suppose that $f(0) = 0$ and show that, at least for sufficiently large values of the disturbance attenuation level γ, (1.6) has a classical supersolution of the form $V(x) = K\|x\|^2$, with a suitable constant $K > 0$. Since A is monotone and $(0,0) \in A$, we have $\langle y, x\rangle \geq 0$ for every $(x,y) \in A$, and thus to verify that V is a classical supersolution of (1.6) it is sufficient to prove the pointwise relation

$$-\langle f(x), DV(x)\rangle + \frac{1}{4}\|B^* DV(x)\|^2 - \frac{1}{4\gamma^2}\|C^* DV(x)\|^2 \geq g(x)$$

for $x \in \overline{D(A)}$. Easy computations, using the fact that f is L-Lipschitz continuous and $f(0) = 0$, show that it is enough to find $K > 0$ satisfying

$$K^2 \left(\inf \left\{ \|B^* x\|^2 \colon \ \|x\| = 1 \right\} - \frac{1}{\gamma^2} \|C^*\|^2 \right) - 2LK - M \geq 0.$$

It is clear that we will succeed for any choice of γ such that

$$\gamma \inf \left\{ \|B^* x\| \colon \ \|x\| = 1 \right\} > \|C^*\|. \tag{2.5}$$

By applying Theorem 2.5 we then get the following result.

Corollary 2.6. *Assume that* (1.2)*,* (1.3)*, and* (1.4) *hold, and that* $f(0) = 0$*. Let* g *have at most quadratic growth. If* $\inf\{\|B^* x\| \colon \ \|x\| = 1\} > 0$*, then for any disturbance attenuation level* γ *satisfying* (2.5) *we can find a constant* $K > 0$ *(which can be explicitly obtained from the computations above) such that the linear feedback law* $u(x) = -KB^* x$ *solves the* $\mathcal{H}_\infty$ *suboptimal problem for the nonlinear system* (2.4)*.*

We resume to say what our plan is for the proof of the main statements. Theorems 2.3 and 2.4 are eventually proved in section 5 and will be achieved through the following series of statements that also have independent interest. The first one refers to the regularity of the value function V_γ. It is proved in section 3.

Proposition 2.7. *Assume that* (1.2), (1.3), (1.4)*, and* (2.1) *hold. Then the value function* V_γ *in* (1.8) *has optimal strategies at any point. In particular,* (1.5) *holds at* $x \in \overline{\mathrm{dom}(g)}$ *for some* $\alpha \in \Delta$ *and* $K(x) < +\infty$ *if and only if* $V_\gamma(x) < +\infty$*. Thus, the* L^2*-gain condition* (1.5) *with* $K(0) = 0$ *can be satisfied if and only if* V_γ *is finite on* $\overline{\mathrm{dom}(g)}$ *and* $V_\gamma(0) = 0$*. Moreover,* V_γ *is lower semicontinuous.*

The next statement concerns the relationships between V_γ and (1.6). It is proved at the end of section 3.

Proposition 2.8. *Assume that* (1.2), (1.3), (1.4)*, and* (2.1) *hold. Then* V_γ *is a viscosity supersolution of* (1.6)*. If, moreover, there is* $M \geq 0$ *such that*

$$V_\gamma(x),\, g(x) \leq M(1 + \|x\|^2) \quad \text{for all} \ \ x \in \overline{D(A)}, \tag{2.6}$$

then ${V_\gamma}^*$ *is a viscosity subsolution of* (1.6) *in* $\{x \in \overline{D(A)} \colon \ {V_\gamma}^*(x) > 0\}$*.*

The next proposition states a representation formula for viscosity solutions (optimality principle). This is the main step of the proof and is proved in section 5.

Proposition 2.9. *Assume that* (1.2), (1.3), (1.4)*, and* (2.1) *hold and suppose that the function* $U \colon \ \overline{D(A)} \to \mathbb{R} \cup \{+\infty\}$ *is a lower semicontinuous, bounded-from-below viscosity supersolution of* (1.6)*. Then for every* $x \in \mathrm{dom}(U)$ *we have*

$$\begin{aligned} U(x) \quad &= \inf_{\alpha \in \Delta} \sup_{w \in L^2_{\mathrm{loc}}(0,+\infty;W)} \sup_{T \geq 0} \left\{ \int_0^T \left(g(y(t)) + \|\alpha[w](t)\|^2 \right) dt \right. \\ &\quad \left. - \int_0^T \gamma^2 \|w(t)\|^2 dt + U(y(T)) \right\}, \end{aligned} \tag{2.7}$$

where $y(\cdot) = y(\cdot, x, \alpha[w], w)$*. In particular, if* U *is nonnegative, then* $U \geq V_\gamma$ *on* $\overline{D(A)}$*.*

We conclude this section with the definition of viscosity solution for first order, nonlinear Hamilton–Jacobi equations with unbounded, nonlinear terms. For a function $\Phi \colon \ \overline{D(A)} \to \mathbb{R}$ and $\hat{x} \in \overline{D(A)}$, let

$$D_A^- \Phi(\hat{x}) = \liminf_{\substack{x \to \hat{x} \\ h \downarrow 0}} \frac{\Phi(x) - \Phi\left(S(h)x\right)}{h}, \quad D_A^+ \Phi(\hat{x}) = \limsup_{\substack{x \to \hat{x} \\ h \downarrow 0}} \frac{\Phi(x) - \Phi\left(S(h)x\right)}{h}.$$

We refer the reader to [10] and [19] for the basic properties of such operators. Given an extended real-valued function u on $\overline{D(A)}$, we denote by u^* and u_* its upper and lower semicontinuous envelopes, respectively, i.e., for $x \in \overline{D(A)}$

$$u^*(x) = \limsup_{\overline{D(A)} \ni y \to x} u(y), \quad u_*(x) = \liminf_{\overline{D(A)} \ni y \to x} u(y). \tag{2.8}$$

Next, we introduce the *test functions* used to define viscosity solutions.

Definition 2.10. *We will say that* $\Phi = \varphi + \psi \in C^1(H) + Lip(H)$ *is a subtest (supertest, resp.) function if*

$$\varphi(Px) \le \ (\ge,\ \textit{resp.})\ \varphi(x) \qquad \textit{and} \ \ \psi(Px) \le \ (\ge)\ \psi(x) \quad \textit{for} \ \ x \in H.$$

We are now ready to define solutions of the equation (1.6).

Definition 2.11. *Assume that* (1.2), (1.3), *and* (1.4) *hold. A possibly extended real-valued function* $U \in USC(\overline{D(A)})$ $(U \in LSC(\overline{D(A)})$, *resp.) is a viscosity subsolution (resp., supersolution) of* (1.6) *if for every subtest (resp., supertest) function* $\Phi = \varphi + \psi \in C^1(H) + Lip(H)$ *and local maximum (resp., minimum)* $\hat{x} \in \mathrm{dom}(U)$ *of* $U - \Phi$ *relative to* $\overline{D(A)}$ *we have*

$$\begin{aligned} D_A^- \Phi(\hat{x}) \quad & + \inf_{w \in W} \sup_{u \in U} \{ -\langle f(\hat{x}) + Bu + Cw, D\varphi(\hat{x}) \rangle \\ & -L(\psi)(\|f(\hat{x})\| + \|Bu\| + \|Cw\|) - \|u\|^2 + \gamma^2 \|w\|^2 \} \le g^*(\hat{x}) \\ \Big(D_A^+ \Phi(\hat{x}) \quad & + \inf_{w \in W} \sup_{u \in U} \{ -\langle f(\hat{x}) + Bu + Cw, D\varphi(\hat{x}) \rangle \\ & +L(\psi)(\|f(\hat{x})\| + \|Bu\| + \|Cw\|) - \|u\|^2 + \gamma^2 \|w\|^2 \} \ge g(\hat{x}), \text{resp.} \Big) \end{aligned}$$

A function U *(not necessarily continuous) defined on* $\overline{D(A)}$ *is a solution of* (1.6) *if* U^* *is a subsolution and* U_* *is a supersolution of* (1.6).

We refer the reader to [10] and to [17] for some more details and discussion of the notion of viscosity solution. Just recall after [17] that without loss of generality one can assume that the test function Φ (equivalently, φ) appearing in Definition 2.11 is globally Lipschitz continuous: $\Phi \in Lip(H)$.

3. Value function and the dynamic programming equation.

In this section we study the value function in (1.8) and prove its relations with the HJI equation (1.6).

First of all, we will obtain some useful estimates concerning the trajectories of the system. Gronwall's inequality immediately shows that for all $x, z \in \overline{D(A)}$, $u \in L^1_{\mathrm{loc}}(0, \infty; U)$, and $w \in L^1_{\mathrm{loc}}(0, \infty; W)$ we have

$$\|y(t, x, u, w) - y(t, z, u, w)\| \le \|x - z\| e^{L(f)\, t} \quad \text{for all} \ \ t \ge 0. \tag{3.1}$$

From (1.2), (1.3), and Hölder's inequality, for all $t \ge 0$ we also have

$$\begin{aligned} \|y(t) - x\| & \le \|S(t)x - x\| + \int_0^t \|f(y(s)) + Bu(s) + Cw(s)\| ds \\ & \le \|S(t)x - x\| + t\|f(x)\| \\ & +K \left(\sqrt{t} \|u\|_{L^2(0,t;U)} + \sqrt{t} \|w\|_{L^2(0,t;W)} + \int_0^t \|y(s) - x\| ds \right) \end{aligned} \tag{3.2}$$

with a constant $K = K(f, B, C)$. Then by Gronwall's inequality, for all $t \geq 0$ we obtain

$$\begin{aligned} \|y(t) - x\| &\leq K\Big(t\|f(x)\| + \sup_{0\leq s\leq t} \|S(s)x - x\| \\ &+\sqrt{t}\|u\|_{L^2(0,t;U)} + \sqrt{t}\|w\|_{L^2(0,t;W)}\Big)e^{Kt}, \end{aligned} \tag{3.3}$$

with a possibly new constant K. In particular, we conclude that for every $x \in \overline{D(A)}$

$$\begin{aligned} &y(t, x, u, w) \to x \ \text{ as } t \downarrow 0, \quad \text{uniformly for } u \text{ and } w \ \text{ bounded in} \\ &L^2(0, t; U) \text{ and } L^2(0, t; W), \quad \text{respectively.} \end{aligned} \tag{3.4}$$

Computations similar to those in (3.2) show that, if f is L-Lipschitz continuous, then for every $t \geq 0$

$$\begin{aligned} \|y(t) - S(t)x\| &\leq \left(\left(L \sup_{0\leq s\leq t} \|S(s)x - x\| + \|f(x)\|\right) t \right. \\ &\left. +\|B\| \int_0^t \|u\| + \|C\| \int_0^t \|w\|\right) e^{Lt}, \end{aligned} \tag{3.5}$$

where $y(\cdot) = y(\cdot, x, u, w)$.

We can now turn to study the value function. The next statement is the superoptimality part of the dynamic programming principle whose proof is standard. (For the suboptimality part, see the final statement in the proof of Proposition 2.8.) In the following for $t > 0$, $x \in \overline{D(A)}$, $u \in L^2(0, t; U)$, and $w \in L^2(0, t; W)$, we put

$$J(t, x, u, w) = \int_0^t \left(g(y(s, x, u, w)) + \|u(s)\|^2 - \gamma^2\|w(s)\|^2\right) ds.$$

PROPOSITION 3.1. *For any $x \in \overline{D(A)}$ and $t \geq 0$*

$$V_\gamma(x) \geq \inf_{\alpha\in\Delta} \sup_{w\in L^2(0,t;W)} \{J(t, x, \alpha[w], w) + V_\gamma(y(t, x, \alpha[w], w))\}. \tag{3.6}$$

COROLLARY 3.2. *Assume* (1.2), (1.3), *and* (1.4), *and suppose that $V_\gamma(x) < +\infty$ and $t \geq 0$. Then for every $w^* \in W$ and $M > \sqrt{V_\gamma(x)}$*

$$V_\gamma(x) \geq \inf \left\{J(t, x, u, w^*) + V_\gamma(y(t, x, u, w^*)) \colon \ u \in B^{L^2(0,t;U)}_{M+\gamma\sqrt{t}\|w^*\|}(0)\right\}.$$

Proof. Fix $x \in \overline{D(A)}$, $w^* \in W$, $t > 0$, and let M be as above. From Proposition 3.1

$$\begin{aligned} V_\gamma(x) &\geq \inf_{\alpha\in\Delta} \{J(t, x, \alpha[w^*], w^*) + V_\gamma(y(t, x, \alpha[w^*], w^*))\} \\ &\geq \inf_{u\in L^2(0,t;U)} \{J(t, x, u, w^*) + V_\gamma(y(t, x, u, w^*))\}. \end{aligned}$$

Suppose that $\|u\|_{L^2(0,t;U)} > \gamma\sqrt{t}\|w^*\| + M$. Since $V_\gamma, g \geq 0$, then

$$J(t, x, u, w^*) + V_\gamma(y(t, x, u, w^*)) \geq \int_0^t \|u(s)\|^2 \, ds - \gamma^2 t\|w^*\|^2 > M^2,$$

and hence the infimum of the left-hand side over all $\|u\|_{L^2(0,t;U)} > \gamma\sqrt{t}\|w^*\| + M$ is at least $M^2 > V_\gamma(x)$. □

To prove that the value function is lower semicontinuous, we will use the following general lemma that constructs optimal strategies in a variety of situations.

LEMMA 3.3. *Suppose that* $g, g_n, \varphi, \varphi_n \in LSC(\overline{D(A)})$, $(\varphi_n)_n$, $(g_n)_n$ *are nondecreasing sequences,* $\sup_n \varphi_n = \varphi$ *and* $\sup_n g_n = g$. *If* (2.1) *holds and*

$$M \equiv \sup_{n\geq 1} \inf_{\alpha\in\Delta} \sup_{w\in L^2_{\mathrm{loc}}} \sup_{t\geq 0} \left\{ \int_0^t \left(g_n(y(s)) + \|\alpha[w](s)\|^2 - \gamma^2\|w(s)\|^2 \right) ds + \varphi_n(y(t)) \right\},$$

where $x \in \overline{D(A)}$ *is fixed and* $y(\cdot) = y(\cdot, x, \alpha[w], w)$, *then there exists* $\alpha^\# \in \Delta$ *such that*

$$\begin{aligned} M = \sup_{w\in L^2_{\mathrm{loc}}} \sup_{t\geq 0} \quad & \left\{ \int_0^t \left(g(y(s,x,\alpha^\#[w],w)) + \|\alpha^\#[w](s)\|^2 \right) ds \right. \\ & \left. -\gamma^2 \int_0^t \|w(s)\|^2 ds + \varphi(y(t,x,\alpha^\#[w],w)) \right\}. \end{aligned} \tag{3.7}$$

Proof. (1) It is clear that M is not bigger than the right-hand side of (3.7) for any choice of $\alpha^\#$. To prove the reverse inequality, we may assume that $M < +\infty$. For every n there exists $\alpha_n \in \Delta$ such that for every $t \geq 0$ and $w \in L^2_{\mathrm{loc}}(0,\infty;W)$ denoting $y_n(\cdot) \equiv y(\cdot, x, \alpha_n[w], w)$, we have

$$\begin{aligned} M + \tfrac{1}{n} \quad & \geq \int_0^t \left(g_n(y_n(s)) + \|\alpha_n[w](s)\|^2 - \gamma^2\|w(s)\|^2 \right) ds + \varphi_n(y_n(t)) \\ & \geq \int_0^t \left(\|\alpha_n[w](s)\|^2 - \gamma^2\|w(s)\|^2 \right) ds. \end{aligned}$$

Hence for all $w \in L^2_{\mathrm{loc}}$ and $t \geq 0$ we have

$$\|\alpha_n[w]\|^2_{L^2(0,t;U)} \leq M + 1 + \gamma^2\|w\|^2_{L^2(0,t;W)}. \tag{3.8}$$

(2) For $T > 0$, every $\alpha \in \Delta$ can be viewed as an element of the product space

$$\left(L^2(0,T;U)\right)^{L^2(0,T;W)}.$$

From (3.8) we have that, up to an obvious identification, $\alpha_n \in \Pi_T$ for all T, n, where

$$\Pi_T = \{\alpha\colon\ L^2(0,T;W) \to L^2(0,T;U)\colon \|\alpha[w]\|^2_{L^2(0,T;U)} \leq M + 1 + \gamma^2\|w\|^2_{L^2(0,T;W)}\}.$$

Since bounded closed balls in $L^2(0,T;U)$ are weakly compact, it follows that Π_T equipped with the product topology is compact. Therefore, for $T = 1$, there is a subnet $\{\alpha^1_\lambda\}_{\lambda\in\Lambda_1}$ of $\{\alpha_n\}$ converging to some $\alpha^1 \in \Pi_1$. Given a positive integer k and a subnet $\{\alpha^k_\lambda\}_{\lambda\in\Lambda_k}$ of $\{\alpha_n\}$ converging to $\alpha^k \in \Pi_k$, there exists a subnet $\{\alpha^{k+1}_\lambda\}_{\lambda\in\Lambda_{k+1}}$ of $\{\alpha^k_\lambda\}_{\lambda\in\Lambda_k}$ converging to $\alpha^{k+1} \in \Pi_{k+1}$. Define $\alpha^\#\colon\ L^2_{\mathrm{loc}}(0,\infty;W) \to L^2_{\mathrm{loc}}(0,\infty;U)$ according to the rule

$$\alpha^\#[w](t) = \alpha^{k+1}[w](t), \quad k \leq t < k+1.$$

It is not difficult to see that nonanticipating strategies form a closed subset of any Π_T, and therefore $\alpha^\# \in \Delta$.

(3) We will finally show that $\alpha^\#$ has the desired property. Let $w \in L^2_{\mathrm{loc}}(0,\infty;W)$ and $t > 0$. By construction there is a subsequence n_k (possibly depending on w and t) such that

$$\alpha_{n_k}[w] \rightharpoonup \alpha^\#[w], \ \text{ weakly in } L^2(0,t;U), \text{ as } k \to \infty. \tag{3.9}$$

By the stability assumption (2.1) the trajectories $y_{n_k}(\cdot)$ converge pointwise to $y^\#(\cdot) \equiv y(\cdot, x, \alpha^\#[w], w)$. Using the lower semicontinuity of φ_n's, g_n's, and the norm, from Fatou's lemma, we deduce

$$\begin{aligned}\liminf_{k\to\infty} \quad & \int_0^t \left(g_{n_k}(y_{n_k}(s)) + \|\alpha_{n_k}[w](s)\|^2 - \gamma^2\|w(s)\|^2\right) ds + \varphi_{n_k}(y_{n_k}(t)) \\ & \geq \int_0^t \left(g(y^\#(s)) + \|\alpha^\#[w](s)\|^2 - \gamma^2\|w(s)\|^2\right) ds + \varphi(y^\#(t)).\end{aligned}$$

Therefore,

$$M \geq \int_0^t \left(g(y^\#(s)) + \|\alpha^\#[w](s)\|^2 - \gamma^2\|w(s)\|^2\right) ds + \varphi(y^\#(t))$$

for every $t \geq 0$ and $w \in L^2_{\text{loc}}(0, \infty; W)$, and hence the conclusion. □

Proof of Proposition 2.7. To prove that for all $x \in \overline{D(A)}$ there are optimal strategies for the value function $V_\gamma(x)$, we need just to apply Lemma 3.3 with $g_n \equiv g$ and $\varphi_n \equiv 0$. The statements concerning the L^2-gain condition then follow trivially. En-passant we note that, using the equivalent definition of V_γ with L^2 disturbances instead of L^2_{loc}, one could simplify this part of the proof by working directly with $L^2(0, \infty; W)$ and $L^2(0, \infty; U)$, which avoids the diagonal argument in the proof of Lemma 3.3.

In order to prove that the value function V_γ is lower semicontinuous, let $x_n \to x$, $x_n, x \in \overline{D(A)}$. Without loss of generality we may assume that $V_\gamma(x_n)$ converges to a finite limit, say, L. Let $\alpha_n \in \Delta$ be an optimal strategy for $V_\gamma(x_n)$, i.e., for every $t \geq 0$ and $w \in L^2(0, t; W)$

$$V_\gamma(x_n) \geq \int_0^t \left(g(y(s, x_n, \alpha_n[w], w)) + \|\alpha_n[w](s)\|^2 - \gamma^2\|w(s)\|^2\right) ds. \tag{3.10}$$

It follows that, if n is sufficiently big, for any $t \geq 0$ and $w \in L^2(0, t; W)$

$$\|\alpha_n[w]\|^2_{L^2(0,t;U)} \leq L + 1 + \gamma^2\|w\|^2_{L^2(0,t;W)}.$$

Arguing as in the proof of Lemma 3.3, there exists $\alpha^\# \in \Delta$ satisfying (3.9). Taking $\liminf$ in (3.10) as $n \to \infty$ as in Lemma 3.3 gives

$$L \geq \int_0^t \left(g(y(s, x, \alpha^\#[w], w)) + \|\alpha^\#[w](s)\|^2 - \gamma^2\|w(s)\|^2\right) ds$$

for every $t \geq 0$ and $w \in L^2(0, t; W)$, which yields $L \geq V_\gamma(x)$, completing the proof of Proposition 2.7. □

We end this section by proving that V_γ solves the HJI equation.

Proof of Proposition 2.8. We start by proving that V_γ is a viscosity supersolution. Recall that, by Proposition 2.7, V_γ is lower semicontinuous. Let $\Phi = \varphi + \psi$ be a Lipschitz continuous supertest function, and assume that $V_\gamma - \Phi$ attains a local minimum at $\hat{x} \in \text{dom}(V_\gamma)$. Note that it is conceivable that $g(\hat{x}) = +\infty$. If $D^+_A\Phi(\hat{x}) = +\infty$, there is nothing to prove. Otherwise, we argue by contradiction. Suppose that there are $\theta > 0$ and $w^* \in W$ such that for every $u \in U$

$$\begin{aligned}D^+_A\Phi(\hat{x}) - \langle f(\hat{x}) \quad & + Bu + Cw^*, D\varphi(\hat{x})\rangle + L(\psi)(\|f(\hat{x})\| + \|Bu\| + \|Cw^*\|) \\ & - \|u\|^2 + \gamma^2\|w^*\|^2 < g(\hat{x}) - 2\theta.\end{aligned} \tag{3.11}$$

First, we will show that

$$D_A^+\Phi(\hat{x}) > -\infty. \tag{3.12}$$

For $t > 0$ choose in the dynamic programming principle, Proposition 3.1, any t-optimal strategy α_t and denote $u_t = \alpha_t[0]$. Since $g \geq 0$, it follows that

$$V_\gamma(\hat{x}) + t \geq \int_0^t \|u_t(s)\|^2 ds + V_\gamma(y(t)), \tag{3.13}$$

where $y(\cdot) = y(\cdot, \hat{x}, u_t, 0)$. Since $V_\gamma \geq 0$, (3.13) yields a uniform bound on $\|u_t\|_{L^2(0,t;U)}$ and by (3.4) the assumption on Φ gives $V_\gamma(\hat{x}) - V_\gamma(y(t)) \leq \Phi(\hat{x}) - \Phi(y(t))$ for all sufficiently small t. From (3.13) and (3.5), for small t we then get

$$\begin{aligned}\Phi(S(t)\hat{x}) - \Phi(\hat{x}) \quad &\leq t - \int_0^t \|u_t\|^2 + \Phi(S(t)\hat{x}) - \Phi(y(t)) \\ &\leq t + L(\Phi)\|y(t) - S(t)\hat{x}\| - \int_0^t \|u_t\|^2 \leq Kt,\end{aligned}$$

where the last inequality also uses the Hölder inequality and an appropriate constant K. Hence $\Phi(S(t)\hat{x}) - \Phi(\hat{x}) \leq O(t)$ as $t \downarrow 0$, which yields (3.12).

Let $M = \sqrt{V_\gamma(\hat{x})} + 1$. Suppose that $\|u\|_{L^2(0,t;U)} \leq M$, $t > 0$, and for $s \in [0,t]$ put $y(s) = y(s, \hat{x}, u, w^*)$. From (3.4), (3.11), the coercivity in u of the left-hand side of (3.11), and the lower semicontinuity of g, it follows that, if t is sufficiently small, then for all $s \in [0,t]$

$$D_A^+\Phi(\hat{x}) - \langle v(s), D\varphi(y(s))\rangle + L(\psi)\|v(s)\| - \|u(s)\|^2 + \gamma^2\|w^*\|^2 < g(y(s)) - \theta,$$

where we denoted $v(s) = f(y(s)) + Bu(s) + Cw^*$. Integrating from 0 to t and using Tataru's result (see [27] and Corollary 4.8 in [19]) yields (note that it works just like formally except for the term $o(t)$)

$$t\theta \leq \int_0^t \left(g(y(s)) + \|u(s)\|^2 - \gamma^2\|w^*\|^2\right) ds + \Phi(y(t)) - \Phi(\hat{x}) + o(t),$$

as $t \to 0$, uniformly for all $\|u\|_{L^2(0,t;U)} \leq M$. By the local minimum property of Φ we obtain

$$t\theta \leq \int_0^t \left(g(y(s)) + \|u(s)\|^2 - \gamma^2\|w^*\|^2\right) ds + V_\gamma(y(t)) - V_\gamma(\hat{x}) + o(t)$$

for all $\|u\|_{L^2(0,t;U)} \leq M$. Then, from Corollary 3.2 (recall the choice of M), we get $t\theta \leq o(t)$ and a contradiction when choosing t sufficiently small.

To prove that, under the additional assumption (2.6), the function ${V_\gamma}^*$ is a subsolution of (1.6) in $\{x \in \overline{D(A)}\colon\ {V_\gamma}^*(x) > 0\}$, we can argue in a similar way as in section 4, where we prove the corresponding result for an auxiliary value function, which is needed in the most delicate step of the proof of Theorem 2.3. Specifically, the suboptimality part of the dynamic programming principle we need follows along the lines of the proof of Proposition 4.6, although adapting this proof requires the assumption (2.6). The fact that ${V_\gamma}^*$ solves the equation can then be adapted from the argument of the proof of Proposition 4.7; see also Proposition 2.4 in [17] for discontinuous value functions. We will therefore omit the details of this part. □

4. The auxiliary problem. Since our problem has unbounded controls and possibly a singular value function, we need to construct a suitable, more regular, auxiliary problem and then use the ideas of the theory of viscosity solutions to compare supersolutions of HJI equations and the corresponding value function. We start by introducing a change of variables. Define $\rho\colon\ \mathbb{R}\cup\{+\infty\}\to(0,\pi]$ as $\rho(t)=\frac{\pi}{2}+\tan^{-1}(t)$, $\rho(+\infty)=\pi$. Note that ρ is strictly increasing and 1-Lipschitz continuous on $\mathbb{R}$. Also, $\rho^{-1}\colon\ (0,\pi]\to\mathbb{R}\cup\{+\infty\}$ is given by $\rho^{-1}(s)=-\cot(s),\ s\in(0,\pi)$.

Let (1.2) and (1.3) hold and let $g\colon\ \overline{D(A)}\to[0,\infty)$ be continuous. Suppose that $U\in LSC(\overline{D(A)})$ is a bounded-from-below supersolution (in the sense of Definition 2.11) of (1.6). Define $W\colon\ \overline{D(A)}\times\mathbb{R}\to(0,\pi]$ according to

$$W(x,r)=\rho(U(x)+r). \tag{4.1}$$

Note that $U(x)=+\infty$ if and only if $W(x,r)=\pi$. Formal computations suggest that W ought to be a supersolution of

$$\begin{aligned}\langle Ax,D_xW\rangle\ &+\inf_{w\in W}\sup_{u\in U}\{-\langle f(x)+Bu+Cw,D_xW\rangle\\ &+\left(\gamma^2\|w\|^2-\|u\|^2-g(x)\right)D_rW\}=0.\end{aligned}\tag{4.2}$$

This is indeed the case if the solutions of (4.2) are meant in an appropriate viscosity sense, modeled on the one introduced in [16] and [21] (equations with separated variables) adapted to differential games in the spirit of Definition 2.11; see also [17]. Note that the Hamiltonian in (4.2) is not necessarily real valued unless $D_rW>0$.

Definition 4.1. *Let $\Omega\subseteq\overline{D(A)}\times\mathbb{R}$. A possibly extended real-valued function $W\in USC(\overline{D(A)}\times\mathbb{R})$ is a viscosity subsolution of* (4.2) *on Ω if for every subtest function Φ as in Definition* 2.10*, $\eta\in C^1(\mathbb{R})$, and local maximum $(\hat{x},\hat{r})\in\Omega\cap\mathrm{dom}(W)$ of $W(x,r)-\Phi(x)-\eta(r)$ relative to Ω, we have*

$$\begin{aligned}&D_A^-\Phi(\hat{x})+\inf_{w\in W}\sup_{u\in U}\{-\langle f(\hat{x})+Bu+Cw,D\varphi(\hat{x})\rangle\\ &-L(\psi)(\|f(\hat{x})\|+\|Bu\|+\|Cw\|)+\left(\gamma^2\|w\|^2-\|u\|^2-g(x)\right)\eta'(\hat{r})\}\le 0.\end{aligned}\tag{4.3}$$

Supersolutions and solutions are defined similarly, following Definition 2.11.

The precise result on the change of variables is the following. Its proof parallels that of Lemmas 3.8 and 3.9 in Kocan–Soravia [17], and we skip it.

Proposition 4.2. *Let $g\colon\ \overline{D(A)}\to[0,\infty)$ be bounded and continuous. Suppose that the function $U\colon\ \overline{D(A)}\to(-\infty,+\infty]$ is a bounded-from-below supersolution (in the sense of Definition* 2.11*) of* (1.6)*. Define $W\colon\ \overline{D(A)}\times\mathbb{R}\to(0,\pi]$ according to* (4.1)*. Then W is a supersolution of* (4.2) *on $\overline{D(A)}\times\mathbb{R}$ in the sense of Definition* 4.1.

At this point we study the auxiliary differential game that corresponds to the change of variables we introduced. Suppose that $g\colon\ \overline{D(A)}\to\mathbb{R}$ and $W\colon\ \overline{D(A)}\times\mathbb{R}\to(0,\infty)$ are bounded. For $\lambda>0$ we will consider an auxiliary value function

$$\begin{aligned}(V_{g,W}^\lambda\equiv)V^\lambda(x,r)\quad&=\inf\nolimits_{\alpha\in\Delta}\sup\nolimits_{w\in L^2_{\mathrm{loc}}(0,\infty;W)}\\ &\quad\sup\nolimits_{t\ge0}e^{-\lambda k(t,\alpha[w],w)}W\left(y(t,x,\alpha[w],w),r+r(t,x,\alpha[w],w)\right),\end{aligned}\tag{4.4}$$

where $(x,r)\in\overline{D(A)}\times\mathbb{R}$ and

$$\begin{aligned}k(t,u,w)\quad&=\int_0^t(1+\|u\|^2+\|w\|^2),\\ r(t,x,u,w)\quad&=\int_0^t\left(g(y(s,x,u,w))+\|u(s)\|^2-\gamma^2\|w(s)\|^2\right)ds.\end{aligned}$$

Choosing $t = 0$ in the definition (4.4) shows that $V^\lambda \geq W > 0$. Moreover, if $U\colon \overline{D(A)} \to \mathbb{R}$ and W is as in (4.1), then $0 < V^\lambda \leq \pi$. Note that for fixed $u \in L^2_{\mathrm{loc}}(0,\infty;U)$ and $w \in L^2_{\mathrm{loc}}(0,\infty;W)$, the function $k(\cdot,u,w)$ is strictly increasing from $[0,+\infty)$ to $[0,+\infty)$ and $t \leq k(t,u,w)$ for all u,w. We will denote by $t = t(\cdot,u,w)$ its inverse function, so, in particular, $k(t(\tau,u,w),u,w) = \tau$, $t(k(t,u,w),u,w) = t$ and $t(\tau,u,w) \leq \tau$; therefore, $t(\tau,u,w) \to 0$ as $\tau \to 0$, uniformly in $u \in L^2_{\mathrm{loc}}$ and $w \in L^2_{\mathrm{loc}}$. We will need to reparametrize the trajectories with the new parameter $\tau = \tau(t,u,w)$, which depends upon the trajectory, in order to get estimates which are uniform on the controls.

We start showing some estimates and regularity of the auxiliary value function.

Lemma 4.3. *Suppose that $g\colon \overline{D(A)} \to [0,\infty)$ and $U\colon \overline{D(A)} \to \mathbb{R}$ are bounded and let W be as in* (4.1). *Then $V^\lambda = V^\lambda_{g,W}$ satisfies*

$$\lim_{r\to-\infty} V^\lambda(x,r) = 0, \quad \textit{uniformly for all } x \in \overline{D(A)}. \tag{4.5}$$

Moreover, for every $M > 0$, $\sup\{V^\lambda(x,r)\colon\ x \in \overline{D(A)},\ r \leq M\} < \pi$.

Proof. Fix any $u^* \in U$. By definition (4.4), choosing $\alpha \equiv u^*$, we have

$$V^\lambda(x,r) \leq \sup_{t\geq 0} e^{-\lambda t}\rho\left(\|U\|_{L^\infty} + r + t(\|g\|_{L^\infty} + \|u^*\|^2)\right), \tag{4.6}$$

and evoking the elementary Lemma 4.4 below, this shows (4.5). If $r < M$, then from (4.6), $V^\lambda(x,r) \leq \rho\left(\|U\|_{L^\infty} + M + \|g\|_{L^\infty} + \|u^*\|^2\right) \vee \pi e^{-\lambda} < \pi$. □

Lemma 4.4. *Let $K,\lambda > 0$. For $r \in \mathbb{R}$ define $h(r) = \sup_{t\geq 0} e^{-\lambda t}\rho(Kt + r)$. Then* $\lim_{r\to-\infty} h(r) = 0$.

Proof. Clearly, $h \geq 0$. Let $\mu = \frac{\lambda}{K} > 0$; by changing variables it follows that $h(r) = e^{\mu r}\sup_{t\geq r} e^{-\mu t}\rho(t)$. Denoting $f(t) = e^{-\mu t}\rho(t)$, we have $f'(t) = e^{-\mu t}(\rho'(t) - \mu\rho(t))$, and since $\lim_{t\to-\infty}\frac{\rho'(t)}{\rho(t)} = 0$, there is t_0 such that $f' < 0$ on $(-\infty,t_0)$. Thus

$$\begin{aligned} h(r) &\leq e^{\mu r}\left(f(r) \vee \sup_{t\geq t_0} f(t)\right) \leq e^{\mu r}\left(e^{-\mu r}\rho(r) \vee \pi e^{-\mu t_0}\right) \\ &\leq \rho(r) \vee \pi e^{\mu(r-t_0)} \to 0 \quad \text{as} \quad r \to -\infty. \qquad \square \end{aligned}$$

Lemma 4.5. *Suppose that $g\colon \overline{D(A)} \to [0,\infty)$ and $U\colon \overline{D(A)} \to \mathbb{R}$ are bounded and Lipschitz continuous, let W be as in* (4.1) *and let $V^\lambda = V^\lambda_{g,W}$.*

(1) *Suppose that $(x_i,r_i) \in \overline{D(A)}\times\mathbb{R}$, $i = 1,2$, and $V^\lambda(x_1,r_1) > V^\lambda(x_2,r_2) > \delta > 0$. Then there exists $M = M(g,U,f,\lambda,\delta) > 0$ such that*

$$V^\lambda(x_1,r_1) - V^\lambda(x_2,r_2) \leq M\|x_1 - x_2\| + r_1 - r_2.$$

(2) *V^λ is Lipschitz continuous in a neighborhood of every point.*

(3) *Suppose that $(\hat{x},\hat{r}) \in \overline{D(A)}\times\mathbb{R}$ is such that $V^\lambda(\hat{x},\hat{r}) > \delta > 0$ and $\hat{r} < 1/\delta$. Then*

$$\liminf_{h\downarrow 0}\frac{V^\lambda(\hat{x},\hat{r}+h) - V^\lambda(\hat{x},\hat{r})}{h} \geq \beta > 0,$$

where $\beta = \beta(g,U,\lambda,\delta)$.

Proof. (1) By definition, for every $\sigma > 0$ there exists $\alpha_\sigma \in \Delta$ such that for all $t \geq 0$ and $w \in L^2(0,t;W)$ we have

$$e^{-\lambda k(t,\alpha_\sigma[w],w)}W\left(y(t,x_2,\alpha_\sigma[w],w), r_2 + r(t,x_2,\alpha_\sigma[w],w)\right) < V^\lambda(x_2,r_2) + \sigma.$$

Given α_σ, there are also $t_\sigma \geq 0$ and $w_\sigma \in L^2(0, t_\sigma; W)$ such that

$$\begin{aligned}&e^{-\lambda k(t_\sigma,\alpha_\sigma[w_\sigma],w_\sigma)} W\left(y(t_\sigma, x_1, \alpha_\sigma[w_\sigma], w_\sigma), r_1 + r(t_\sigma, x_1, \alpha_\sigma[w_\sigma], w_\sigma)\right)\\ &\quad > V^\lambda(x_1, r_1) - \sigma.\end{aligned}$$

Since ρ is bounded, it follows that for every $\sigma < \delta/2$ we have $k(t_\sigma, \alpha_\sigma[w_\sigma], w_\sigma) \leq T = T(\lambda, \delta)$. If for $i = 1, 2$ we denote $y_i(\cdot) = y(\cdot, x_i, \alpha_\sigma[w_\sigma], w_\sigma)$ and $r_i(\cdot) = r_i + r(\cdot, x_i, \alpha_\sigma[w_\sigma], w_\sigma)$, we also have

$$\begin{aligned}&V^\lambda(x_1, r_1) - V^\lambda(x_2, r_2) - 2\sigma\\ &\quad < e^{-\lambda k(t_\sigma,\alpha_\sigma[w_\sigma],w_\sigma)} \left(W(y_1(t_\sigma), r_1(t_\sigma)) - W(y_2(t_\sigma), r_2(t_\sigma))\right).\end{aligned} \tag{4.7}$$

If σ is sufficiently small, then the left-hand side in (4.7) is positive, and then, since ρ is increasing and 1-Lipschitz continuous, the right-hand side is at most $U(y_1(t_\sigma)) + r_1(t_\sigma) - U(y_2(t_\sigma)) - r_2(t_\sigma)$. Therefore, using (3.1), we compute

$$\begin{aligned}&V^\lambda(x_1, r_1) - V^\lambda(x_2, r_2) \leq L(U)\|x_1 - x_2\|e^{L(f)T} + r_1 - r_2\\ &+L(g)\int_0^{t_\sigma} \|y_1(s) - y_2(s)\|ds + 2\sigma\\ &\leq (L(U) + L(g)T)\, e^{L(f)T}\|x_1 - x_2\| + r_1 - r_2 + 2\sigma,\end{aligned}$$

uniformly for σ sufficiently small. Letting $\sigma \downarrow 0$ gives the result.

(2) This follows immediately from (1) since $V^\lambda \geq W > 0$ and W is continuous.

(3) To prove the lower bound on the derivative of the value function V^λ in the last variable we proceed as in the first part of the proof of (1). Let $\hat{x}, \hat{r}$ be given and $V^\lambda(\hat{x}, \hat{r}) \geq \delta > 0$. We fix $\epsilon > 0$ and for $\sigma = \epsilon h$, $r_1 = \hat{r}, r_2 = \hat{r} + h, h > 0$, $x_1 = x_2 = \hat{x}$, we find $\alpha_\sigma, w_\sigma, t_\sigma$ such that, putting $y(\cdot) \equiv y_1(\cdot) \equiv y_2(\cdot)$ and $r(\cdot) = r(\cdot, \hat{x}, \alpha_\sigma[w_\sigma], w_\sigma)$, we have an analogue of (4.7), namely,

$$\begin{aligned}V^\lambda(\hat{x}, \hat{r} + h) &- V^\lambda(\hat{x}, \hat{r}) + 2h\epsilon > e^{-\lambda k(t_\sigma,\alpha_\sigma[w_\sigma],w_\sigma)}(W(y(t_\sigma), \hat{r} + h + r(t_\sigma))\\ &-W(y(t_\sigma), \hat{r} + r(t_\sigma))).\end{aligned} \tag{4.8}$$

Note that, by construction, $\|\alpha_\sigma[w_\sigma]\|^2_{L^2(0,t_\sigma)} \leq k(t_\sigma, \alpha_\sigma[w_\sigma], w_\sigma) \leq T = T(\lambda, \delta)$ for all $\sigma \leq \delta/2$. This implies

$$U(y(t_\sigma)) + \hat{r} + h + r(t_\sigma) \leq \|U\|_\infty + \hat{r} + 1 + (1 + \|g\|_\infty)T$$

for h small enough. Moreover, since for $\sigma < \delta/2$

$$0 < \delta/2 \leq V^\lambda(\hat{x}, \hat{r}) - \sigma < e^{-\lambda k(t_\sigma,\alpha_\sigma[w_\sigma],w_\sigma)} W(y(t_\sigma), \hat{r} + r(t_\sigma)) \leq W(y(t_\sigma), \hat{r} + r(t_\sigma)),$$

we also get

$$U(y(t_\sigma)) + \hat{r} + r(t_\sigma) \geq \rho^{-1}(\delta/2).$$

Then there is $R = R(g, U, \lambda, \delta) > 0$ such that for all h sufficiently small we have

$$U(y(t_\sigma)) + \hat{r} + h + r(t_\sigma) \in [-R, R].$$

Observe that the right-hand side of (4.8) is positive, since ρ is increasing and $h > 0$, and we can now conclude by the mean value theorem that

$$\begin{aligned}&V^\lambda(\hat{x}, \hat{r} + h) - V^\lambda(\hat{x}, \hat{r}) + 2h\epsilon\\ &\quad > e^{-\lambda T}(\rho(U(y(t_\sigma)) + \hat{r} + h + r(t_\sigma)) - \rho(U(y(t_\sigma)) + \hat{r} + r(t_\sigma))) \geq Kh,\end{aligned}$$

where $K = e^{-\lambda T}/(1+R^2)$. Dividing then by h and letting $h \to 0$, we finally obtain

$$\liminf_{h\to 0} \frac{V^\lambda(\hat{x},\hat{r}+h) - V^\lambda(\hat{x},\hat{r})}{h} \geq K - 2\epsilon,$$

and the proof is completed by choosing ϵ sufficiently small. □

We will now obtain a dynamic programming principle for the auxiliary value function.

PROPOSITION 4.6. *Suppose that $g\colon \overline{D(A)} \to [0,\infty)$ and $U\colon \overline{D(A)} \to \mathbb{R}$ are continuous and bounded and let W be as in* (4.1). *Denote $V^\lambda = V^\lambda_{g,W}$ and let $u^* \in U$. If $(V^\lambda)^*(\hat{x},\hat{r}) > W(\hat{x},\hat{r})$, then there is $\epsilon > 0$ such that if $\|x-\hat{x}\|$, $|r-\hat{r}|$, $|V^\lambda(x,r) - (V^\lambda)^*(\hat{x},\hat{r})| < \epsilon$, $t \in (0,\epsilon)$ and $\tau \in (0,\epsilon)$, then*

(4.9)
$$V^\lambda(x,r) \leq e^{-\lambda\tau} \sup_{w\in L^2_{\text{loc}}(0,\infty;W)} V^\lambda\left(y(t(\tau,u^*,w),x,u^*,w), r + r(t(\tau,u^*,w),x,u^*,w)\right),$$

$$V^\lambda(x,r) \leq \sup\{e^{-\lambda k(t,u^*,w)} V^\lambda\left(y(t,x,u^*,w), r + r(t,x,u^*,w)\right) : \ \|w\|_{L^2(0,t;W)} \leq M\}.$$

Proof. (1) To prove (4.9) we argue by contradiction. Suppose that $x_n \to \hat{x}$, $r_n \to \hat{r}$, $\tau_n \downarrow 0$, $V^\lambda(x_n,r_n) \to (V^\lambda)^*(\hat{x},\hat{r})$, and $\epsilon_n > 0$ are such that

$$V^\lambda(x_n,r_n) - \epsilon_n > \sup_{w\in L^2_{\text{loc}}} e^{-\lambda\tau_n} V^\lambda\left(y_n(t_n), r_n(t_n)\right), \tag{4.10}$$

where $y_n(\cdot) = y(\cdot,x_n,u^*,w)$, $r_n(\cdot) = r_n + r(\cdot,x_n,u^*,w)$, and $t_n = t(\tau_n,u^*,w)$.

For a fixed n, first suppose that

$$e^{-\lambda s} W(y_n(t(s,u^*,w)), r_n(t(s,u^*,w))) \leq V^\lambda(x_n,r_n) - \epsilon_n \tag{4.11}$$

for all $s \in [0,\tau_n)$, $w \in L^2(0,t(s,u^*,w);W)$. By definition, for every $\alpha \in \Delta$ there exists $w_\alpha \in L^2_{\text{loc}}$ such that

$$V^\lambda(x_n,r_n) - \frac{\epsilon_n}{2} < \sup_{t\geq 0} e^{-\lambda k(t,\alpha[w_\alpha],w_\alpha)} W\left(y(t,x_n,\alpha[w_\alpha],w_\alpha), r_n + r(t,x_n,\alpha[w_\alpha],w_\alpha)\right).$$

Combining this with (4.11), it follows that for every $\alpha \in \Delta$ with the property that $\alpha[w](t) = u^*$ for all $w \in L^2_{\text{loc}}$ and $t \in [0, t(\tau_n,u^*,w))$, there exists $w_\alpha \in L^2_{\text{loc}}$ such that

$$\begin{aligned} V^\lambda(x_n,r_n) \quad -\tfrac{\epsilon_n}{2} < \sup_{t\geq t(\tau_n,u^*,w_\alpha)} & e^{-\lambda k(t,\alpha[w_\alpha],w_\alpha)} W(y(t,x_n,\alpha[w_\alpha],w_\alpha), r_n \\ & + r(t,x_n,\alpha[w_\alpha],w_\alpha)). \end{aligned} \tag{4.12}$$

By definition, for every $x \in \overline{D(A)}$, $r \in \mathbb{R}$, there exists $\alpha = \alpha(x,r,n) \in \Delta$ such that

$$V^\lambda(x,r) + \frac{\epsilon_n}{2} > \sup_{w\in L^2_{\text{loc}}} \sup_{t\geq 0} e^{-\lambda k(t,\alpha[w],w)} W\left(y(t,x,\alpha[w],w), r + r(t,x,\alpha[w],w)\right).$$

Consider a strategy $\bar{\alpha}$ defined by

$$\bar{\alpha}[w](t) = \begin{cases} u^* & \text{if } 0 \leq t < t_n = t(\tau_n,u^*,w), \\ \alpha_n[w(\cdot - t_n)](t-t_n) & \text{for } t \geq t_n, \end{cases}$$

where $\alpha_n = \alpha(y(t_n, x_n, u^*, w), r_n + r(t_n, x_n, u^*, w), n)$. Observe that $\bar{\alpha}$ is nonanticipating (see Remark 4.2 in Soravia [23]), so $\bar{\alpha} \in \Delta$. There is $\bar{w} = w_{\bar{\alpha}} \in L^2_{\text{loc}}$ as in (4.12), and then by the definition of α_n

$$V^\lambda(x_n, r_n) - \frac{\epsilon_n}{2} < e^{-\lambda k(t_n, \bar{\alpha}[\bar{w}], \bar{w})} \Big(V^\lambda\left(y(t_n, x_n, u^*, \bar{w}), r_n + r(t_n, x_n, u^*, \bar{w})\right) + \frac{\epsilon_n}{2} \Big),$$

which contradicts (4.10).

Hence (4.11) must fail and thus for every n we can find $s_n \in [0, \tau_n)$ and a disturbance $w_n \in L^2(0, t(s_n, u^*, w_n); W)$ such that

$$\begin{aligned} V^\lambda(x_n, r_n) - \epsilon_n < e^{-\lambda s_n} W(y(t(s_n, u^*, w_n), x_n, u^*, w_n), r_n \\ + r(t(s_n, u^*, w_n), x_n, u^*, w_n)). \end{aligned} \tag{4.13}$$

Since $(V^\lambda)^*(\hat{x}, \hat{r}) > W(\hat{x}, \hat{r}) \geq 0$, without loss of generality we may assume that there is $\delta > 0$ such that $V^\lambda(x_n, r_n) - \epsilon_n > \delta$ for all n. From (4.13), since ρ is increasing and $t(s_n, u^*, w_n) \leq s_n$, we get

$$0 < \delta < \rho\left(\|U\|_{L^\infty} + r_n + s_n(\|g\|_{L^\infty} + \|u^*\|^2) - \gamma^2 \int_0^{t(s_n, u^*, w_n)} \|w_n\|^2 \right),$$

and it follows that there is $M > 0$ such that $\int_0^{t(s_n, u^*, w_n)} \|w_n\|^2 \leq M$ for all n. Then from (3.3) we obtain $y(t(s_n, u^*, w_n), x_n, u^*, w_n) \to \hat{x}$ as $n \to \infty$ and, taking $\limsup$ as $n \to \infty$ in (4.13), we conclude

$$\begin{aligned} &(V^\lambda)^*(\hat{x}, \hat{r}) \\ &\leq \limsup_{n\to\infty} \quad \rho\left(U(y(t(s_n, u^*, w_n), x_n, u^*, w_n)) + r_n + s_n(\|g\|_{L^\infty} + \|u^*\|^2)\right) \\ &\qquad\qquad\quad = \rho(U(\hat{x}) + \hat{r}) = W(\hat{x}, \hat{r}), \end{aligned}$$

and we have again a contradiction, completing the proof.

(2) To complete the proof of Proposition 4.6, we first argue as in the proof of (4.9) and show that there is $\epsilon > 0$ such that if $\|x - \hat{x}\|$, $|r - \hat{r}|$, $|V^\lambda(x, r) - (V^\lambda)^*(\hat{x}, \hat{r})| < \epsilon$, and $t \in (0, \epsilon)$ then

$$V^\lambda(x, r) \leq \sup_{w \in L^2(0,t;W)} e^{-\lambda k(t, u^*, w)} V^\lambda\left(y(t, x, u^*, w), r + r(t, x, u^*, w)\right). \tag{4.14}$$

Next, since $(V^\lambda)^*(\hat{x}, \hat{r}) > W(\hat{x}, \hat{r}) \geq 0$, we may assume that ϵ is so small that the left-hand side in (4.14) is always bigger than δ for some fixed $\delta > 0$. Therefore, in the supremum in (4.14) only $w \in L^2(0, t; W)$ satisfying

$$V^\lambda\left(y(t, x, u^*, w), r + r(t, x, u^*, w)\right) > \delta$$

are relevant. From Lemma 4.3 there is $N \in \mathbb{R}$ such that for every such w we have $r + r(t, x, u^*, w) > N$, and hence

$$\begin{aligned} \gamma^2 \int_0^t \|w\|^2 &< r + t(\|g\|_{L^\infty} + \|u^*\|^2) - N \\ &\leq \hat{r} + \epsilon(1 + \|g\|_{L^\infty} + \|u^*\|^2) - N. \qquad \square \end{aligned}$$

Proposition 4.6 enables us to show that the value function V^λ in (4.4) is a subsolution of the quasi-variational inequality

$$\begin{aligned}&\min\Big\{V^\lambda - W,\ \lambda V^\lambda + \langle Ax, D_x V^\lambda\rangle + \sup_{u\in U}\inf_{w\in W}\\ &\qquad\big\{-\langle f(x)+Bu+Cw, D_x V^\lambda\rangle + \big(\gamma^2\|w\|^2 - \|u\|^2 - g(x)\big)\, D_r V^\lambda\big\}\Big\} = 0\end{aligned} \tag{4.15}$$

on $\overline{D(A)}\times\mathbb{R}$. Subsolutions and supersolutions of the equation in separated variables (4.15) are defined by modifying Definition 4.1 in an obvious way. Note that the Hamiltonian in (4.15) may not be real valued unless $D_r V^\lambda > 0$, which by Lemma 4.5 holds in the viscosity sense. The following is a delicate step in our method.

PROPOSITION 4.7. *Suppose that g: $\overline{D(A)}\to[0,\infty)$ and U: $\overline{D(A)}\to\mathbb{R}$ are bounded and Lipschitz continuous; let W be as in* (4.1) *and denote $V^\lambda = V^\lambda_{g,W}$. Then V^λ is a subsolution of* (4.15).

Proof. First, observe that, by Lemma 4.5, the value function V^λ is continuous and that by definition $V^\lambda \ge W$. Let $\Phi = \varphi + \psi$ be a subtest function, which we may assume to be Lipschitz continuous; let $\eta \in C^1(\mathbb{R})$ and $V^\lambda - \Phi - \eta$ have a local maximum equal to 0 at $(\hat x,\hat r)\in\overline{D(A)}\times\mathbb{R}$, where $V^\lambda(\hat x,\hat r) > W(\hat x,\hat r)$. Note that by Lemma 4.5 (3) we have $\eta'(\hat r) > 0$, since $V^\lambda(\hat x,\cdot) - \eta(\cdot)$ has a local maximum point at $\hat r$. By modifying η off a small open interval containing $\hat r$, without loss of generality we may assume that there is $\sigma > 0$ such that

$$\eta'(r) > \sigma \quad \text{for all}\ \ r\in\mathbb{R}. \tag{4.16}$$

By definition of viscosity subsolution, we need to show that

$$\begin{aligned}&\lambda V^\lambda(\hat x,\hat r) + D_A^-\Phi(\hat x) + \sup_{u\in U}\inf_{w\in W}\{-\langle f(\hat x)+Bu+Cw, D\varphi(\hat x)\rangle\\ &-L(\psi)(\|f(\hat x)\| + \|Bu\| + \|Cw\|) + \big(\gamma^2\|w\|^2 - \|u\|^2 - g(\hat x)\big)\,\eta'(\hat r)\} \le 0.\end{aligned} \tag{4.17}$$

(1) If $D_A^-\Phi(\hat x) = -\infty$, there is nothing to do. First, we will prove that

$$D_A^-\Phi(\hat x) < +\infty. \tag{4.18}$$

Choose any $u^*\in U$. From Proposition 4.6 there are $M>0$ and $\epsilon>0$ such that for all $t\in(0,\epsilon)$

$$V^\lambda(\hat x,\hat r) \le \sup\{e^{-\lambda k(t,u^*,w)}V^\lambda(y(t),r(t))\colon\ \|w\|_{L^2(0,t;W)}\le M\}, \tag{4.19}$$

where for given w we wrote $y(\cdot) = y(\cdot,\hat x,u^*,w)$ and $r(\cdot) = \hat r + r(\cdot,\hat x,u^*,w)$. Making ϵ smaller if necessary, we can assume that $\|S(t)\hat x - \hat x\| < 1$ and $V^\lambda(S(t)\hat x,\hat r) > V^\lambda(\hat x,\hat r)/2$ for every $t\in(0,\epsilon)$. Suppose that

$$t\in(0,\epsilon),\quad \|w\|_{L^2(0,t;W)}\le M,\quad \text{and}\ \ V^\lambda(\hat x,\hat r)/2 \le V^\lambda(y(t),r(t)). \tag{4.20}$$

Then, from Lemma 4.5, either $V^\lambda(y(t),r(t)) \le V^\lambda(S(t)\hat x,\hat r)$ or

$$\begin{aligned}V^\lambda(y(t),r(t)) - V^\lambda(S(t)\hat x,\hat r) &\le K\|y(t)-S(t)\hat x\| + r(t) - \hat r\\ &\le K\|y(t)-S(t)\hat x\| + t(\|g\|_{L^\infty} + \|u^*\|^2) - \gamma^2\int_0^t \|w\|^2,\end{aligned}$$

with K depending only on $V^\lambda(\hat{x}, \hat{r})$ and λ. From this, (3.5), and the Hölder inequality we obtain that always

$$V^\lambda(y(t), r(t)) \leq V^\lambda(S(t)\hat{x}, \hat{r}) + Ct \quad \text{as} \quad t \downarrow 0,$$

with a constant C uniform for all t and w. Then from (4.19) it follows that $V^\lambda(\hat{x}, \hat{r}) \leq V^\lambda(S(t)\hat{x}, \hat{r}) + Ct$, and, consequently, $\Phi(\hat{x}) - \Phi(S(t)\hat{x}) \leq Ct$ as $t \downarrow 0$, and we obtain (4.18).

(2) In this step, arguing by contradiction, we will show that

$$\begin{aligned} \lambda V^\lambda(\hat{x}, \hat{r}) \quad &+ \sup_{u \in U} \inf_{w \in W} \{[D_A^- \Phi(\hat{x}) - \langle f(\hat{x}) + Bu + Cw, D\varphi(\hat{x})\rangle \\ &-L(\psi)(\|f(\hat{x})\| + \|Bu\| + \|Cw\|) \\ &+ \left(\gamma^2\|w\|^2 - \|u\|^2 - g(\hat{x})\right) \eta'(\hat{r})]/(1 + \|u\|^2 + \|w\|^2)\} \leq 0. \end{aligned} \tag{4.21}$$

Suppose then that there are $u^* \in U$ and $\theta > 0$ such that for every $w \in W$

$$\begin{aligned} \lambda(\Phi(\hat{x}) \quad &+\eta(\hat{r}))\left(1 + \|u^*\|^2 + \|w\|^2\right) + D_A^-\Phi(\hat{x}) - \langle f(\hat{x}) + Bu^* + Cw, D\varphi(\hat{x})\rangle \\ &-L(\psi)(\|f(\hat{x})\| + \|Bu^*\| + \|Cw\|) \\ &+ \left(\gamma^2\|w\|^2 - \|u^*\|^2 - g(\hat{x})\right) \eta'(\hat{r}) > 2\theta(1 + \|u^*\|^2 + \|w\|^2). \end{aligned} \tag{4.22}$$

Fix $\tau > 0$ and let $w \in L^2_{\text{loc}}(0, \infty; W)$. Denote $y(\cdot) = y(\cdot, \hat{x}, u^*, w)$ and $r(\cdot) = \hat{r} + r(\cdot, \hat{x}, u^*, w)$. Note that

$$t(\tau, u^*, w) + \|w\|^2_{L^2(0, t(\tau, u^*, w); W)} \leq k(t(\tau, u^*, w), u^*, w) = \tau. \tag{4.23}$$

Therefore, from (4.22), (4.16), and (3.4) it follows that if τ is sufficiently small, then for all $s \in [0, t(\tau, u^*, w)]$ (we may suppose $\theta < \sigma\gamma^2$)

$$\begin{aligned} D_A^-\Phi(\hat{x}) \quad &+\lambda\left(\Phi(y(s)) + \eta(r(s))\right)\left(1 + \|u^*\|^2 + \|w(s)\|^2\right) - \langle v(s), D\varphi(y(s))\rangle \\ &\qquad -L(\psi)\|v(s)\| + \left(\gamma^2\|w(s)\|^2 - \|u^*\|^2 - g(y(s))\right)\eta'(r(s)) \\ &> \theta(1 + \|u^*\|^2 + \|w(s)\|^2), \end{aligned} \tag{4.24}$$

with $v(s) = f(y(s)) + Bu^* + Cw(s)$. Multiplying by $e^{-\lambda k(s, u^*, w)}$, integrating from 0 to $t(\tau, u^*, w)$, and using a version of Tataru's result (see Corollaries 4.8 and 4.9 in the paper by the authors and Święch [19]) as in the proof of Proposition 2.8, we obtain

$$\begin{aligned} &\Phi(\hat{x}) + \eta(\hat{r}) - e^{-\lambda\tau}[\Phi(y(t(\tau, u^*, w))) + \eta(r(t(\tau, u^*, w)))] \\ &\geq (1 - e^{-\lambda\tau})\tfrac{\theta}{\lambda} - o(t(\tau, u^*, w)) \geq (1 - e^{-\lambda\tau})\tfrac{\theta}{\lambda} - o(\tau), \end{aligned} \tag{4.25}$$

as $\tau \to 0$, uniformly for all $w \in L^2_{\text{loc}}$. Using (4.23), (3.4), and the maximum property of $\Phi + \eta$, we have

$$\begin{aligned} \sup \quad &\left\{e^{-\lambda\tau}V^\lambda(y(t(\tau, u^*, w)), r(t(\tau, u^*, w))) \colon \ w \in L^2(0, t; W)\right\} \\ &\leq V^\lambda(\hat{x}, \hat{r}) - (1 - e^{-\lambda\tau})\tfrac{\theta}{\lambda} + o(\tau). \end{aligned}$$

Hence by Proposition 4.6, for a sufficiently small $\tau > 0$, this gives $(1 - e^{-\lambda\tau})\theta/\lambda \leq o(\tau)$ and a contradiction when $\tau \downarrow 0$. Thus (4.21) is proved.

(3) Let $\epsilon > 0$. From (4.21) for every $u \in U$ there is $w_u \in W$ such that

$$\begin{aligned} \lambda V^\lambda(\hat{x}, \hat{r}) \quad &(1 + \|u\|^2 + \|w_u\|^2) + D_A^-\Phi(\hat{x}) - \langle f(\hat{x}) + Bu + Cw_u, D\varphi(\hat{x})\rangle \\ &-L(\psi)(\|f(\hat{x})\| + \|Bu\| + \|Cw_u\|) \\ &+ \left(\gamma^2\|w_u\|^2 - \|u\|^2 - g(\hat{x})\right)\eta'(\hat{r}) \leq \epsilon(1 + \|u\|^2 + \|w_u\|^2), \end{aligned} \tag{4.26}$$

and therefore

$$\begin{aligned}(4.27)\quad \|u\|^2 &\big(\lambda V^\lambda(\hat{x},\hat{r})-\eta'(\hat{r})-\epsilon\big)-\langle Bu,D\varphi(\hat{x})\rangle-L(\psi)\|Bu\|\\ &\le \epsilon-D_A^-\Phi(\hat{x})-\lambda V^\lambda(\hat{x},\hat{r})+\langle f(\hat{x}),D\varphi(\hat{x})\rangle+L(\psi)\|f(\hat{x})\|+g(\hat{x})\eta'(\hat{r})\\ &+\|w_u\|^2\big(\epsilon-\lambda V^\lambda(\hat{x},\hat{r})-\gamma^2\eta'(\hat{r})\big)+\langle Cw_u,D\ gf(\hat{x})\rangle+L(\psi)\|Cw_u\|.\end{aligned}$$

If $\epsilon<\sigma\gamma^2$, then the coefficient of $\|w_u\|^2$ on the right-hand side in (4.27) is negative and hence the supremum over all $w_u\in W$ is finite. Then taking the supremum over all $u\in U$ on the left-hand side shows that $\lambda V^\lambda(\hat{x},\hat{r})-\eta'(\hat{r})-\epsilon\le 0$ for every small enough $\epsilon>0$, i.e.,

$$\lambda V^\lambda(\hat{x},\hat{r})\le\eta'(\hat{r}).$$

(4) Using (4.16), it is easy to see that there is $R>0$ such that in the supremum (4.17) it is not restrictive to take $\|u\|\le R$. Now if $\epsilon<\gamma^2\sigma$, then taking the infimum over all $\|u\|\le R$ in (4.27) shows that there is $K=K(R)>0$ independent of ϵ small enough such that $\|w_u\|\le K$ for every $\|u\|\le R$. From (4.26) we then have

$$\begin{aligned}\lambda V^\lambda(\hat{x},\hat{r})\quad &+D_A^-\Phi(\hat{x})+\sup_{u\in U}\inf_{w\in W}\{-\langle f(\hat{x})+Bu+Cw,D\varphi(\hat{x})\rangle\\ &\qquad-L(\psi)(\|f(\hat{x})\|+\|Bu\|+\|Cw\|)+\big(\gamma^2\|w\|^2-\|u\|^2-g(\hat{x})\big)\,\eta'(\hat{r})\}\\ &\le\epsilon(1+K^2+R^2),\end{aligned}$$

and (4.17) follows, as ϵ is arbitrary. □

5. Proofs of the main results. In this section we put everything together to obtain proofs of the main statements. We first deal with the key point of the sufficiency part of our theorems.

Proof of Proposition 2.9. Suppose that $U\in LSC(\overline{D(A)})$ bounded from below is a supersolution of (1.6). Note that we need only to show that U is bigger than the right-hand side of (2.7), by choosing $t=0$. (1) As explained in [17], we can construct two nondecreasing sequences $(g_n)_n$, $g_n\ge 0$, and $(U_n)_n$ of bounded, globally Lipschitz functions defined on H such that on $\overline{D(A)}$ $g=\sup_n g_n$ and $U=\sup_n U_n$. For every n, put $W_n(x,r)=\rho(U_n(x)+r)$, so that $W_n\colon\ H\times\mathbb{R}\to(0,\pi)$ and $W_n\in Lip(H\times\mathbb{R})$. Observe that $W\ge W_n$ for every n. By Proposition 4.2, for every n and $\lambda>0$, $W(x,r)=\rho(U(x)+r)$ is a supersolution of (4.2) on $\overline{D(A)}\times\mathbb{R}$.

(2) Denoting $V_n=V^\lambda_{g_n,W_n}$, by Proposition 4.7 and Lemma 4.5 we know that V_n is a continuous subsolution of

$$\begin{aligned}(5.1)\quad &\min\Big\{V_n-W_n,\ \lambda V_n+\langle Ax,D_xV_n\rangle+\sup_{u\in U}\inf_{w\in W}\\ &\qquad\big\{-\langle f(x)+Bu+Cw,D_xV_n\rangle+\big(\gamma^2\|w\|^2-\|u\|^2-g(x)\big)\,D_rV_n\big\}\Big\}=0\end{aligned}$$

on $\overline{D(A)}\times\mathbb{R}$. We will show that

$$(5.2)\qquad V_n\le W\quad\text{on}\quad\overline{D(A)}\times\mathbb{R}.$$

The proof of (5.2) is adapted from a similar argument of Lemma 5.5 in [17]. We argue by contradiction and suppose that $V_n(\hat{z},\hat{t})-W(\hat{z},\hat{t})\equiv 3\tau>0$ for some $\hat{z}\in\overline{D(A)}$ and $\hat{t}\in\mathbb{R}$. For the sake of simplicity we will assume that $(0,0)\in A$; see step 3 in the proof of Lemma 5.5 in [17] for the small technical modifications necessary in the general case. For $\alpha, \delta, \beta, \kappa > 0$ and $x, y \in \overline{D(A)}$ and $r, s \in \mathbb{R}$ let

$$\Phi(x,y,r,s) = V_n(x,r) - W(y,s) - \frac{\alpha}{2}\|x-y\|^2 - \frac{\beta}{2}(r-s)^2 - \delta(\|x\|^2 + \|y\|^2) - \kappa(r^2+s^2).$$

If δ, κ are sufficiently small, then

$$\sup \Phi \geq V_n(\hat{z},\hat{t}) - W(\hat{z},\hat{t}) - 2\delta\|\hat{z}\|^2 - 2\kappa\hat{t}^2 \geq 2\tau.$$

By the perturbed optimization result of the first author and Święch [20], for every $\epsilon > 0$ one can find $\hat{x}, \hat{y} \in \overline{D(A)}$, $\hat{r}, \hat{s} \in \mathbb{R}$ and $a, b \in \mathbb{R}$ such that $|a|, |b| < \epsilon$, $\Phi(\hat{x},\hat{y},\hat{r},\hat{s}) \geq \sup\Phi - \epsilon$ and the map

$$\Phi(x,y,r,s) - \epsilon d(x,\hat{x}) - \epsilon d(y,\hat{y}) - ar - bs$$

has a strict global maximum at $(\hat{x},\hat{y},\hat{r},\hat{s})$. The function d that we are using above is the so-called Tataru distance

$$d(x,y) = \inf_{t\geq 0}\{t + \|x - S(t)x\|\}.$$

For its properties, besides those mentioned in [20], we refer the reader to [26], [27], [10]. Note that $\Phi(\hat{x},\hat{y},\hat{r},\hat{s}) \geq \sup\Phi - \epsilon \geq 2\tau - \epsilon \geq \tau$ for all sufficiently small ϵ, and then

$$\begin{aligned} V_n(\hat{x},\hat{r}) \geq{}& W(\hat{y},\hat{s}) + \frac{\alpha}{2}\|\hat{x}-\hat{y}\|^2 + \frac{\beta}{2}|\hat{r}-\hat{s}|^2 \\ &+ \delta\left(\|\hat{x}\|^2 + \|\hat{y}\|^2\right) + \kappa\left(|\hat{r}|^2 + |\hat{s}|^2\right) + \tau. \end{aligned} \tag{5.3}$$

In particular, $\hat{y} \in \mathrm{dom}(U)$. Consider two cases. If $V_n(\hat{x},\hat{r}) > W_n(\hat{x},\hat{r})$, then we use (4.2) and (5.1), and from the doubling Theorem 3.1 in Crandall–Lions [10] (see also [16] or [21]; note that our notions of viscosity subsolutions and supersolutions slightly differ from those employed in [10], yet the proof of the doubling theorem can be easily adapted to our setting) we obtain

$$\begin{aligned} &\lambda V_n(\hat{x},\hat{r}) \leq \sup_{u\in U}\inf_{w\in W}\{-\langle f(\hat{y}) + Bu + Cw, \alpha(\hat{x}-\hat{y}) - 2\delta\hat{y}\rangle \\ &+\epsilon\left(\|f(\hat{y})\| + \|Bu\| + \|Cw\|\right) + \left(\gamma^2\|w\|^2 - \|u\|^2 - g_n(\hat{y})\right)\left(\beta(\hat{r}-\hat{s}) - 2\kappa\hat{s} - b\right)\} \\ &- \sup_{u\in U}\inf_{w\in W}\{-\langle f(\hat{x}) + Bu + Cw, \alpha(\hat{x}-\hat{y}) + 2\delta\hat{x}\rangle \\ &-\epsilon\left(\|f(\hat{x})\| + \|Bu\| + \|Cw\|\right) + \left(\gamma^2\|w\|^2 - \|u\|^2 - g_n(\hat{x})\right)\left(\beta(\hat{r}-\hat{s}) + 2\kappa\hat{r} + a\right)\} + 2\epsilon. \end{aligned} \tag{5.4}$$

From (5.3) we obtain that $V_n(\hat{x},\hat{r}) \geq \tau$ and $W(\hat{y},\hat{s}) \leq \pi - \tau$, and Lemma 4.3 and the properties of U and ρ, respectively, give that $\hat{r}$ is bounded from below and $\hat{s}$ is bounded from above, uniformly in all parameters. Also, from (5.3), $\beta|\hat{r}-\hat{s}|^2 \leq 2\pi$, and therefore for any fixed β, $\hat{r}$ and $\hat{s}$ are bounded, uniformly in all other parameters. Now Lemma 4.5 (3) together with the fact that $V_n(\hat{x},r) - \frac{\beta}{2}(r-\hat{s})^2 - \kappa r^2 - ar$ has a maximum at $r = \hat{r}$ yield that

$$\beta(\hat{r}-\hat{s}) + 2\kappa\hat{r} + a \geq \nu = \nu(\beta) > 0. \tag{5.5}$$

From (5.3)

$$\kappa|\hat{r}|, \kappa|\hat{s}| \leq \sqrt{\kappa\pi}, \quad \delta\|\hat{x}\|, \delta\|\hat{y}\| \leq \sqrt{\delta\pi}. \tag{5.6}$$

Also recall that, by standard arguments, see, e.g., Lemma 3.5 in [10], we have

$$\begin{aligned}&\limsup_{\beta\to\infty}\limsup_{\alpha\to\infty}\limsup_{\delta\downarrow 0}\limsup_{\kappa\downarrow 0}\\ &\limsup_{\epsilon\downarrow 0}\left(\alpha\|\hat{x}-\hat{y}\|^2+\delta(\|\hat{x}\|^2+\|\hat{y}\|^2)\right)=0.\end{aligned}\tag{5.7}$$

Equation (5.3) also gives that for α, β fixed, $\alpha(\hat{x}-\hat{y})$ and $\beta(\hat{r}-\hat{s})$ remain bounded, uniformly in other parameters. Hence taking the limit as $\epsilon \downarrow 0$ in (5.4), using (5.5), (5.6), and (5.3) and the estimate on the Hamiltonian as in Lemma 5.4 in [17], we have

$$\begin{aligned}\lambda\tau \le \limsup_{\epsilon\downarrow 0}\{H_n(\hat{y},\quad &\alpha(\hat{x}-\hat{y})-2\delta\hat{y},\beta(\hat{r}-\hat{s})-2\kappa\hat{s})\\ &-H_n(\hat{x},\alpha(\hat{x}-\hat{y})+2\delta\hat{x},\beta(\hat{r}-\hat{s})+2\kappa\hat{r})\},\end{aligned}\tag{5.8}$$

where for $(x,p,v)\in\overline{D(A)}\times H\times(0,+\infty)$ we wrote

$$\begin{aligned}H_n(x,p,v) &= \sup_{u\in U}\inf_{w\in W}\{-\langle f(x)+Bu+Cw,p\rangle\\ &\quad+\left(\gamma^2\|w\|^2-\|u\|^2-g_n(x)\right)v\}\\ &= -\langle f(x),p\rangle-g_n(x)v+\tfrac{1}{4v}\|B^*p\|^2-\tfrac{1}{4v\gamma^2}\|C^*p\|^2.\end{aligned}$$

Note that H_n is uniformly continuous on bounded closed subsets of $\overline{D(A)}\times H\times(0,+\infty)$, and thus using (5.5) and (5.6) while taking the limits gives

$$\begin{aligned}\lambda\tau \le \limsup_{\kappa\downarrow 0}\quad &\limsup_{\delta\downarrow 0}\limsup_{\epsilon\downarrow 0}(\beta L(g_n)|\hat{r}-\hat{s}|\|\hat{x}-\hat{y}\|\\ &+\alpha\|f(\hat{x})-f(\hat{y})\|\|\hat{x}-\hat{y}\|+2\delta\left(|\langle f(\hat{x}),\hat{x}\rangle|+|\langle f(\hat{y}),\hat{y}\rangle|\right))\end{aligned}$$

and then (1.3) and (5.7) yield a contradiction.

The second case applies if $V_n(\hat{x},\hat{r})=W_n(\hat{x},\hat{r})$. Therefore by (5.3) we have

$$\tau\le V_n(\hat{x},\hat{r})-W(\hat{y},\hat{s})\le W_n(\hat{x},\hat{r})-W_n(\hat{y},\hat{s}),\tag{5.9}$$

and then the fact that $W_n\in Lip(H\times\mathbb{R})$ and (5.3) also yield a contradiction as $\alpha,\beta\to\infty$, and thus (5.2) is proved.

(3) So far we proved that for every $\lambda>0$ and $n\ge 1$

$$\begin{aligned}&W(x,r)\ge\inf_{\alpha\in\Delta}\sup_{w\in L^2_{\text{loc}}}\\ &\sup_{t\ge 0}e^{-\lambda k(t,\alpha[w],w)}W_n(y(t,x,\alpha[w],w),r+r(t,x,\alpha[w],w))\\ &=\inf_{\alpha\in\Delta}\sup_{w\in L^2_{\text{loc}}}\sup_{\tau\ge 0}e^{-\lambda\tau}W_n(y(t(\tau,\alpha[w],w)),r+r(t(\tau,\alpha[w],w)))\end{aligned}\tag{5.10}$$

for $x\in\overline{D(A)}$, $U(x)<+\infty$, and $r\in\mathbb{R}$. Fix $x\in\overline{D(A)}$, $U(x)<+\infty$. Letting $\lambda\downarrow 0$ in (5.10), we obtain for every fixed $T>0$

$$\begin{aligned}&W(x,r)=\rho(U(x)+r)\ge\inf_{\alpha\in\Delta}\sup_{w\in L^2_{\text{loc}}(0,\infty;W)}\sup_{\tau\in[0,T]}\\ &\{W_n(y(t(\tau,\alpha[w],w)),r+r(t(\tau,\alpha[w],w)))\}=\rho\Bigg(\inf_{\alpha\in\Delta}\sup_{w\in L^2_{\text{loc}}}\sup_{\tau\in[0,T]}\\ &\left\{\int_0^{t(\tau,\alpha[w],w)}\left(g_n(y(s))+\|\alpha[w](s)\|^2-\gamma^2\|w(s)\|^2\right)ds+r+U_n(y(t(\tau,\alpha[w],w)))\right\}\Bigg),\end{aligned}$$

which implies, since ρ is increasing, for all $n \geq 1$ and $T > 0$

$$(5.11)\quad U(x) \geq \inf_{\alpha\in\Delta} \sup_{w\in L^2_{\text{loc}}} \sup_{\tau\in[0,T]} \left\{ \int_0^{t(\tau,\alpha[w],w)} (g_n(y(s)) + \|\alpha[w](s)\|^2 - \gamma^2\|w(s)\|^2)ds + U_n(y(t(\tau,\alpha[w],w))) \right\}.$$

We can pass to the limit as $n \to \infty$ in (5.11), with an argument similar to the one in Lemma 3.3. Since by definition, for all $\alpha \in \Delta$ and $w \in L^2_{\text{loc}}(0,\infty;W)$, we have $\|\alpha[w]\|^2_{L^2(0,t(T,\alpha[w],w))} \leq T$ and $t(T,\alpha[w],w) \leq T$, in (5.11), we can always limit ourselves to using strategies verifying $\|\alpha[w]\|^2_{L^2(0,\infty)} \leq T$ for all $w \in L^2_{\text{loc}}$. Similarly, we can think here that disturbances $w \in L^2_{\text{loc}}$ satisfy $\|w\|_{L^2(0,\infty)} \leq T$. We can then construct a limit strategy $\alpha^\#$ as in Lemma 3.3, part (2), starting with a family of strategies $\{\alpha_n\}$, where α_n is $1/n$-optimal in (5.11), i.e., satisfies

$$(5.12)\qquad \begin{aligned} U(x) + \frac{1}{n} \geq \int_0^{t(\tau,\alpha_n[w],w)} &(g_n(y(s)) + \|\alpha_n[w](s)\|^2 - \gamma^2\|w(s)\|^2)ds \\ &+ U_n(y(t(\tau,\alpha_n[w],w))), \end{aligned}$$

for all $w(\cdot)$ and $\tau \in [0,T]$. Moreover, the fact that the sequence U_n is uniformly bounded from below (with no loss of generality) gives

$$U(x) + \frac{1}{n} + M \geq \int_0^{t(\tau,\alpha_n[w],w)} (\|\alpha_n[w](s)\|^2 - \gamma^2\|w(s)\|^2)ds;$$

hence, in particular, for $\tau = T$ and allowing only disturbances such that $\|w\|_2^2 \leq C$,

$$t(T,\alpha_n[w],w) + U(x) + \frac{1}{n} + M + (\gamma^2+1)C \geq T.$$

Thus $t_n = t(T,\alpha_n[w],w) \geq 1$ for $T \geq T(U,x,C)$ large enough. From now on we then fix $T \geq T(U,x,C)$. Given $\tau \in [0,T]$, we set

$$\tau_n = k(t(\tau,\alpha^\#,w),\alpha_n,w).$$

Note that $\tau_n \leq k(\tau,\alpha_n[w],w) \leq k(1,\alpha_n[w],w) \leq T$ for $\tau \in [0,1]$ if $t_n \geq 1$, and therefore we can replace τ by τ_n in (5.12) and get

$$\begin{aligned} U(x) + \frac{1}{n} \geq \int_0^{t(\tau,\alpha^\#[w],w)} &\left(g_n(y(s)) + \|\alpha_n[w](s)\|^2 - \gamma^2\|w(s)\|^2\right) ds \\ &+ U_n(y(t(\tau,\alpha^\#[w],w))) \end{aligned}$$

for all $w(\cdot)$, $\|w\|_2^2 \leq C$, and $\tau \in [0,1]$. Using the assumption (2.1) as in Lemma 3.3, part (3), to pass to the limit as $n \to +\infty$ we get that for every $x \in \overline{D(A)}$ and $C > 0$

$$(5.13)\quad U(x) \geq \inf_{\alpha\in\Delta} \sup_{w\in L^2_{loc},\|w\|_2^2\leq C} \sup_{\tau\in[0,1]} \left\{ \int_0^{t(\tau,\alpha[w],w)} \left(g(y(s)) + \|\alpha[w](s)\|^2 - \gamma^2\|w(s)\|^2\right) ds + U(y(t(\tau,\alpha[w],w))) \right\}.$$

A further diagonal argument, again using the machinery of Lemma 3.3 in a similar fashion, is then needed to remove in (5.13) the restriction on the L^2 norm of the disturbances (letting $C \to +\infty$), but we skip the details. In order to take the horizon to ∞ in (5.13) (but with $C = +\infty$), we proceed as follows. For a given $\varepsilon > 0$ and $w \in L^2_{\mathrm{loc}}(0, \infty; W)$, we apply (5.13) and find $\alpha_0 \in \Delta$ (as a matter of fact (5.13) has optimal strategies, and we can even choose α_0 as a function of x) such that, for $x_0 = x$ and $w_0 = w$,

$$\begin{aligned} U(x_0) + \tfrac{\varepsilon}{2} \;\; &\geq \int_0^{t(\tau,\alpha_0[w_0],w_0)} \left(g(y(s, x, \alpha_0[w_0], w_0)) + \|\alpha_0[w_0](s)\|^2 - \gamma^2\|w_0(s)\|^2\right) ds \\ &+U(y(t(\tau, \alpha_0[w_0], w_0))) \end{aligned}$$

for $\tau \in [0, 1]$. Then we apply (5.13) with $C = +\infty$ again at $x_1 = y(1, x, \alpha_0[w_0], w_0)$ and find $\alpha_1 \in \Delta$ such that, for $w_1(\cdot) = w_0(\cdot + 1)$,

$$\begin{aligned} U(x_1) + \tfrac{\varepsilon}{2^2} \;\; &\geq \int_0^{t(\tau,\alpha_1[w_1],w_1)} \left(g(y(s, x_1, \alpha_1[w_1], w_1)) + \|\alpha_1[w_1](s)\|^2 - \gamma^2\|w_1(s)\|^2\right) ds \\ &+U(y(t(\tau, \alpha_1[w_1], w_1), x_1, \alpha_1[w_1], w_1)) \end{aligned}$$

for $\tau \in [0, 1]$, and so forth. We proceed recursively and define the strategy α by the position $\alpha[w] = \overline{u}$, where $\overline{u}$ is the control defined by setting

$$\overline{u}(s) = \begin{cases} \alpha_0[w_0](s), & s \in [0, t(1, \alpha_0[w_0], w_0)), \\ \alpha_1[w_1](s - t(1, \alpha_0[w_0], w_0)), & s \in I_1, \\ \dots . & \end{cases}$$

$I_1 = [t(1, \alpha_0[w_0], w_0), t(1, \alpha_0[w_0], w_0) + t(1, \alpha_1[w_1], w_1))$. The strategy α is a causal functional by construction, since $t(\tau, u, w)$ is a causal functional of the controls for any fixed $\tau > 0$, and, moreover,

$$t(1, \alpha_0[w_0], w_0) + t(1, \alpha_1[w_1], w_1) = t(2, \alpha[w], w),$$

and so forth (see Remark 4.2 in [23] for both statements). Then α satisfies

$$\begin{aligned} U(x) + \varepsilon \;\; &\geq \int_0^{t(\tau,\alpha[w],w)} \left(g(y(s, x, \alpha[w], w)) + \|\alpha[w](s)\|^2 - \gamma^2\|w(s)\|^2\right) ds \\ &+U(y(t(\tau, \alpha[w], w), x, \alpha[w], w)) \end{aligned}$$

for all $\tau \geq 0$ and $w \in L^2_{\mathrm{loc}}(0, \infty; W)$; therefore, in particular, $\alpha[w] \in L^2_{\mathrm{loc}}(0, \infty; U)$ for all $w \in L^2_{\mathrm{loc}}(0, \infty; W)$, since g and U are nonnegative and $t(\tau, u, w) \to +\infty$ as $\tau \to +\infty$. Hence, since ϵ is arbitrary, we obtain for all $x \in \overline{D(A)}$

$$U(x) \geq \inf_{\alpha\in\Delta} \sup_{w\in L^2_{\mathrm{loc}}(0,\infty;W)} \sup_{t\geq 0} \left\{ \int_0^t (g(y(s)) + \|\alpha[w](s)\|^2 - \gamma^2\|w(s)\|^2) ds + U(y(t)) \right\},$$

which concludes the proof since the opposite inequality follows immediately by choosing $t = 0$ in the right-hand side. □

Proof of Theorem 2.4. To prove the Lyapunov stability at 0 of the undisturbed system, let $U\colon \overline{D(A)} \to [0, +\infty]$, $\mathrm{dom}(U) \supset \overline{\mathrm{dom}(g)}$, be a lower semicontinuous supersolution of the HJI equation (1.6), vanishing and continuous at the origin. Suppose that U is locally positive definite at 0 and that σ, $\omega(\cdot)$ are as in Definition 2.2. From

Proposition 2.9, (2.7) holds, and, moreover, the right-hand side of (2.7) has optimal strategies, as it follows applying Lemma 3.3 with $g_n \equiv g$ and $\varphi_n \equiv U$. For all $x \in \operatorname{dom}(U)$ we can then find $\alpha_x \in \Delta$ such that

$$U(x) \geq \int_0^T \left(g(y(t)) + \|\alpha_x[w](t)\|^2 - \gamma^2\|w(t)\|^2\right) dt + U(y(T)) \tag{5.14}$$

for all $T \geq 0$ and $w \in L^2_{\text{loc}}(0, +\infty; W)$. Then (1.5) is satisfied by $\{\alpha_x\}_x$ and $K = U$. By choosing $w \equiv 0$ we then get

$$U(x) \geq U(y(T, x, \alpha_x[0], 0))$$

for all $T \geq 0$, so, in particular, the sublevel sets $\mathcal{V}_\delta = \{x \in \overline{D(A)}\colon\ U(x) < \delta\}$ of U are invariant for the undisturbed system. Moreover, since U is continuous at 0 and $U(0) = 0$, every $\mathcal{V}_\delta$ is a neighborhood (relative to $\overline{D(A)}$) of 0.

Now suppose that $\mathcal{U} = B^H_\epsilon(0) \cap \overline{D(A)}$, where $0 < \epsilon \leq \sigma$. Consider $\mathcal{V} = \mathcal{U} \cap \mathcal{V}_{\omega(\epsilon)}$, a relative neighborhood of 0. If $x \in \mathcal{V}$ and $\|y(t, x, \alpha_x[0], 0)\| = \epsilon$ for some $t > 0$, then $U(y(t, x, \alpha_x[0], 0)) \geq \omega(\epsilon)$, contradicting the invariance of $\mathcal{V}_{\omega(\epsilon)}$. It follows that if $x \in \mathcal{V}$, then $y(t, x, \alpha_x[0], 0) \in \mathcal{V} \subset \mathcal{U}$ for all $t \geq 0$, and the stability at 0 of the undisturbed system follows.

The above argument in fact shows that the family of sets $\mathcal{U}_\epsilon = \{x \in \overline{D(A)}\colon\ \|x\| < \epsilon,\ U(x) < \omega(\epsilon)\}$, $0 < \epsilon \leq \sigma$, is invariant for the undisturbed system and, moreover, is a base of neighborhoods (relative to $\overline{D(A)}$) of 0.

To prove asymptotic stability at 0 when g is locally positive definite at 0, we need a further argument. Let σ, $\omega(\cdot)$ be as in Definition 2.2; it is not restrictive to assume they are the same as above. For $0 < \epsilon \leq \sigma$ put

$$\rho(\epsilon) = \inf\{g(x)\colon\ x \in B^H_\sigma \cap \overline{D(A)} \setminus \mathcal{U}_\epsilon\} > 0.$$

Choosing $w \equiv 0$ in (5.14), we get

$$U(x) \geq \int_0^T g(y(t, x, \alpha_x[0], 0))\, dt$$

for all $T \geq 0$. Suppose that $x \in \mathcal{U}_\sigma$. If $0 < \epsilon < \sigma$ and $y(t, x, \alpha_x[0], 0) \notin \mathcal{U}_\epsilon$ for $t \in [0, T]$, then $T \leq U(x)/\rho(\epsilon)$. Thus $y(t, x, \alpha_x[0], 0) \in \mathcal{U}_\epsilon$ for some $t \geq 0$, and we conclude by the invariance of $\mathcal{U}_\epsilon$. □

REFERENCES

[1] J.A. Ball, J.W. Helton, and M.L. Walker, *H_∞ control for nonlinear systems with output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 546–559.

[2] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leiden, The Netherlands, 1976.

[3] V. Barbu, *Analysis and Control of Infinite Dimensional Systems*, Academic Press, Boston, 1993.

[4] V. Barbu, *H_∞-control for semilinear systems in Hilbert spaces*, Systems Control Lett., 21 (1993), pp. 65–72.

[5] V. Barbu, *H_∞-boundary control with state feedback: The hyperbolic case*, SIAM J. Control Optim., 33 (1995), pp. 684–701.

[6] V. Barbu, *The H_∞-problem for infinite dimensional semilinear systems*, SIAM J. Control Optim., 33 (1995), pp. 1017–1027.

[7] T. Basar and P. Bernhard, *$\mathcal{H}_\infty$ Optimal Control and Related Minimax Design Problems*, 2nd ed., Birkhäuser Boston, Boston, MA, 1995.

[8] H. Brézis, *Operateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert*, North-Holland, Amsterdam, 1973.

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

[10] M.G. Crandall and P.-L. Lions, *Hamilton-Jacobi equations in infinite dimensions.* VI. *Nonlinear A and Tataru's method refined*, in Evolution Equations, Control Theory, and Biomathematics, Lecture Notes in Pure and Appl. Math. 155, P. Clément and G. Lumer, eds., Marcel Dekker, New York, 1994, pp. 51–89.

[11] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

[12] J. Doyle, K. Glover, P.P. Khargonekar, and B.A. Francis, *State space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[13] R.J. Elliott and N.J. Kalton, *The Existence of Value in Differential Games*, Mem. Amer. Math. Soc., 126 (1972).

[14] L.C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, Amer. Math. Soc., Providence, RI, 1998.

[15] K. Glover and J. Doyle, *State space formulas for all stabilizing controllers that satisfy an H_∞ norm bound and relations to risk-sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[16] M. Kocan, *Some Aspects of the Theory of Viscosity Solutions of Fully Nonlinear PDE's in Infinite Dimensions*, Ph.D. thesis, University of California at Santa Barbara, 1994.

[17] M. Kocan and P. Soravia, *A viscosity approach to infinite-dimensional Hamilton–Jacobi equations arising in optimal control with state constraints*, SIAM J. Control Optim., 36 (1998), pp. 1348–1375.

[18] M. Kocan and P. Soravia, *Nonlinear, dissipative, infinite dimensional systems*, in Stochastic Analysis, Control, Optimization, and Applications, W.H. Fleming et al., eds., Birkhäuser Boston, Boston, MA, 1998, pp. 75–93.

[19] M. Kocan, P. Soravia, and A. Święch, *On differential games for infinite dimensional systems with nonlinear, unbounded operators*, J. Math. Anal. Appl., 211 (1997), pp. 395–423.

[20] M. Kocan and A. Święch, *Perturbed optimization on product spaces*, Nonlinear Anal., 26 (1996), pp. 81–90.

[21] M. Kocan and A. Święch, *Second order unbounded parabolic equations in separated form*, Studia Math., 115 (1995), pp. 291–310.

[22] C. Scherer, *H_∞-control by state feedback: An alternative algorithm and characterization of high-gain occurrence*, Systems Control Lett., 12 (1989), pp. 383–391.

[23] P. Soravia, *H_∞ control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.

[24] P. Soravia, *Nonlinear H_∞ control*, in Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations, M. Bardi and I. Capuzzo-Dolcetta, eds., Birkhäuser, Basel, Boston, 1998, pp. 504–531.

[25] P. Soravia, *Equivalence between nonlinear H_∞ control problems and existence of viscosity solutions of Hamilton-Jacobi equations*, Appl. Math. Optim., 39 (1999), pp. 17–32.

[26] D. Tataru, *Viscosity solutions for Hamilton-Jacobi equations with unbounded nonlinear terms*, J. Math. Anal. Appl., 163 (1992), pp. 345–392.

[27] D. Tataru, *Viscosity solutions for Hamilton-Jacobi equations with unbounded nonlinear term: A simplified approach*, J. Differential Equations, 111 (1994), pp. 123–146.

[28] A.J. van der Schaft, *L_2 gain analysis for nonlinear systems and nonlinear H_∞ control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

[29] K. Van Keulen, M. Peters, and R. Curtain, *H_∞-control with state feedback: The infinite dimensional case*, J. Math. Syst. Estim. Control, 3 (1993), pp. 1–39.

[30] K. Van Keulen, *H_∞-control with measurement feedback for linear infinite dimensional systems*, J. Math. Syst. Estim. Control, 3 (1993), pp. 373–411.

[31] T.J. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[32] I.I. Vrabie, *Compactness Methods for Nonlinear Evolution*, 2nd ed., Longman, London, 1995.

[33] G. Zames, *Feedback optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301–320.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1323–1350

CONVEXITY IN HAMILTON–JACOBI THEORY I: DYNAMICS AND DUALITY*

R. TYRRELL ROCKAFELLAR† AND PETER R. WOLENSKI‡

Abstract. Value functions propagated from initial or terminal costs and constraints by way of a differential inclusion, or more broadly through a Lagrangian that may take on ∞, are studied in the case where convexity persists in the state argument. Such value functions, themselves taking on ∞, are shown to satisfy a subgradient form of the Hamilton–Jacobi equation which strongly supports properties of local Lipschitz continuity, semidifferentiability and Clarke regularity. An extended "method of characteristics" is developed which determines them from the Hamiltonian dynamics underlying the given Lagrangian. Close relations with a dual value function are revealed.

Key words. convex value functions, dual value functions, subgradient Hamilton–Jacobi equations, extended method of characteristics, nonsmooth Hamiltonian dynamics, viscosity solutions, variational analysis, optimal control, generalized problems of Bolza

AMS subject classifications. Primary, 49L25; Secondary, 93C10, 49N15

PII. S0363012998345366

1. Introduction.

Fundamental to optimal control and the calculus of variations are value functions $V : [0,\infty) \times \mathbb{R}^n \to \overline{\mathbb{R}} := [-\infty,\infty]$ of the type

$$(1.1)\; V(\tau,\xi) := \inf\left\{ g\big(x(0)\big) + \int_0^\tau L\big(x(t),\dot{x}(t)\big)dt \,\middle|\, x(\tau)=\xi \right\}, \qquad V(0,\xi) = g(\xi),$$

which propagate an initial cost function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ forward from time 0 in a manner dictated by a Lagrangian function $L : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$. The possible extended-real-valuedness of g and L serves in the modeling of the constraints and dynamics involved in this propagation, such as restrictions on $x(0)$ and on $\dot{x}(t)$ relative to $x(t)$. The minimization takes place over the arc space $\mathcal{A}^1_n[0,\tau]$, in the general notation that $\mathcal{A}^p_n[\tau_0,\tau_1]$ consists of all absolutely continuous $x(\cdot) : [\tau_0,\tau_1] \to \mathbb{R}^n$ with derivative $\dot{x}(\cdot) \in \mathcal{L}^p_n[\tau_0,\tau_1]$.

Value functions of the "cost-to-go" type, which propagate a terminal cost function backward from a time T, are covered by (1.1) through time reversal; this is the usual setting in optimal control. The fact that problems in optimal control can be treated in terms of an integral functional as in (1.1) for a choice of an extended-real-valued Lagrangian L has been recognized since [1] and has long been the subject of developments in nonsmooth optimization; for more on how control fits in, see, e.g., [2], [3], [4]. This is parallel to, and subsumes, the notion that differential equations with controls can be treated in terms of differential inclusions with the controls suppressed. Value functions are of interest in optimal control especially because of potential connections with feedback rules.

*Received by the editors October 2, 1998; accepted for publication (in revised form) July 21, 2000; published electronically December 28, 2000.
http://www.siam.org/journals/sicon/39-5/34536.html

†Department of Mathematics, Box 354350, University of Washington, Seattle, WA 98195-4350 (rtr@math.washington.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9803089.

‡Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918 (wolenski@marais.math.lsu.edu). The research of this author was supported by grant DMS-9623406.

An important issue in Hamilton–Jacobi theory is the extent to which V can be characterized in terms of the Hamiltonian function $H : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ associated with L, as defined through the Legendre–Fenchel transform by

$$H(x, y) := \sup_v \Big\{ \langle v, y\rangle - L(x, v)\Big\}. \tag{1.2}$$

Under the properties of this transform, $H(x, y)$ is sure to be convex in y. When $L(x, v)$ is convex, proper, and lower semicontinuous in v, as is natural for the existence of optimal arcs in (1.1), the reciprocal formula holds that

$$L(x, v) = \sup_y \Big\{ \langle v, y\rangle - H(x, y)\Big\}, \tag{1.3}$$

so L and H are completely dual to each other.

It is well recognized that a function V given by (1.1) can fail to be smooth despite any degree of smoothness of g and L, or for that matter, H. Much of modern Hamilton–Jacobi theory has revolved around this fact, especially in coming up with generalizations of the Hamilton–Jacobi PDE that might pin down V, which of course was the historical motivation for that equation. Except for the case in which H is independent of x, little attention has been paid to ascertaining circumstances in which $V(\tau, \xi)$ is convex in ξ for each $\tau \geq 0$, and to exploring the consequences of such convexity. The convex case merits study for several reasons, however.

Convexity is a crucial marker in classifying optimization problems, and it's often accompanied by interesting phenomena of duality. It can provide powerful support in matters of computation and approximation. Moreover, it has a prospect here of enabling V to be characterized via H in other ways, complementary to the Hamilton–Jacobi PDE, such as versions of the method of characteristics in which convex analysis can be brought to bear. Efforts in the convex case could therefore shed light on topics in nonsmooth Hamilton–Jacobi theory that so far have been overshadowed by PDE extensions.

The convexity of $V(\tau, \xi)$ in ξ entails, for $\tau = 0$, the convexity of the initial function g, but what does it need from the Lagrangian L? The simplest, and in a certain sense the only robust assumption for this is the joint convexity of $L(x, v)$ in x and v, which corresponds under (1.2) and (1.3) to pairing the natural convexity of $H(x, y)$ in y with the concavity of $H(x, y)$ in x. This is what we work with, along with mild conditions of semicontinuity and growth that can readily be dualized.

In optimal control, problems of convex type have roughly the same status within general control theory that linear differential equations have in the general theory of differential equations. They form the backbone for many control applications, covering traditional linear-quadratic control and its modifications to incorporate constraints and penalties (cf. [5]), but also numerous problem models in areas such as economics and operations research.

From the technical standpoint, our convexity assumptions ensure that the optimization problem appearing in (1.1) fits the theory of generalized problems of Bolza of convex type as developed in Rockafellar [1], [6], [7], [8]. That duality theory, dating from the early 1970s and based entirely on convex analysis [6], hasn't previously been utilized in the Hamilton–Jacobi setting. It had to wait for advances toward handling robustly, by means of subgradients, not only the convexity of $V(\tau, \xi)$ in ξ but also its nonconvexity in (τ, ξ). Such advances have since been made through the labor of many researchers, and the time is therefore ripe for investigating the Hamilton–Jacobi aspects of convexity and duality beyond the very special Hopf–Lax case treated in the past, where L and H don't depend on the x argument.

Relying on the background of variational analysis in [10], we make progress in several ways. We demonstrate the existence of a dual value function $\tilde{V}$, propagated by a dual Lagrangian $\tilde{L}$, such that the convex functions $V(\tau, \cdot)$ and $\tilde{V}(\tau, \cdot)$ are conjugate to each other under the Legendre–Fenchel transform for every τ. We use this in particular to derive a subgradient Hamilton–Jacobi equation satisfied directly by V, and a dual one for $\tilde{V}$, despite the unboundedness of these functions and their pervasive ∞ values. At the same time we establish a new subgradient form of the "method of characteristics" for determining these functions from the Hamiltonian H.

Central to our approach is a generalized Hamiltonian ODE associated with H, which is actually a differential inclusion in terms of subgradients instead of gradients. By focusing on $V_\tau = V(\tau, \cdot)$ as a convex function on $\mathbb{R}^n$ that varies with τ, we bring to light the remarkable fact that the graph of the subgradient mapping ∂V_τ evolves through nothing more nor less than its "drift" in the (set-valued) flow in $\mathbb{R}^n \times \mathbb{R}^n$ induced by this generalized Hamiltonian dynamical system.

Our treatment of V, although limited to the convex case, contrasts with other work in Hamilton–Jacobi theory which, in coping with ∞ values, has required $H(x, y)$ to be a special kind of *globally* Lipschitz continuous, convex function of y for each x; see Frankowska [11], [12] and Clarke et al. [13], where ∞ is admitted directly, or Bardi and Capuzzo-Dolcetta [14, Chapter V, section 5], where ∞ is suppressed by nonlinear rescaling (a maneuver incompatible with maintaining convexity). These authors take $H(x, y)$ to be positively homogeneous in y, but a standard trick (passing from Lipschitzian running costs to a Mayer formulation) allows extension to a somewhat broader class of Hamiltonians (of unknown characterization).

While the interior of the set of points where $V < \infty$ could be empty, we prove that if it isn't, then properties of semidifferentiability, Clarke regularity, and local Lipschitz continuity hold for V on that open set under our assumptions. Also, we identify through duality the situations in which coercivity or global finiteness is preserved for all $\tau > 0$.

For simplicity and to illuminate clearly the new features stemming from convexity, we keep to the case of a time-independent Lagrangian L, although extensions of the results to accommodate time dependence ought to be possible.

2. Hypotheses and main results.

In formulating the conditions that will be invoked throughout this paper, we abbreviate lower semicontinuous by "lsc" and refer to an extended-real-valued function as *proper* when it's not the constant function ∞ yet nowhere takes on $-\infty$. Thus, a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper if and only if its effective domain $\operatorname{dom} f := \{v \mid f(v) < \infty\}$ is nonempty and, on this set, f is finite. Equivalently, f is proper if and only if its epigraph, the set $\operatorname{epi} f := \{(v, s) \mid s \in \mathbb{R},\ f(v) \le s\}$, is nonempty and contains no (entire) vertical lines. Convexity of f corresponds to the convexity of $\operatorname{epi} f$, while lower semicontinuity of f corresponds to the closedness of $\operatorname{epi} f$. Convexity of f implies convexity of $\operatorname{dom} f$, but lower semicontinuity of f need not entail closedness of $\operatorname{dom} f$ (as for instance when $f(v)$ approaches ∞ as v approaches the boundary of $\operatorname{dom} f$ from within).

We denote the Euclidean norm by $|\cdot|$ and call f *coercive* when it is bounded from below and has $f(v)/|v| \to \infty$ as $|v| \to \infty$. Coercivity of a proper nondecreasing function θ on $[0, \infty)$ means that $\theta(s)/s \to \infty$ as $s \to \infty$. For a proper convex function f on $\mathbb{R}^n$, coercivity is equivalent to the finiteness of the conjugate convex function f^* on $\mathbb{R}^n$ under the Legendre–Fenchel transform, $f^*(y) := \sup_v \{\langle v, y\rangle - f(v)\}$.

Basic assumptions (A).

(A0) The initial function g is convex, proper, and lsc on $\mathbb{R}^n$.

(A1) The Lagrangian function L is convex, proper, and lsc on $\mathbb{R}^n \times \mathbb{R}^n$.

(A2) The set $F(x) := \operatorname{dom} L(x, \cdot)$ is nonempty for all x, and there is a constant ρ such that $\operatorname{dist}(0, F(x)) \leq \rho\big(1 + |x|\big)$ for all x.

(A3) There are constants α and β and a coercive, proper, nondecreasing function θ on $[0, \infty)$ such that $L(x, v) \geq \theta\big(\max\{0, |v| - \alpha|x|\}\big) - \beta|x|$ for all x and v.

The joint convexity of L with respect to x and v in (A1) contrasts with the more common assumption of convexity merely with respect to v. It is vital to our duality-based methodology. In combination with the convexity in (A0), it ensures that the functional

$$J_\tau\big(x(\cdot)\big) \;:=\; g\big(x(0)\big) + \int_0^\tau L\big((x(t), \dot{x}(t)\big)dt \tag{2.1}$$

is convex on $\mathcal{A}_n^1[0, \tau]$. It also, as a side benefit, guarantees that J_τ is well defined. That follows because $L(x(t), \dot{x}(t))$ is measurable in t when L is lsc, whereas L majorizes at least one affine function on $\mathbb{R}^n \times \mathbb{R}^n$ through its convexity and properness. Then there exist $(w, y) \in \mathbb{R}^n \times \mathbb{R}^n$ and $c \in \mathbb{R}$ with $L(x(t), \dot{x}(t)) \geq \langle x(t), w\rangle + \langle \dot{x}(t), y\rangle - c$, the expression on the right being summable in t. The integral thus has an unambiguous value in $(-\infty, \infty]$, and so then does $J_\tau(x(\cdot))$.

In (A2), the mapping F gives the differential inclusion that's implicit in the Lagrangian L. Obviously $J_\tau(x(\cdot)) = \infty$ unless the arc $x(\cdot)$ satisfies the constraints

$$\dot{x}(t) \in F(x(t)) \text{ almost everywhere (a.e.) } t, \text{ with } x(0) \in D := \operatorname{dom} g. \tag{2.2}$$

Note that the graph of F, which is the set $\operatorname{dom} L \subset \mathbb{R}^n \times \mathbb{R}^n$, is convex by (A1), although not necessarily closed. Similarly, the initial set D in these implicit constraints is convex by (A0), but need not be closed. Of course, in the special case where L is finite everywhere, the graph of F is all of $\mathbb{R}^n \times \mathbb{R}^n$ and the condition $\dot{x}(t) \in F(x(t))$ trivializes; likewise, if g is finite everywhere, the condition $x(0) \in D$ trivializes.

The nonempty-valuedness of F in (A2) means that there are no state constraints implicitly imposed by L. State constraints are definitely of interest in some applications, but in order to handle them we would have to pass from our duality framework of absolutely continuous trajectories to one in which dual trajectories or perhaps even primal trajectories might have to be merely of bounded variation; cf. [15], [16], [17]. That could be possible, but the technical complications would be more formidable and additional groundwork might have to be laid, so we forgo such an extension for now.

The growth condition in (A2) will be seen to imply that the differential inclusion in (2.2) has no "forced escape time": from any point it provides at least one trajectory over the infinite time interval $[0, \infty)$. The nonemptiness of $F(x)$ didn't really have to be mentioned separately from this growth condition, inasmuch as the distance to $\emptyset$ is ∞.

The function $L(x, \cdot)$ on $\mathbb{R}^n$, which for each x is convex by (A1) and proper by (A2), is coercive under the growth condition in (A3). Note that this growth condition is much weaker than the commonly imposed Tonelli-type condition in which $L(x, v) \geq \theta(|v|)$ for a coercive, proper, nondecreasing function θ. For instance, it covers the case of $L(x, v) = L_0(v - Ax) + L_1(x)$ for coercive L_0 and a function L_1 that does not go down to $-\infty$ at more than a linear rate, whereas the Tonelli-type condition would not do that unless $A = 0$ and L_1 is bounded from below.

The following consequence of our assumptions sets the stage for our analysis of the value function V as giving a "continuously moving" convex function on $\mathbb{R}^n$.

Theorem 2.1 (value function convexity and epi-continuity). *Under* (A), *the function $V_\tau = V(\tau, \cdot)$ is proper, lsc, and convex on $\mathbb{R}^n$ for each $\tau \in [0, \infty)$. Moreover, V_τ depends epi-continuously on τ. In particular, V is proper and lsc as a function on $[0, \infty) \times \mathbb{R}^n$, and V_τ epi-converges to g as $\tau \searrow 0$.*

This theorem will be proved in section 5. The epi-continuity in its statement refers to the continuity of the set-valued mapping $\tau \mapsto \operatorname{epi} V_\tau$ with respect to Painlevé–Kuratowski set convergence. It amounts to the following assertion (here, as elsewhere in this paper, we consistently use superscript $\nu = 1, 2, \ldots \to \infty$ in describing sequences):

$$\text{whenever } \tau^\nu \to \tau \text{ with } \tau^\nu \geq 0, \text{ one has}$$
$$\left\{\begin{array}{ll} \liminf_\nu V(\tau^\nu, \xi^\nu) \geq V(\tau, \xi) & \text{for every sequence } \xi^\nu \to \xi, \\ \limsup_\nu V(\tau^\nu, \xi^\nu) \leq V(\tau, \xi) & \text{for some sequence } \xi^\nu \to \xi, \end{array}\right. \tag{2.3}$$

where the first limit property is the lower semicontinuity of V on $[0, \infty) \times \mathbb{R}^n$. An exposition of the theory of epi-convergence of functions on $\mathbb{R}^n$ is available in Chapter 7 of [10].

Observe that the epi-convergence in Theorem 2.1 answers the question of how the initial condition $V_0 = g$ should be coordinated with the behavior of V when $\tau > 0$. Pointwise convergence of V_τ to V_0 as $\tau \searrow 0$ isn't a suitable property for a context of semicontinuity and extended-real-valuedness.

Epi-convergence has implications also for the subgradients of the functions V_τ. Recall that for a proper convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point x, a vector $y \in \mathbb{R}^n$ is a *subgradient in the sense of convex analysis* if

$$f(x') \;\geq\; f(x) + \langle y, \, x' - x \rangle \quad \text{for all } x' \in \mathbb{R}^n. \tag{2.4}$$

The set of such subgradients is denoted by $\partial f(x)$. (This is, in particular, empty when $x \notin \operatorname{dom} f$ but nonempty when $x \in \operatorname{ri} \operatorname{dom} f$, the relative interior of the convex set $\operatorname{dom} f$; see [9], [10].) The *subgradient mapping* $\partial f : x \mapsto \partial f(x)$ has graph

$$\operatorname{gph} \partial f := \big\{ (x, y) \,\big|\, y \in \partial f(x) \big\} \subset \mathbb{R}^n \times \mathbb{R}^n. \tag{2.5}$$

When f is lsc as well as proper and convex, ∂f is a maximal monotone mapping, and $\operatorname{gph} \partial f$ is therefore a globally Lipschitzian manifold of dimension n in $\mathbb{R}^n \times \mathbb{R}^n$; see [10, Chapter 12]. Furthermore, epi-convergence of functions corresponds in this picture to graphical convergence of their subgradient mappings, i.e., Painlevé–Kuratowski set convergence of their graphs; [10, 12.35].

Corollary 2.2 (subgradient manifolds). *Under* (A), *the graph of the subgradient mapping ∂V_τ is, for each $\tau \in [0, \infty)$, a globally Lipschitzian manifold of dimension n in $\mathbb{R}^n \times \mathbb{R}^n$. Moreover this set* $\operatorname{gph} \partial V_\tau$ *depends continuously on τ.*

The epigraphical continuity in the motion of V_τ in Theorem 2.1 thus corresponds to continuity graphically in the motion of ∂V_τ. Not just "continuous" aspects of this motion, but "differential" aspects need to be understood, however. For that purpose the Hamiltonian function H in (1.2) is an indispensable tool.

A better grasp of the nature of H under our assumptions is essential. Because $L(x, \cdot)$ is lsc, proper, and convex under (A1) and (A2), the reciprocal formula in (1.3) does hold, and every property of L must accordingly have some exact counterpart for H. The following fact will be verified in section 3. It describes the class of functions

H such that, when L is defined from H by (1.3), L will be the unique Lagrangian for which (A1), (A2), and (A3) hold, and for which H is the associated Hamiltonian expressed by (1.2).

THEOREM 2.3 (identification of the Hamiltonian class). *A function $H : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is the Hamiltonian for a Lagrangian L satisfying* (A1), (A2), *and* (A3) *if and only if $H(x, y)$ is everywhere finite, concave in x, convex in y, and the following growth conditions hold, where* (a) *corresponds to* (A3), *and* (b) *corresponds to* (A2)*:*

(a) *There are constants α and β and a finite, convex function φ such that*

$$H(x, y) \ \le \ \varphi(y) + (\alpha|y| + \beta)|x| \quad \textit{for all } x,\ y.$$

(b) *There are constants γ and δ and a finite, concave function ψ such that*

$$H(x, y) \ \ge \ \psi(x) - (\gamma|x| + \delta)|y| \quad \textit{for all } x,\ y.$$

The finite concavity-convexity in Theorem 2.3 implies that H is locally Lipschitz continuous; cf. [9, section 35].

Concave-convex Hamiltonian functions first surfaced as a significant class in connection with generalized problems of Bolza and Lagrange of convex type; cf. [6]. In the study of such problems, a subgradient form of Hamiltonian dynamics turned out to be crucial in characterizing optimality. Only subgradients of convex analysis are needed in expressing such dynamics. The generalized Hamiltonian *system* is

$$\dot{x}(t) \in \partial_y H(x(t), y(t)), \qquad -\dot{y}(t) \in \tilde{\partial}_x H(x(t), y(t)), \tag{2.6}$$

with $\partial_y H(x, y)$ the usual set of "lower" subgradients of the convex function $H(x, \cdot)$ at y, but $\tilde{\partial}_x H(x, y)$ the analogously defined set of "upper" subgradients of the concave function $H(\cdot, y)$ at x. A Hamiltonian *trajectory* over $[\tau_0, \tau_1]$ is an arc $(x(\cdot), y(\cdot)) \in \mathcal{A}^1_{2n}[\tau_0, \tau_1]$ that satisfies (2.6) for almost every t. The associated Hamiltonian *flow* is the one-parameter family of (generally) set-valued mappings S_τ for $\tau \ge 0$ defined by

$$S_\tau(\xi_0, \eta_0) := \big\{(\xi, \eta) \,\big|\, \exists \text{ Hamiltonian trajectory over } [0, \tau] \text{ from } (\xi_0, \eta_0) \text{ to } (\xi, \eta)\big\}. \tag{2.7}$$

Details and alternative expressions of the dynamics in (2.6) will be worked out in section 6. Appropriate extensions to nonsmooth Hamiltonians $H(x, y)$ that aren't concave in x, and thus correspond to Lagrangians $L(x, v)$ that aren't jointly convex in x and v, can be found in [3], [18], [19], and [20]. Hamiltonian trajectories are featured as necessary conditions in these works, but as will be recalled in Theorem 4.1 below, our assumptions yield symmetric relationships between the x and y elements. Here, we confine ourselves to stating how, under our assumptions, the graph of the subgradient mapping ∂V_τ, namely

$$\operatorname{gph} \partial V_\tau := \big\{(\xi, \eta) \,\big|\, \eta \in \partial V_\tau(\xi)\big\} \subset \mathbb{R}^n \times \mathbb{R}^n, \tag{2.8}$$

evolves through such dynamics from the graph of the subgradient mapping $\partial V_0 = \partial g$.

THEOREM 2.4 (Hamiltonian evolution of subgradients). *Under* (A)*, one has $\eta \in \partial V_\tau(\xi)$ if and only if, for some $\eta_0 \in \partial g(\xi_0)$, there is a Hamiltonian trajectory $(x(\cdot), y(\cdot))$ over $[0, \tau]$ with $(x(0), y(0)) = (\xi_0, \eta_0)$ and $(x(\tau), y(\tau)) = (\xi, \eta)$. Thus, the graph of ∂V_τ is the image of the graph of ∂g under the flow mapping S_τ:*

$$\operatorname{gph} \partial V_\tau = S_\tau(\operatorname{gph} \partial g) \quad \textit{for all } \tau \ge 0. \tag{2.9}$$

It will be shown in Theorem 6.3 that in the circumstances of Theorem 2.4, $x(\cdot)$ is an optimal trajectory for the minimization problem defining $V(\tau, \xi)$ in (1.1). At the same time, $y(\cdot)$ is optimal for a certain dual problem, and such optimality of $x(\cdot)$ and $y(\cdot)$ is actually equivalent to the condition in Theorem 2.4.

Theorem 2.4 is the basis for a generalized *method of characteristics* for determining V uniquely from g and H. It will be proved in section 6, where the method will be laid out in full. Especially noteworthy is the global nature of the complete description in Theorem 2.4, which is a by-product of convexity and underscores why the convex case deserves special attention. The classical method of characteristics (which requires the continuous differentiability of g and H) gives an equivalent description of V satisfying the Hamilton–Jacobi equation, but is valid only locally.

Subbotin [21] pioneered a global characteristic method for quite general nonlinear problems by introducing a (nonunique) characteristic *inclusion*, the weak invariance (viability) of which he used to define the concept of a *minimax* solution to the Hamilton–Jacobi equation. In such a general situation, *one* solution of the differential inclusion plays the role of a characteristic trajectory, whereas under our convexity assumptions, *every* solution of (2.6) plays such a role. Further recent work in generalized characteristics for nonlinear first order PDEs can be found in [22] and [23].

To go from the characterization in Theorem 2.4 to a description of the motion of V_τ in terms of a generalized Hamilton–Jacobi PDE, we need to bring in subgradients beyond those of convex analysis. The notation and terminology of the book [10] will be adopted.

Consider any function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and let x be any point at which $f(x)$ is finite. A vector $y \in \mathbb{R}^n$ is a *regular subgradient* of f at x, written $y \in \hat{\partial} f(x)$, if

$$f(x') \;\geq\; f(x) + \langle y,\, x' - x\rangle + o(|x' - x|). \tag{2.10}$$

It is a (*general*) *subgradient* of f at x, written $y \in \partial f(x)$, if there is a sequence of points $x^\nu \to x$ with $f(x^\nu) \to f(x)$ for which regular subgradients $y^\nu \in \hat{\partial} f(x^\nu)$ exist with $y^\nu \to y$.

These definitions refer to "lower" subgradients, which are usually all that we need. To keep the notation uncluttered, we take "lower" for granted, and in the few situations where "upper" subgradient sets (analogously defined) are called for, we express them by

$$\tilde{\partial} f(x) = -\partial[-f](x), \qquad \tilde{\hat{\partial}} f(x) = -\hat{\partial}[-f](x). \tag{2.11}$$

For a convex function f, $\hat{\partial} f(x)$ and $\partial f(x)$ reduce to the subgradient set defined earlier through (2.4). In the case of the value function V, the "partial subgradient" notation

$$\partial_\xi V(\tau, \xi) = \big\{\eta \,\big|\, \eta \in \partial V_\tau(\xi)\big\} \quad \text{for } V_\tau = V(\tau, \cdot)$$

can thus, through Theorem 2.1, be interpreted equally in any of the senses above.

THEOREM 2.5 (generalized Hamilton–Jacobi equation). *Under* (A)*, the subgradients of V on $(0, \tau) \times \mathbb{R}^n$ have the property that*

$$\begin{aligned} (\sigma, \eta) \in \partial V(\tau, \xi) \quad &\Longleftrightarrow \quad (\sigma, \eta) \in \hat{\partial} V(\tau, \xi) \\ &\Longleftrightarrow \quad \eta \in \partial_\xi V(\tau, \xi), \;\; \sigma = -H(\xi, \eta). \end{aligned} \tag{2.12}$$

In particular, therefore, V satisfies the generalized Hamilton–Jacobi equation

$$\sigma + H(\xi, \eta) = 0 \quad \textit{for all} \quad (\sigma, \eta) \in \partial V(\tau, \xi) \quad \textit{when} \quad \tau > 0. \tag{2.13}$$

This theorem will be proved in section 7. By the first equivalence in (2.12), the equation in (2.13) could be stated with $\hat{\partial}V(\tau,\xi)$ in place of $\partial V(\tau,\xi)$ (or in terms of the proximal subgradients emphasized in the book of Clarke, Ledyaev, Stern, and Wolenski [24]), but we prefer the ∂V version because general subgradients dominate in the variational analysis and subdifferential calculus in [10]. The $\hat{\partial}V$ version would effectively turn (2.13) into the one-sided "viscosity" form of Hamilton–Jacobi equation used for lsc functions by Barron and Jensen [25] and Frankowska [12], in distinction to earlier forms for continuous functions that rested on pairs of inequalities; cf. Crandall, Evans, and Lions [26]. The book of Bardi and Capuzzo-Dolcetta [14] gives a broad view of viscosity theory in its current state, including the relationships between such different forms. A Hamilton–Jacobi equation is called a Hamilton–Jacobi–*Bellman* equation when H is expressed by the max in (1.2).

The extent to which (2.13) (or its viscosity version) and the initial condition on V_0 might suffice to determine V uniquely isn't fully understood yet in the framework of lsc solutions that can take on ∞ when $\tau > 0$. So far, the strongest result directly available in such a framework is the one obtained by Frankowska [12]; for problems satisfying our convexity assumptions, it covers only the case where $L(x,v)$ is the indicator $\delta_C(v-Ax)$ corresponding to a differential inclusion $\dot{x}(t) \in Ax(t)+C$ for some matrix A and nonempty, compact, convex set C. Through a Mayer reformulation, her result could be made to cover the case where a finite, convex function of (x,v) is added to this indicator. How far one could go by such reformulation—and nonlinear rescaling to get rid of ∞—with the results presented by Bardi and Capuzzo-Dolcetta [14, Chapter V, section 5] is unclear.

The arcs $y(\cdot)$ that are paired with the arcs $x(\cdot)$ in the Hamiltonian dynamics are related to the forward propagation of the conjugate initial function g^*, satisfying

$$g^*(y) := \sup_x\Big\{\langle x,y\rangle - g(x)\Big\}, \qquad g(x) := \sup_y\Big\{\langle x,y\rangle - g^*(y)\Big\}, \tag{2.14}$$

with respect to the *dual* Lagrangian $\tilde{L}$, satisfying

$$\begin{aligned} \tilde{L}(y,w) = L^*(w,y) &= \sup_{x,v}\Big\{\langle x,w\rangle + \langle v,y\rangle - L(x,v)\Big\},\\ L(x,v) = \tilde{L}^*(v,x) &= \sup_{y,w}\Big\{\langle x,w\rangle + \langle v,y\rangle - \tilde{L}(y,w)\Big\}. \end{aligned} \tag{2.15}$$

The reciprocal formulas here follow from (A0) and (A1). We'll prove in section 5 that the value function $\tilde{V}$ defined as in (1.1), but with g^* and $\tilde{L}$ in place of g and L, has $\tilde{V}_\tau$ conjugate to V_τ for every τ. This duality will be a workhorse in our analysis of other basic properties.

An advantage of our assumptions (A) is that they carry over symmetrically to the dual setting. Alternative assumptions could fail in that respect. To put this another way, the class of Hamiltonians that we work with, as described in Theorem 2.3, is no accident, but carefully tuned to obtaining the broadest possible results of duality in Hamilton–Jacobi theory (here in the time-independent case).

3. Elaboration of the convexity and growth conditions. Conditions (A1), (A2), and (A3) can be viewed from several different angles, and a better understanding of them is required before we can proceed. Their Hamiltonian translation in Theorem 2.3 has to be verified, but also they will be useful as applied to functions other than L, so a broader, not merely Lagrangian, perspective on them must be attained.

We'll draw on some basic concepts of variational analysis, and convex analysis in particular. For any nonempty subset $C \subset \mathbb{R}^n$, the *horizon cone* is the closed cone

$$C^\infty := \limsup_{\lambda \searrow 0} \lambda C = \big\{ w \in \mathbb{R}^n \,\big|\, \exists\, x^\nu \in C,\, \lambda^\nu \searrow 0, \text{ with } \lambda^\nu x^\nu \to w \big\}.$$

When C is convex, C^∞ is convex and, for any $\bar{x} \in \operatorname{ri} C$ (the relative interior of C) it consists simply of the vectors w such that $\bar{x} + \lambda w \in C$ for all $\lambda > 0$. When C is convex and closed, C^∞ coincides with the "recession cone" of C. See [9, section 6], [10, Chapter 3].

For any function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $f \not\equiv \infty$, the *horizon function* f^∞ is the function having as its epigraph the set $(\operatorname{epi} f)^\infty$, where $\operatorname{epi} f$ is the epigraph of f itself. This function is always lsc and positively homogeneous. When f is convex, f^∞ is convex as well and, for any $\bar{x} \in \operatorname{ri}(\operatorname{dom} f)$, is given by $f^\infty(w) = \lim_{\lambda\to\infty} f(\bar{x} + \lambda w)/\lambda$. When f is convex and lsc, f^∞ is the "recession function" of f in convex analysis. Again, see [9, section 6], [10, Chapter 3].

It will be important in the context of conditions (A1), (A2), and (A3) to view L not just as a function on $\mathbb{R}^n \times \mathbb{R}^n$ but in terms of the associated function-valued mapping $x \mapsto L(x,\cdot)$ that assigns to each $x \in \mathbb{R}^n$ the function $L(x,\cdot) : \mathbb{R}^n \to \overline{\mathbb{R}}$. A function-valued mapping is a "bifunction" in the terminology of [9].

DEFINITION 3.1 (regular convex bifunctions). *A function-valued mapping from $\mathbb{R}^n$ to the space of extended-real-valued functions on $\mathbb{R}^n$, as specified in the form $x \mapsto \Lambda(x,\cdot)$ by a function $\Lambda : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$, is called a regular convex bifunction if*

(a1) *Λ is proper, lsc, and convex as a function on $\mathbb{R}^n \times \mathbb{R}^n$;*

(a2) *for each $w \in \mathbb{R}^n$ there is a $z \in \mathbb{R}^n$ with $(w,z) \in (\operatorname{dom}\Lambda)^\infty$;*

(a3) *there is no $z \neq 0$ with $(0,z) \in \operatorname{cl}(\operatorname{dom}\Lambda^\infty)$.*

PROPOSITION 3.2 (bifunction duality). *For $\Lambda : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$, suppose that the mapping $x \mapsto \Lambda(x,\cdot)$ is a regular convex bifunction. Then for the conjugate function $\Lambda^* : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$, the mapping $y \mapsto \Lambda^*(\cdot,y)$ is a regular convex bifunction.*

Indeed, conditions (a2) *and* (a3) *of Definition* 3.1 *are dual to each other in the sense that, under* (a1)*, Λ satisfies* (a2) *if and only if Λ^* satisfies* (a3)*, whereas Λ satisfies* (a3) *if and only if Λ^* satisfies* (a2)*.*

Proof. This was shown as part of Theorem 4 of [8]; for the duality between (a2) and (a3), see the proof of that theorem. □

LEMMA 3.3 (domain selections). *For a function $\Lambda : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ satisfying condition* (a1) *of Definition* 3.1*, condition* (a2) *is equivalent to the existence of a matrix $A \in \mathbb{R}^{n\times n}$ and vectors $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^n$ such that*

$$\big(x,\, Ax + |x|a + b\big) \in \operatorname{ri}(\operatorname{dom}\Lambda) \text{ for all } x \in \mathbb{R}^n. \tag{3.1}$$

Proof. See the first half of the proof of Theorem 5 of [8] for the necessity. The sufficiency is clear because (3.1) implies $\big(x,\, Ax + |x|a\big) \in (\operatorname{dom}\Lambda)^\infty$ for all $x \in \mathbb{R}^n$. □

PROPOSITION 3.4 (Lagrangian growth characterization). *A function $L : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ satisfies* (A1)*,* (A2)*, and* (A3) *if and only if the mapping $x \mapsto L(x,\cdot)$ is a regular convex bifunction. Specifically in the context of Definition* 3.1 *with $\Lambda = L$,* (A1) *corresponds to* (a1)*, and then one has the equivalence of* (A2) *with* (a2) *and that of* (A3) *with* (a3)*.*

Proof. When $\Lambda = L$, (A1) is identical to (a1). Assuming this property now, we argue the other equivalences.

(A2) $\Rightarrow$ (a2). For any $w \in \mathbb{R}^n$ and any integer $\nu > 0$ there exists by (A2) some $v^\nu \in F(\nu w)$ with $|v^\nu| \le \rho(1+\nu|w|)$. Let $x^\nu = \nu w$ and $\lambda^\nu = 1/\nu$. We have $(x^\nu, v^\nu) \in \operatorname{dom} L = \operatorname{dom}\Lambda$ and $\lambda^\nu(x^\nu, v^\nu) = (w, (1/\nu)v^\nu)$ with $(1/\nu)|v^\nu| \le \rho(1+|w|)$. The sequence of pairs $\lambda^\nu(x^\nu, v^\nu)$ is therefore bounded in $\mathbb{R}^n \times \mathbb{R}^n$ and has a cluster point, which necessarily is of the form (w,z) for some $z \in \mathbb{R}^n$. Furthermore $(w,z) \in (\operatorname{dom}\Lambda)^\infty$ by definition. Thus, (a2) is fulfilled.

(a2) $\Rightarrow$ (A2). Applying Lemma 3.3, we get the existence of a matrix A and vectors a and b such that $Ax+|x|a+b \in F(x)$ for all x. Then $\operatorname{dist}(0,F(x)) \le |A||x|+|x||a|+|b|$, so we can get the bound in (A2) by taking $\rho \ge \max\{|b|,\ |A|+|a|\}$.

(A3) $\Rightarrow$ (a3). Let $(\bar{x},\bar{v}) \in \operatorname{ri}(\operatorname{dom} L) = \operatorname{ri}(\operatorname{dom}\Lambda)$. For any (w,z) we have $\Lambda^\infty(w,z) = \lim_{\lambda\to\infty}\Lambda(\bar{x}+\lambda w, \bar{v}+\lambda z)/\lambda$. On the basis of (A3) this yields, in the notation $[s]_+ = \max\{0,s\}$,

$$\begin{aligned}\Lambda^\infty(w,z) &\ge \lim_{\lambda\to\infty} \lambda^{-1}\big[\theta\big([|\bar{v}+\lambda z| - \alpha|\bar{x}+\lambda w|]_+\big) - \beta|\bar{x}+\lambda w|\big]\\ &= \lim_{\lambda\to\infty} \big[\lambda^{-1}\theta\big(\lambda[|\lambda^{-1}\bar{v}+z| - \alpha|\lambda^{-1}\bar{x}+w|]_+\big)\big] - \beta|\lambda^{-1}\bar{x}+w|\big]\\ &= \begin{cases} -\beta|w| & \text{if } [|z|-\alpha|w|]_+ = 0,\\ \infty & \text{if } [|z|-\alpha|w|]_+ > 0.\end{cases}\end{aligned}$$

Hence $\operatorname{dom}\Lambda^\infty \subset \big\{(w,z)\,\big|\,|z| \le \alpha|w|\big\}$. Any $(0,z) \in \operatorname{cl}(\operatorname{dom}\Lambda^\infty)$ then has $|z| \le \alpha|0|$, hence $z=0$, so (a3) holds.

(a3) $\Rightarrow$ (A3). According to Proposition 3.2, condition (a3) on the mapping $x \mapsto \Lambda(x,\cdot)$ is equivalent to condition (a2) on the mapping $y \mapsto \Lambda^*(\cdot,y)$. By Lemma 3.3, the latter provides the existence of a matrix A and vectors a and b such that

$$(Ay+|y|a+b, y) \in \operatorname{ri}(\operatorname{dom}\Lambda^*) \text{ for all } y \in \mathbb{R}^n.$$

Any convex function is continuous over the relative interior of its effective domain, so the function $y \mapsto \Lambda^*(Ay+|y|a+b,y)$ is (finite and) continuous on $\mathbb{R}^n$ (although not necessarily convex). Define the function ψ on $[0,\infty)$ by $\psi(r) = \max\big\{\Lambda^*(Ay+|y|a+b,y)\,\big|\,|y| \le r\big\}$. Then ψ is finite, continuous, and nondecreasing. Because

$$\Lambda(x,v) = \Lambda^{**}(x,v) = \sup\nolimits_{z,y}\Big\{\langle x,z\rangle + \langle v,y\rangle - \Lambda^*(z,y)\Big\}$$

under (a1), we have

$$\begin{aligned}\Lambda(x,v) &\ge \sup\nolimits_y\Big\{\langle x,\ Ay+|y|a+b\rangle + \langle v,y\rangle - \Lambda^*(Ay+|y|a+b,y)\Big\}\\ &\ge \sup\nolimits_y\Big\{-|x|(|A||y|+|y||a|+|b|) + \langle v,y\rangle - \psi(|y|)\Big\}\\ &= \sup\nolimits_y\Big\{-|x||y|(|A|+|a|) - |x||b| + |v||y| - \psi(|y|)\Big\}\\ &= -|x||b| + \sup\nolimits_{r\ge 0}\Big\{r\big[\,|v| - (|A|+|a|)|x|\,\big] - \psi(r)\Big\}\\ &= \psi^*\big([\,|v|-(|A|+|a|)|x|\,]_+\big) - |b||x|,\end{aligned}$$

where again $[s]_+ := \max\{0,s\}$. Let $\alpha = |A|+|a|$, $\beta = |b|$, and $\theta = \psi^*$ on $[0,\infty)$. Then the inequality in (A3) holds for $L = \Lambda$. The function θ has $\theta(0) = -\psi(0)$ (finite) and is the pointwise supremum of a collection of affine functions of the form $s \mapsto rs - \psi(r)$ with $r \ge 0$ and $\psi(r)$ always finite. Hence θ is convex, proper, nondecreasing and in addition has $\lim_{s\to\infty}\theta(s)/s \ge r$ for all $r \ge 0$, which implies coercivity. ∎

PROPOSITION 3.5 (Lagrangian dualization). *If the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ satisfies* (A1), (A2), *and* (A3), *then so too does the dual Lagrangian $\tilde{L} : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ in* (2.15). *Indeed,* (A1) *for L yields* (A1) *for $\tilde{L}$ and the reciprocal formula in* (2.15), *and then* (A2) *for L corresponds to* (A3) *for $\tilde{L}$, whereas* (A3) *for L corresponds to* (A2) *for $\tilde{L}$. Furthermore, the dual Hamiltonian*

$$\tilde{H}(y,x) := \sup_w \Big\{ \langle x, w\rangle - \tilde{L}(y,w) \Big\} \tag{3.2}$$

associated with $\tilde{L}$ is then related to the Hamiltonian H for L by

$$\tilde{H}(y,x) = -H(x,y). \tag{3.3}$$

Proof. Combine Proposition 3.4 with Proposition 3.2 to get the dualization of (A1), (A2), and (A3) to $\tilde{L}$. Note next that since $L(x,\cdot)$ is by (A1), (A2), and (A3) a proper, lsc, convex, and coercive function on $\mathbb{R}^n$, its conjugate function, which is $H(x,\cdot)$, is finite on $\mathbb{R}^n$. The joint convexity of $L(x,v)$ in x and v corresponds to $H(x,y)$ being not just convex in y, as always, but also concave in x; see [9, 33.3] or [10, 11.48]. For the Hamiltonian relationship in (3.3), observe through (2.15) and the formula (1.2) for H that

$$\tilde{L}(y,w) = \sup_{x,v} \Big\{ \langle x, w\rangle + \langle v, z\rangle - L(x,v) \Big\} = \sup_x \Big\{ \langle x, w\rangle + H(x,y) \Big\}. \tag{3.4}$$

Fix any y and let $h(\cdot) = -H(\cdot,y)$, noting that $h(\cdot)$ is a finite convex function on $\mathbb{R}^n$. According to (3.4), we have $\tilde{L}(y,\cdot) = h^*(\cdot)$, and from (3.3) we then have $h^{**}(\cdot) = \tilde{H}(y,\cdot)$. The finiteness and convexity of h ensures that $h^{**} = h$, so that $\tilde{H}(y,\cdot) = -H(\cdot,y)$ as claimed in (3.3). □

Proof of Theorem 2.3. Finite convex functions correspond under the Legendre–Fenchel transform to the proper convex functions that are coercive. Having $H(x,\cdot)$ be a finite convex function on $\mathbb{R}^n$ for each $x \in \mathbb{R}^n$ is equivalent therefore to having H be the Hamiltonian associated by (1.2) with a Lagrangian L such that $L(x,\cdot)$ is, for each $x \in \mathbb{R}^n$, a proper, convex function that is coercive; the function L is recovered from H by (1.3). Concavity of $H(x,y)$ in x corresponds then to joint convexity of $L(x,v)$ in x and v, as already pointed out in the proof of Proposition 3.5; see [9, 33.3] or [10, 11.48].

Thus in particular, any finite, concave-convex function H is the Hamiltonian for some Lagrangian L satisfying (A1), while on the other hand, if L satisfies (A3) along with (A1) (and therefore has $L(x,\cdot)$ always coercive), its Hamiltonian H is finite concave-convex.

It will be demonstrated next that in the case of a Lagrangian L satisfying (A1), condition (A3) is equivalent to the growth condition in (a). This will yield through the duality in Proposition 3.5 the equivalence (A2) with the growth condition in (b), and all claims will thereby be justified. Starting with (a), define $\psi(r) = \max\big\{\varphi(y) \,\big|\, |y| \leq r\big\}$ to get a finite, nondecreasing, convex function ψ on $[0,\infty)$. The inequality in (a) yields $H(x,y) \leq \psi(|y|) + (\alpha|y| + \beta)|x|$ and consequently through (1.3) that

$$\begin{aligned} L(x,v) &\geq \sup_y \Big\{ \langle v, y\rangle - \psi(|y|) - (\alpha|y| + \beta)|x| \Big\} \\ &= \sup_{r\geq 0} \sup_{|y|\leq r} \Big\{ \langle v, y\rangle - \psi(|y|) - (\alpha|y| + \beta)|x| \Big\} \\ &= \sup_{r\geq 0} \Big\{ |v|r - \psi(r) - (\alpha r + \beta)|x| \Big\} = \psi^*\big([\,|v| - \alpha|x|\,]_+\big) - \beta|x|, \end{aligned}$$

where ψ^* is coercive, proper, and nondecreasing on $[0,\infty)$. Taking $\theta = \psi^*$, we get (A3).

Conversely from (A3), where it can be assumed without loss of generality that $\alpha \geq 0$, we can retrace this pattern by estimating through (1.2) that

$$\begin{aligned} H(x,y) &\leq \sup_v \Big\{ \langle v,y\rangle - \theta\big(\big[|v| - \alpha|x|\big]_+\big) + \beta|x| \Big\} \\ &= \sup_{s\geq 0} \sup_{|v|\leq s} \Big\{ \langle v,y\rangle - \theta\big([\,|v| - \alpha|x|\,]_+\big) + \beta|x| \Big\} \\ &= \sup_{s\geq 0} \Big\{ s|y| - \theta\big([s - \alpha|x|\,]_+\big) + \beta|x| \Big\}, \end{aligned}$$

and on changing to the variable $r = s - \alpha|x|$ obtain

$$\begin{aligned} H(x,y) &\leq \sup_{r\geq -\alpha|x|} \Big\{ (r+\alpha|x|)|y| - \theta\big([r]_+\big) + \beta|x| \Big\} \\ &= \sup_{r\geq 0} \Big\{ r|y| - \theta(r) \Big\} + (\alpha|y| + \beta)|x| = \theta^*(|y|) + (\alpha|y|+\beta)|x|, \end{aligned}$$

where θ^* is finite, convex and nondecreasing. The function $\varphi(y) = \theta^*(|y|)$ is then convex on $\mathbb{R}^n$ (see [9, 15.3] or [10, 11.21]). Thus, we have the growth condition in (a). □

4. Consequences for Bolza problem duality.

The properties we have put in place for L and H lead to stronger results about duality for the generalized problems of Bolza of convex type. These improvements, which we lay out next, will be a platform for our investigation of value function duality in section 5.

The duality theory in [1] and [7], as expressed over a fixed interval $[0,\tau]$, centers (in the autonomous case) on a problem of the form

$$(\mathcal{P}) \quad \text{minimize } J\big(x(\cdot)\big) := \int_0^\tau L\big(x(t),\dot{x}(t)\big)dt + l\big(x(0),x(\tau)\big) \text{ over } x(\cdot) \in \mathcal{A}_n^1[0,\tau],$$

where the endpoint function $l : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lsc, and convex, and on the corresponding dual problem

$$(\tilde{\mathcal{P}}) \quad \text{minimize } \tilde{J}\big(y(\cdot)\big) := \int_0^\tau \tilde{L}\big(y(t),\dot{y}(t)\big)dt + \tilde{l}\big(y(0),y(\tau)\big) \text{ over } y(\cdot) \in \mathcal{A}_n^1[0,\tau],$$

where the dual endpoint function $\tilde{l} : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ is generated through conjugacy:

$$(4.1) \quad \begin{aligned} \tilde{l}(\eta,\eta') &= l^*(\eta,-\eta') = \sup_{\xi',\xi} \Big\{ \langle \eta,\xi'\rangle - \langle \eta',\xi\rangle - l(\xi',\xi) \Big\}, \\ l(\xi',\xi) &= \tilde{l}^*(\xi',-\xi) = \sup_{\eta,\eta'} \Big\{ \langle \eta,\xi'\rangle - \langle \eta',\xi\rangle - \tilde{l}(\eta',\xi) \Big\}. \end{aligned}$$

A major role in characterizing optimality in the generalized Bolza problems $(\mathcal{P})$ and $(\tilde{\mathcal{P}})$ is played by the *generalized Euler–Lagrange condition*

$$(4.2) \qquad (\dot{y}(t),y(t)) \in \partial L(x(t),\dot{x}(t)) \text{ for a.e. } t,$$

which can also be written in the dual form $(\dot{x}(t),x(t)) \in \partial\tilde{L}(y(t),\dot{y}(t))$ for a.e. t. The Euler–Lagrange conditions are known to be equivalent in turn to the *generalized*

Hamiltonian condition (2.6) being satisfied over the time interval $[0,\tau]$; cf. [6]. They act in combination with the *generalized transversality condition*

$$(y(0),-y(\tau)) \in \partial l(x(0),x(\tau)), \tag{4.3}$$

which likewise has an equivalent dual form, $(x(0),-x(\tau)) \in \partial\tilde{l}(y(0),y(\tau))$. The basic facts about optimality are the following.

THEOREM 4.1 ([1], [6] optimality conditions). *For any functions L and l that are proper, lsc, and convex on $\mathbb{R}^n\times\mathbb{R}^n$, the optimal values in $(\mathcal{P})$ and $(\tilde{\mathcal{P}})$ satisfy* $\inf(\mathcal{P}) \le -\inf(\tilde{\mathcal{P}})$. *Moreover, for arcs $x(\cdot)$ and $y(\cdot)$ in $\mathcal{A}^1_n[0,\tau]$, the following properties are equivalent:*

(a) $(x(\cdot),y(\cdot))$ *is a Hamiltonian trajectory satisfying the transversality condition;*

(b) $x(\cdot)$ *solves* $(\mathcal{P})$, $y(\cdot)$ *solves* $(\tilde{\mathcal{P}})$, *and* $\inf(\mathcal{P}) = -\inf(\tilde{\mathcal{P}})$.

Proof. Basically this is Theorem 5 of [1], but we've used Theorem 1 of [6] to translate the Euler–Lagrange condition to the Hamiltonian condition. □

Theorem 4.1 gives us the sufficiency of the Hamiltonian condition and transversality condition for optimality of arcs in $(\mathcal{P})$ and $(\tilde{\mathcal{P}})$ but not the necessity. We can get that to the extent that we are able to establish that optimal arcs do exist for these problems, and $\inf(\mathcal{P}) = -\inf(\tilde{\mathcal{P}})$. Criteria for that have been furnished in [7] in terms of certain "constraint qualifications," but this is where we can make improvements now in consequence of our working assumptions.

The issue concerns the *fundamental kernel* $E : [0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n \to \overline{\mathbb{R}}$ defined for the Lagrangian L by

$$\begin{aligned} &E(\tau,\xi',\xi) := \inf\left\{ \int_0^\tau L\big(x(t),\dot{x}(t)\big)dt \,\middle|\, x(0)=\xi',\ x(\tau)=\xi \right\}, \\ &E(0,\xi',\xi) := \begin{cases} 0 & \text{if } \xi'=\xi, \\ \infty & \text{if } \xi'\ne\xi, \end{cases} \end{aligned} \tag{4.4}$$

where the minimization is over all arcs $x(\cdot)\in\mathcal{A}^1_n[0,\tau]$ satisfying the initial and terminal conditions. At the same time it concerns the dual fundamental kernel associated with the dual Lagrangian $\tilde{L}$, namely the function $\tilde{E} : [0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n \to \overline{\mathbb{R}}$ defined by

$$\begin{aligned} &\tilde{E}(\tau,\eta',\eta) := \inf\left\{ \int_0^\tau \tilde{L}\big(y(t),\dot{y}(t)\big)dt \,\middle|\, y(0)=\eta',\ y(\tau)=\eta \right\}, \\ &\tilde{E}(0,\eta',\eta) := \begin{cases} 0 & \text{if } \eta'=\eta, \\ \infty & \text{if } \eta'\ne\eta, \end{cases} \end{aligned} \tag{4.5}$$

with the minimization taking place over $y(\cdot)\in\mathcal{A}^1_n[0,\tau]$. The constraint qualifications in [7] are stated in terms of the sets

$$C_\tau := \big\{(\xi',\xi) \,\big|\, E(\tau,\xi',\xi)<\infty\big\}, \qquad \tilde{C}_\tau := \big\{(\eta',\eta) \,\big|\, \tilde{E}(\tau,\eta',\eta)<\infty\big\}. \tag{4.6}$$

They revolve around the overlap between these sets and the sets $\operatorname{dom} l$ and $\operatorname{dom}\tilde{l}$. In this respect the next result provides vital information.

PROPOSITION 4.2 (growth of the fundamental kernel). *Suppose* (A1), (A2), *and* (A3) *hold. Then the following properties of $E(\tau,\cdot,\cdot)$ hold for all $\tau\ge 0$ and guarantee that for all ξ and ξ' the functions $E(\tau,\xi',\cdot)$ and $E(\tau,\cdot,\xi)$ are proper, lsc, convex, and coercive:*

(a) $E(\tau,\cdot,\cdot)$ *is proper, lsc, and convex on* $\mathbb{R}^n\times\mathbb{R}^n$.

(b) *There is a constant* $\rho(\tau) \in (0, \infty)$ *such that*

$$\begin{aligned} \operatorname{dist}\big(0, \operatorname{dom} E(\tau, \xi', \cdot)\big) &\le \rho(\tau)(1 + |\xi'|) \textit{ for all } \xi' \in \mathbb{R}^n, \\ \operatorname{dist}\big(0, \operatorname{dom} E(\tau, \cdot, \xi)\big) &\le \rho(\tau)(1 + |\xi|) \textit{ for all } \xi \in \mathbb{R}^n. \end{aligned}$$

(c) *There are constants* $\alpha(\tau)$*,* $\beta(\tau)$*, and a coercive, proper, nondecreasing function* $\theta(\tau, \cdot)$ *on* $[0, \infty)$ *such that*

$$\left.\begin{aligned} E(\tau, \xi', \xi) &\ge \theta\big(\tau, [\,|\xi| - \alpha(\tau)|\xi'|\,]_+\big) - \beta(\tau)|\xi'| \\ E(\tau, \xi', \xi) &\ge \theta\big(\tau, [\,|\xi'| - \alpha(\tau)|\xi|\,]_+\big) - \beta(\tau)|\xi| \end{aligned}\right\} \textit{ for all } \xi', \xi \in \mathbb{R}^n.$$

Proof. When the mapping $x \mapsto L(x, \cdot)$ is a regular convex bifunction, both of the mappings $\xi' \mapsto E(\tau, \xi', \cdot)$ and $\xi \mapsto E(\tau, \cdot, \xi)$ are regular convex bifunctions as well, for all $\tau \ge 0$. For $\tau > 0$, this was proved as part of Theorem 5 of [8]. For $\tau = 0$, it is obvious from formula (4.5). On this basis we can appeal to Proposition 3.2 for each of the three function-valued mappings. In the conditions in (a) and (b), we get separate constants to work for $E(\tau, \xi', \cdot)$ and $E(\tau, \cdot, \xi)$, but then by taking a max we can get constants that work simultaneously for both, so as to simplify the statements. □

COROLLARY 4.3 (growth of the dual fundamental kernel). *When* L *satisfies* (A1), (A2)*, and* (A3)*, the function* $\tilde{E}$ *likewise has the properties in Proposition* 4.2*.*

Proof. Apply Proposition 4.2 to $\tilde{L}$ instead of L, using the fact from Proposition 3.5 that $\tilde{L}$, like L, satisfies (A1), (A2), and (A3). □

COROLLARY 4.4 (reachable endpoint pairs). *Under* (A1)*,* (A2)*, and* (A3)*, the sets* C_τ *and* $\tilde{C}_\tau$ *in* (4.6) *have the following property for every* $\tau > 0$*:*

(a) *The image of* C_τ *under the projection* $(\xi', \xi) \mapsto \xi'$ *is all of* $\mathbb{R}^n$*. Likewise, the image of* C_τ *under the projection* $(\xi', \xi) \mapsto \xi$ *is all of* $\mathbb{R}^n$*.*

(b) *The image of* $\tilde{C}_\tau$ *under the projection* $(\eta', \eta) \mapsto \eta'$ *is all of* $\mathbb{R}^n$*. Likewise, the image of* $\tilde{C}_\tau$ *under the projection* $(\eta', \eta) \mapsto \eta$ *is all of* $\mathbb{R}^n$*.*

Proof. We get (a) from the property in Proposition 4.2(b). We get (b) then out of the preceding corollary. □

Some generalizations of the conditions in Proposition 4.2 to the case of functions E coming from Lagrangians L that are not fully convex are available in [27].

THEOREM 4.5 (strengthened duality for Bolza problems). *Consider* $(\mathcal{P})$ *and* $(\tilde{\mathcal{P}})$ *under the assumption that the Lagrangian* L *satisfies* (A1)*,* (A2)*, and* (A3)*, whereas the endpoint function* l *is proper, lsc, and convex.*

(a) *If there exists* ξ *such that* $l(\cdot, \xi)$ *is finite, or there exists* ξ' *such that* $l(\xi', \cdot)$ *is finite, then* $\inf(\mathcal{P}) = -\inf(\tilde{\mathcal{P}})$*. This value is not* ∞*, and if it also is not* $-\infty$*, there is an optimal arc* $y(\cdot) \in \mathcal{A}^1_n[0, \tau]$ *for* $(\tilde{\mathcal{P}})$*. In particular the latter holds if an optimal arc* $x(\cdot) \in \mathcal{A}^1_n[0, \tau]$ *exists for* $(\mathcal{P})$*, and in that case both* $x(\cdot)$ *and* $y(\cdot)$ *must actually belong to* $\mathcal{A}^\infty_n[0, \tau]$*.*

(b) *If there exists* η *such that* $\tilde{l}(\eta, \cdot)$ *is finite, or there exists* η' *such that* $\tilde{l}(\cdot, \eta')$ *is finite, then* $\inf(\mathcal{P}) = -\inf(\tilde{\mathcal{P}})$*. This value is not* $-\infty$*, and if it also is not* ∞*, there is an optimal arc* $x(\cdot) \in \mathcal{A}^1_n[0, \tau]$ *for* $(\mathcal{P})$*. In particular the latter holds if an optimal arc* $y(\cdot) \in \mathcal{A}^1_n[0, \tau]$ *exists for* $(\tilde{\mathcal{P}})$*, and in that case both* $x(\cdot)$ *and* $y(\cdot)$ *must actually belong to* $\mathcal{A}^\infty_n[0, \tau]$*.*

Proof. Theorem 1 of [7] will be our vehicle. The conditions referred to as (C_0) and (D_0) in the statement of that result are fulfilled in the case of a finite, time-independent Hamiltonian (cf. p. 11 of [7]), which we have here via Theorem 2.3 (already proved in section 3).

If l satisfies one of the conditions in (a), it is impossible in the face of Corollary 4.4(a) for there to exist a hyperplane that separates the convex sets $\operatorname{dom} l$ and $\operatorname{dom} E(\tau,\cdot,\cdot)$. By separation theory (cf. [9, section 11]), this is equivalent to having $\operatorname{ri} C_\tau \cap \operatorname{ri}\operatorname{dom} l \neq \emptyset$ and $\operatorname{aff} C_\tau \cup \operatorname{dom} l = \mathbb{R}^n \times \mathbb{R}^n$, where "ri" is relative interior as earlier and "aff" denotes affine hull. According to part (b) of Theorem 1 of [7], this pair of conditions guarantees that $\inf(\mathcal{P})$ and $-\inf(\tilde{\mathcal{P}})$ have a common value which is not ∞, and that if this value is also not $-\infty$, then $(\tilde{\mathcal{P}})$ has a solution $y(\cdot) \in \mathcal{A}^1_n[0,\tau]$. We know on the other hand that whenever $\inf(\mathcal{P}) < \infty$ and $(\mathcal{P})$ has a solution $x(\cdot) \in \mathcal{A}^1_n[0,\tau]$, we have $J(x(\cdot))$ finite in $(\mathcal{P})$ (because neither l nor the integral functional in (2.1) can take on $-\infty$), so that $\inf(\mathcal{P})$ is finite. It follows then from Theorem 4.1 that $x(\cdot)$ and $y(\cdot)$ satisfy the generalized Hamiltonian condition, i.e., (2.6). Because H is finite everywhere, this implies by Theorem 2 of [6] that these arcs belong to $\mathcal{A}^\infty_n[0,\tau]$. This proves (a). The claims in (b) are justified in parallel by way of Corollary 4.4(b) and part (a) of Theorem 1 of [7]. □

COROLLARY 4.6 (best-case Bolza duality). *Consider $(\mathcal{P})$ and $(\tilde{\mathcal{P}})$ under the assumption that L satisfies* (A1), (A2), *and* (A3), *whereas l is proper, lsc, and convex. Suppose l has one of the finiteness properties in Theorem* 4.5(a), *while $\tilde{l}$ has one of the finiteness properties in Theorem* 4.5(b). *Then $-\infty < \inf(\mathcal{P}) = -\inf(\tilde{\mathcal{P}}) < \infty$, and optimal arcs $x(\cdot)$ and $y(\cdot)$ exist for $(\mathcal{P})$ and $(\tilde{\mathcal{P}})$. Moreover, any such arcs must belong to $\mathcal{A}^\infty_n[0,\tau]$.*

Proof. This simply combines the conclusions in parts (a) and (b) of Theorem 4.5. □

5. Value function duality. The topic we treat next is the relationship between V and the dual value function $\tilde{V}$ generated by $\tilde{L}$ and g^*:

$$(5.1)\ \tilde{V}(\tau,\eta) := \inf\left\{ g^*\big(y(0)\big) + \int_0^\tau \tilde{L}\big(y(t),\dot{y}(t)\big)dt \,\middle|\, y(\tau)=\eta \right\}, \qquad \tilde{V}(0,\eta) = g^*(\eta),$$

where the minimum is taken over all arcs $y(\cdot) \in \mathcal{A}^1_n[0,\tau]$. Henceforth we assume (A0), (A1), (A2), and (A3) without further mention. Because $\tilde{L}$ and g^* inherit these properties from L and g, everything we prove about V automatically holds in parallel form for $\tilde{V}$.

It will be helpful for our endeavor to note that V can be expressed in terms of E. Indeed, from the definitions of V and E in (1.1) and (4.4) it's easy to deduce the rule that

$$(5.2) \qquad V(\tau,\xi) = \inf_{\xi'}\Big\{V(\tau',\xi') + E(\tau-\tau',\xi',\xi)\Big\} \ \text{ for } 0 \le \tau' \le \tau.$$

By the same token we also have, through (5.1) and (4.5), that

$$(5.3) \qquad \tilde{V}(\tau,\eta) = \inf_{\eta'}\Big\{\tilde{V}(\tau',\eta') + \tilde{E}(\tau-\tau',\eta',\eta)\Big\} \ \text{ for } 0 \le \tau' \le \tau.$$

THEOREM 5.1 (conjugacy). *For each $\tau \ge 0$, the functions $V_\tau := V(\tau,\cdot)$ and $\tilde{V}_\tau := \tilde{V}(\tau,\cdot)$ are proper and conjugate to each other under the Legendre–Fenchel transform*

$$(5.4) \qquad \tilde{V}_\tau(\eta) = \sup_\xi\Big\{\langle\xi,\eta\rangle - V_\tau(\xi)\Big\}, \qquad V_\tau(\xi) = \sup_\eta\Big\{\langle\xi,\eta\rangle - \tilde{V}_\tau(\eta)\Big\}.$$

Hence in particular, the subgradients of these convex functions are related by

$$(5.5) \qquad \eta \in \partial V_\tau(\xi) \iff \xi \in \partial\tilde{V}_\tau(\eta) \iff V_\tau(\xi) + \tilde{V}_\tau(\eta) = \langle\xi,\eta\rangle.$$

Proof. Fix $\tau > 0$ and any vector $\bar\eta \in \mathbb{R}^n$. Let $l(\xi', \xi) = g(\xi') - \langle \xi, \bar\eta\rangle$. The corresponding dual endpoint function $\tilde l$ has $\tilde l(\eta', \eta) = g^*(\eta')$ when $\eta = \bar\eta$, but $\tilde l(\eta', \eta) = \infty$ when $\eta \neq \bar\eta$. In the Bolza problems we then have

$$-\inf(\mathcal{P}) = \sup_\xi\big\{\langle \xi, \bar\eta\rangle - V(\tau, \xi)\big\}, \qquad \inf(\tilde{\mathcal{P}}) = \tilde V(\tau, \bar\eta). \tag{5.6}$$

Because $\operatorname{dom} l$ has the form $C \times \mathbb{R}^n$ for a nonempty convex set C, namely $C = \operatorname{dom} g$, the constraint qualification of Theorem 4.5(a) is satisfied, and we may conclude that $-\inf(\tilde{\mathcal{P}}) = \inf(\mathcal{P}) > -\infty$. This yields the first equation in (5.4)—in the case of $\eta = \bar\eta$—and ensures that $V_\tau \not\equiv \infty$ and $\tilde V_\tau > -\infty$ everywhere. By the symmetry between $(\tilde L, g^*)$ and (L, g), we get the second equation in (5.4) along with $\tilde V_\tau \not\equiv \infty$ and $V_\tau > -\infty$ everywhere.

The subgradient relation translates to this context a property that is known for subgradients of conjugate convex functions in general; cf. [9, 11.3]. □

Proof of Theorem 2.1. Through the conjugacy in Theorem 5.1, we see at once that V_τ is convex and lsc, and of course the same for $\tilde V_\tau$. The remaining task is to demonstrate the epi-continuity property (2.3) of V. It will be expedient to tackle the corresponding property of $\tilde V$ at the same time and appeal to the duality between V and $\tilde V$ in simplifying the arguments. By this approach and by passing to subsequences that tend to τ either from above or from below, we can reduce the challenge to proving that

$$\begin{array}{ll} \text{(a)} & \text{whenever } \tau \geq 0 \text{ and } \tau^\nu \searrow \tau, \text{ one has} \\ & \qquad \begin{cases} \limsup_\nu V(\tau^\nu, \xi^\nu) \leq V(\tau, \xi) & \text{for some sequence } \xi^\nu \to \xi, \\ \liminf_\nu \tilde V(\tau^\nu, \eta^\nu) \geq \tilde V(\tau, \eta) & \text{for every sequence } \eta^\nu \to \eta; \end{cases} \\ \text{(b)} & \text{whenever } \tau > 0 \text{ and } \tau^\nu \nearrow \tau, \text{ one has} \\ & \qquad \begin{cases} \limsup_\nu V(\tau^\nu, \xi^\nu) \leq V(\tau, \xi) & \text{for some sequence } \xi^\nu \to \xi, \\ \liminf_\nu \tilde V(\tau^\nu, \eta^\nu) \geq \tilde V(\tau, \eta) & \text{for every sequence } \eta^\nu \to \eta, \end{cases} \end{array} \tag{5.7}$$

since these "subproperties" yield by duality the corresponding ones with V and $\tilde V$ reversed.

Argument for (a) of (5.7): Fix any $\bar\tau \geq 0$ and $\bar\xi \in \operatorname{dom} V_{\bar\tau}$. We'll verify that the first limit in (a) holds for $(\bar\tau, \bar\xi)$. Take any $\hat\tau > \bar\tau$. By Corollary 4.4(a), the image of the set $C_{\hat\tau - \bar\tau} = \operatorname{dom} E(\hat\tau - \bar\tau, \cdot, \cdot)$ under the projection $(\xi', \xi) \mapsto \xi'$ contains $\bar\xi$. Hence there exists $\hat\xi$ such that $E(\hat\tau - \bar\tau, \bar\xi, \hat\xi) < \infty$. Equivalently, there is an arc $x(\cdot) \in \mathcal{A}^1_n[\bar\tau, \hat\tau]$ such that $\int_{\bar\tau}^{\hat\tau} L(x(t), \dot x(t))dt < \infty$ with $x(\bar\tau) = \bar\xi$ and $x(\hat\tau) = \hat\xi$. Then too for every $\tau \in (\bar\tau, \hat\tau)$ we have $E(\tau - \bar\tau, \bar\xi, x(\tau)) \leq \int_{\bar\tau}^{\tau} L(x(t), \dot x(t))dt < \infty$ and therefore by (5.2) that

$$V(\tau, x(\tau)) \leq V(\bar\tau, \bar\xi) + \alpha(\tau) \text{ for } \alpha(\tau) := \int_{\bar\tau}^{\tau} L(x(t), \dot x(t))dt.$$

Consider any sequence $\tau^\nu \searrow \bar\tau$ in $(\bar\tau, \hat\tau)$. Let $\xi^\nu = x(\tau^\nu)$. Then $\xi^\nu \to \bar\xi$ and we obtain

$$\limsup_\nu V(\tau^\nu, \xi^\nu) \leq \limsup_\nu \big\{V(\bar\tau, \bar\xi) + \alpha(\tau^\nu)\big\} = V(\bar\tau, \bar\xi),$$

as desired. To establish the second limit in (a) as a consequence of this, we note now that the conjugacy in Theorem 5.1 gives $\tilde V(\tau^\nu, \cdot) \geq \langle \xi^\nu, \cdot\rangle - V(\tau^\nu, \xi^\nu)$. For any $\bar\eta$ and sequence $\eta^\nu \to \bar\eta$ this yields

$$\liminf_\nu \tilde V(\tau^\nu, \eta^\nu) \geq \liminf_\nu \big\{\langle \xi^\nu, \eta^\nu\rangle - V(\tau^\nu, \xi^\nu)\big\} \geq \langle \bar\xi, \bar\eta\rangle - V(\bar\tau, \bar\xi). \tag{5.8}$$

But $\bar{\xi}$ was an arbitrary point in $\operatorname{dom} V(\bar{\tau},\cdot)$, so we get the rest of what is needed in (a):

$$\liminf_\nu \tilde{V}(\tau^\nu,\eta^\nu) \;\geq\; \sup_\xi\big\{\langle\xi,\bar{\eta}\rangle - V(\bar{\tau},\xi)\big\} \;=\; \tilde{V}(\bar{\tau},\bar{\eta}). \tag{5.9}$$

Argument for (b) of (5.7): Fix any $\bar{\tau} > 0$ and $\bar{\xi} \in \operatorname{dom} V_{\bar{\tau}}$. We'll verify that the first limit in (a) holds for $(\bar{\tau},\bar{\xi})$. Let $\varepsilon > 0$. Because $V(\bar{\tau},\bar{\xi}) < \infty$, there exists $x(\cdot) \in \mathcal{A}^1_n[0,\bar{\tau}]$ with $x(\bar{\tau}) = \bar{\xi}$ and $g(x(0)) + \int_0^{\bar{\tau}} L(x(t),\dot{x}(t))dt < V(\bar{\tau},\bar{\xi}) + \varepsilon$. Then for all $\tau \in (0,\bar{\tau})$,

$$\begin{aligned} V(\tau,x(\tau)) &\leq g(x(0)) + \int_0^\tau L(x(t),\dot{x}(t))dt \\ &\leq V(\bar{\tau},\bar{\xi}) + \varepsilon - \alpha(\tau) \text{ for } \alpha(\tau) = \int_\tau^{\bar{\tau}} L(x(t),\dot{x}(t))dt. \end{aligned}$$

Consider any sequence $\tau^\nu \nearrow \bar{\tau}$ in $(0,\bar{\tau})$. Let $\xi^\nu = x(\tau^\nu)$. Then $\xi^\nu \to \bar{\xi}$ and we have

$$\limsup_\nu V(\tau^\nu,\xi^\nu) \leq \limsup_\nu\big\{V(\bar{\tau},\bar{\xi}) + \varepsilon - \alpha(\tau^\nu)\big\} \leq V(\bar{\tau},\bar{\xi}) + \varepsilon.$$

We've constructed a sequence with $\xi^\nu \to \bar{\xi}$ with this property for arbitrary ε, so by diagonalization we can get a sequence $\xi^\nu \to \bar{\xi}$ with $\limsup_\nu V(\tau^\nu,\xi^\nu) \leq V(\bar{\tau},\bar{\xi})$, as required. Fixing such a sequence and returning to the inequality $\tilde{V}(\tau^\nu,\cdot) \geq \langle\xi^\nu,\cdot\rangle - V(\tau^\nu,\xi^\nu)$, we obtain now for every sequence $\eta^\nu \to \bar{\eta}$ that (5.8) holds, and hence by the arbitrary choice of $\bar{\xi} \in \operatorname{dom} V_{\bar{\tau}}$ that (5.9) holds as well. □

The duality theory for the Bolza problems in this setting also provides insights into the properties of the optimal arcs associated with V.

THEOREM 5.2 (optimal arcs). *In the minimization problem defining $V_\tau(\xi) = V(\tau,\xi)$, an optimal arc $x(\cdot) \in \mathcal{A}^1_n[0,\tau]$ exists for any $\xi \in \operatorname{dom} V_\tau$. Every such arc $x(\cdot)$ must actually belong to $\mathcal{A}^\infty_n[0,\tau]$ when ξ is such that $\partial V_\tau(\xi) \neq \emptyset$, hence, in particular if $\xi \in \operatorname{ri}\operatorname{dom} V_\tau$.*

Proof. Although the theorem is stated in terms of V alone, its proof will rest on the duality between V and $\tilde{V}$. We'll focus actually on proving the $\tilde{V}$ version, since that ties in better with the foundation already laid in the proof of Theorem 5.1.

Returning to the problems $(\mathcal{P})$ and $(\tilde{P})$ of that proof, we make further use of Theorem 4.5. We showed that our choice of the function l implied $\inf(\tilde{\mathcal{P}}) = -\inf(\mathcal{P}) > -\infty$ in (5.6), but we didn't point out then that it also guarantees through Theorem 4.5(a) that an optimal arc $y(\cdot)$ exists for $(\tilde{\mathcal{P}})$ if, in addition, $\inf(\tilde{P}) < \infty$. Thus, an optimal arc exists for the problem defining $\tilde{V}(\tau,\bar{\eta})$ as long as $\tilde{V}(\tau,\bar{\eta}) < \infty$. Likewise then, an optimal arc exists for the problem defining $V(\tau,\bar{\xi})$ for any $\bar{\xi}$ such that $V(\tau,\bar{\xi}) < \infty$.

Next we use the fact that a vector $\bar{\xi}$ belongs to $\partial\tilde{V}_\tau(\bar{\eta})$ if and only if $\bar{\eta} \in \operatorname{dom}\tilde{V}_\tau$ and $\bar{\xi}$ furnishes the maximum in the expression for $-\inf(\mathcal{P})$ in (5.6). (This is true by (5.4) and (5.5) of Theorem 5.1.) For such a vector $\bar{\xi}$, $V(\tau,\bar{\xi})$ has to be finite, so that there exists, by the argument already furnished, an optimal arc $x(\cdot)$ for the minimizing problem that defined $V(\tau,\bar{\xi})$. That arc $x(\cdot)$ must then be optimal for $(\mathcal{P})$. Theorem 4.5(a) tells us in that case that $x(\cdot)$ and the optimal arc $y(\cdot)$ for $(\tilde{\mathcal{P}})$ are in $\mathcal{A}^\infty_n[0,\tau]$.

To finish up, we merely need to recall that a proper convex function φ has subgradients at every point of $\operatorname{ri}\operatorname{dom}\varphi$, in particular. □

6. Hamiltonian dynamics and method of characteristics. The generalized Hamiltonian ODE in (2.6) now enters the discussion. This dynamical system can be written in the form

$$(\dot{x}(t), \dot{y}(t)) \in G(x(t), y(t)) \quad \text{for a.e. } t \tag{6.1}$$

for the set-valued mapping

$$G : (x, y) \;\mapsto\; \partial_y H(x, y) \times -\tilde{\partial}_x H(x, y), \tag{6.2}$$

which derives from the subgradient mapping $(x, y) \;\mapsto\; \tilde{\partial}_x H(x, y) \times \partial_y H(x, y)$. The latter has traditionally been associated in convex analysis with H as a concave-convex function on $\mathbb{R}^n \times \mathbb{R}^n$. It is known to be nonempty-compact-convex-valued and locally bounded with closed graph (since H is also finite; see [9, section 35]). Hence the same holds for G.

Through these properties of G, the theory of differential inclusions [28] ensures the local existence of a Hamiltonian trajectory through every point. The local boundedness of G makes any trajectory $(x(\cdot), y(\cdot))$ over a time interval $[\tau_0, \tau_1]$ be Lipschitz continuous, i.e., belong to $\mathcal{A}^\infty_n[\tau_0, \tau_1]$. Another aspect of the Hamiltonian dynamics in (2.6), or (6.1)–(6.2), is that $H(x(t), y(t))$ is constant along any trajectory $(x(\cdot), y(\cdot))$. This was proved in [6].

Nowadays there are other concepts of subgradient, beyond those of convex analysis, that can be applied to H without separating it into its concave and convex arguments. The general definition in section 2 directly assigns a subset $\partial H(x, y) \subset \mathbb{R}^n \times \mathbb{R}^n$ to each point $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$. An earlier definition for this purpose, which was used by Clarke in his work on Hamiltonian conditions for optimality in *nonconvex* problems of Bolza (see [29] and its references), relied on H being locally Lipschitz continuous and utilized what we now recognize as the set $\operatorname{con} \partial H(x, y)$ in such circumstances. (Here "con" designates the convex hull of a set.) A more subtle form of "partial convexification" of $\partial H(x, y)$, involving only the x argument in a special way, has been featured in more recent work on Hamiltonians in nonconvex problems of Bolza; cf. [18], [19], and [20].

As a preliminary to our further analysis of the Hamiltonian dynamics, we provide a clarification of the relationships between these concepts.

PROPOSITION 6.1 (subgradients of the Hamiltonian). *On the basis of $H(x, y)$ being finite, concave in x, and convex in y, one has*

$$\operatorname{con} \partial H(x, y) \;=\; \tilde{\partial}_x H(x, y) \times \partial_y H(x, y), \tag{6.3}$$

this set being nonempty and compact. In terms of the set D consisting of the points (x, y) where H is differentiable (the complement of which is of measure zero), one has

$$\operatorname{con} \partial H(x, y) = \partial H(x, y) = \{\nabla H(x, y)\} \quad \textit{for all} \quad (x, y) \in D. \tag{6.4}$$

The gradient mapping ∇H is continuous relative to D, so that H is strictly differentiable on D. Elsewhere,

$$\operatorname{con} \partial H(x, y) \;=\; \operatorname{con}\Big\{(w, v) \;\Big|\; \exists\, (x^\nu, y^\nu) \to (x, y) \;\; \textit{with} \;\; \nabla H(x^\nu, y^\nu) \to (w, v)\Big\}. \tag{6.5}$$

Proof. Formula (6.5) is well known to hold for the subgradients of any locally Lipschitz continuous function; cf. [10, 9.61]. The special property coming out of the concavity-convexity of H is that the set-valued mapping

$$T_H : (x, y) \;\mapsto\; [-\tilde{\partial}_x H(x, y)] \times \partial_y H(x, y) \;=\; \partial_x[-H](x, y) \times \partial_y H(x, y) \tag{6.6}$$

is maximal monotone; cf. [10, 12.27]. The points (x,y) where T_H is single-valued are the ones where $\tilde{\partial}_x H(x,y)$ and $\partial H_y(x,y)$ both reduce to singletons, a property which corresponds to $H(\cdot,y)$ being differentiable at x while $H(x,\cdot)$ is differentiable at y; then actually H is differentiable (jointly in the two arguments) at (x,y); cf. [9, 35.6]. Thus, the subset of $\mathbb{R}^n \times \mathbb{R}^n$ on which T_H is single-valued is D, and on this set we have $T_H(x,y) = (-\nabla_x H(x,y), \nabla_y H(x,y))$. Then by maximal monotonicity, T_H is continuous on D with

$$T_H(x,y) \;=\; \operatorname{con}\Big\{(-w,v) \,\Big|\, \exists\,(x^\nu,y^\nu) \to (x,y) \text{ with } \nabla H(x^\nu,y^\nu) \to (w,v)\Big\};$$

see [10, 12.63, 12.67]. We thereby obtain (6.3) from (6.5) and at the same time have (6.4), from which H must be strictly differentiable on D by [10, 9.18]. □

COROLLARY 6.2 (single-valuedness in the Hamiltonian system). *The mapping G in the differential inclusion* (6.1)–(6.2) *has the form*

$$G(x,y) = \big\{(v,-w) \,\big|\, (w,v) \in \operatorname{con} \partial H(x,y)\big\} \tag{6.7}$$

and is single-valued a.e. Indeed, $G(x,y) = \big\{(\nabla_y H(x,y), -\nabla_x H(x,y))\big\}$ at all points where the Hamiltonian H is differentiable, whereas in general,

$$\begin{aligned} G(x,y) \;=\; \operatorname{con}\Big\{(v,-w) \,\Big|\, &\exists\,(x^\nu,y^\nu) \to (x,y) \textit{ with} \\ &(\nabla_y H(x^\nu,y^\nu), -\nabla_x H(x^\nu,y^\nu)) \to (v,-w)\Big\}. \end{aligned} \tag{6.8}$$

Despite the typical single-valuedness of G, situations exist in which there can be more than one Hamiltonian trajectory from a given starting point. The flow mappings S_τ for this system, as defined in (2.7), can well have values that are not singleton sets, and indeed, can even be nonconvex sets consisting of more than finitely many points. It's rather surprising, then, that they nonetheless capture with precision the behavior of the Lipschitzian manifolds $\operatorname{gph} \partial V_\tau$ in Corollary 2.2. We're prepared now to prove this fact.

Proof of Theorem 2.4. Fix $\tau > 0$ along with any $\bar{\xi}$ and $\bar{\eta}$. The relation $\bar{\eta} \in \partial V_\tau(\bar{\xi})$ is equivalent by Theorem 5.1 to $\bar{\xi} \in \partial \tilde{V}_\tau(\bar{\eta})$, or to having $\bar{\xi} \in \operatorname{argmax}_\xi \big\{\langle \xi, \bar{\eta}\rangle - V_\tau(\xi)\big\}$. We saw in the proof of Theorem 5.2 that this corresponded further, in terms of the special Bolza problems $(\mathcal{P})$ and $(\tilde{\mathcal{P}})$ introduced in the proof of Theorem 5.1, to the existence of optimal arcs $x(\cdot)$ for $(\mathcal{P})$ and $y(\cdot)$ for $(\tilde{\mathcal{P}})$ such that $x(\tau) = \bar{\xi}$.

On the other hand, because $-\inf(\mathcal{P}) = (\tilde{\mathcal{P}})$ for these problems, we know from Theorem 4.1 that arcs $x(\cdot)$ and $y(\cdot)$ solve these problems, respectively, if and only if $(x(\cdot), y(\cdot))$ is a Hamiltonian trajectory over $[0,\tau]$ satisfying the generalized transversality condition $(y(0), -y(\tau)) \in \partial l(x(0), \bar{\xi})$. Since $l(\xi', \xi) = g(\xi') - \langle \xi, \bar{\eta}\rangle$ by definition in this case, the transversality condition comes down to the relations $y(0) \in \partial g(x(0))$ and $y(\tau) = \bar{\eta}$.

In summary, we have $\bar{\eta} \in \partial V_\tau(\bar{\xi})$ if and only if there is a trajectory $(x(\cdot), y(\cdot))$ over $[0,\tau]$ such that $x(\tau) = \bar{\eta}$, $y(0) \in \partial g(x(0))$, and $y(\tau) = \bar{\eta}$. □

Further details about the evolution of the subgradient mappings $\partial V_\tau = \partial_\xi V(\tau,\cdot)$ can now be recorded. The equivalence in the next theorem came out in the preceding proof.

THEOREM 6.3 (optimality in subgradient evolution). *A pair of arcs $x(\cdot)$ and $y(\cdot)$ gives a Hamiltonian trajectory over $[0,\tau]$ that starts in $\operatorname{gph} \partial g$ and ends at a point $(\xi,\eta) \in \operatorname{gph} \partial V_\tau$ if and only if*

(a) $x(\cdot)$ *is optimal in the minimization problem in* (1.1) *that defines* $V(\tau,\xi)$*, and*
(b) $y(\cdot)$ *is optimal in the minimization problem in* (5.1) *that defines* $\tilde{V}(\tau,\eta)$.

Corollary 6.4 (persistence of subgradient relations). *When a Hamiltonian trajectory* $(x(\cdot),y(\cdot))$ *over* $[0,\tau]$ *has* $y(0)\in\partial g(x(0))$*, it has* $y(t)\in\partial_\xi V(t,x(t))$ *for all* $t\in[0,\tau]$.

We turn now, however, to the task of broadening Theorem 2.4 to cover not only the evolution of subgradients but also that of function values. For this, the graph of ∂V_τ in $\mathbb{R}^n\times\mathbb{R}^n$ has to be replaced by an associated subset of $\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}$.

Proposition 6.5 (characteristic manifolds for convex functions). *Let* $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ *be convex, proper, and lsc, and let*

$$M=\big\{(x,y,z)\,\big|\,y\in\partial f(x),\ z=f(x)\big\}\subset\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}. \tag{6.9}$$

Then M *is an* n*-dimensional Lipschitzian manifold in the following terms. There is a one-to-one, locally Lipschitz continuous mapping*

$$F:\mathbb{R}^n\to M,\qquad F(u)=(P(u),Q(u),R(u)),$$

whose range is all of M *and whose inverse is Lipschitz continuous as well, in fact with*

$$F^{-1}(x,y,z)=x+y\quad\textit{for }(x,y,z)\in M.$$

The components of F *are given by*

$$P(u)=\mathrm{argmin}_x\Big\{f(x)+\tfrac{1}{2}|x-u|^2\Big\},\qquad Q=I-P,\qquad R=f{\circ}P, \tag{6.10}$$

where P *and* Q*, like* F^{-1}*, are globally Lipschitz continuous with constant* 1*, and* R *is Lipschitz continuous with constant* r *on the ball* $\big\{u\,\big|\,|u|\le r\big\}$ *for each* $r>0$.

Proof. The mapping $u\mapsto(P(u),Q(u))$ is well known to parameterize the graph of ∂f in the manner described; cf. [10, 12.15]. With this parameterization, the component $z=R(u)$ must be $f(P(u))$, so the additional issue is just the claimed Lipschitz property of this expression. According to the formulas for P and Q in (6.10) we have

$$R(u)=p(u)-\tfrac{1}{2}|Q(u)|^2\quad\text{for}\quad p(u):=\min_x\Big\{f(x)+\tfrac{1}{2}|x-u|^2\Big\}. \tag{6.11}$$

The function p is smooth with gradient $\nabla p(u)=Q(u)$; see [10, 2.26]. Hence R is locally Lipschitz continuous, but what can be said about its Lipschitz modulus? Because P and Q are Lipschitz continuous with constant 1 and satisfy $P+Q=I$, they are differentiable at almost every point u, their Jacobian matrices satisfying $\nabla P(u)+\nabla Q(u)=I$ and having norms at most 1. At any such point u, R is differentiable as well, with $\nabla R(u)=Q(u)-\nabla Q(u)Q(u)=\nabla P(u)Q(u)$, so that $|\nabla R(u)|\le|\nabla P(u)||Q(u)|\le|Q(u)|\le|u|$. Thus, $|\nabla R(u)|\le r$ on the ball $\big\{u\,\big|\,|u|\le r\big\}$, and consequently R is Lipschitz continuous with constant r on that ball. □

The set M in (6.9) will be called the (first-order) *characteristic manifold* for f, and the mapping F its *canonical parameterization.*

Proposition 6.6 (recovery of a function from its manifold). *Let* M *be the characteristic manifold of a convex, proper, lsc function* f*. Then* M *uniquely determines* f *as follows:*

(a) *The image* C *of* M *under the projection* $(x,y,z)\mapsto x$*, namely* $C=\mathrm{dom}\,\partial f$*, satisfies* $\mathrm{ri}\,\mathrm{dom}\,f\subset C\subset\mathrm{cl}\,\mathrm{dom}\,f$ *and thus has* $\mathrm{ri}\,C=\mathrm{ri}\,\mathrm{dom}\,f$ *and* $\mathrm{cl}\,C=\mathrm{cl}\,\mathrm{dom}\,f$.

(b) *For every x in C, the vectors $(x,y,z) \in M$ all have the same z, which equals $f(x)$.*

(c) *For every $x \in \operatorname{cl} C \setminus C$ and any $a \in \operatorname{ri} C$, one has $x + \varepsilon(a - x) \in \operatorname{ri} C$ for all $\varepsilon \in (0,1]$ and $f(x + \varepsilon(a-x)) \to f(x)$ as $\varepsilon \searrow 0$.*

(c) *For every $x \notin \operatorname{cl} C$, $f(x) = \infty$.*

Proof. These facts are evident from the definition of M, the well-known existence of subgradients at points of $\operatorname{ri}\operatorname{dom} f$, and the way that f can be recovered fully from its values on $\operatorname{ri}\operatorname{dom} f$; see [9, section 7, section 23]. □

PROPOSITION 6.7 (convergence of characteristic manifolds). *A sequence of convex, proper, lsc functions f^ν on $\mathbb{R}^n$ epi-converges to another such function f if and only if the associated sequence of characteristic manifolds M^ν in $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ converges (in the Painlevé–Kuratowski sense) to the characteristic manifold M for f.*

Proof. Attouch's theorem on convex functions (cf. [10, 12.35]) says that f^ν epi-converges to f if and only if $\operatorname{gph} \partial f^\nu$ converges to $\operatorname{gph} \partial f$ and, for at least one sequence of points $(x^\nu, y^\nu) \in \operatorname{gph} \partial f^\nu$ converging to a point $(x,y) \in \operatorname{gph} \partial f$, one has $f^\nu(x^\nu) \to f(x)$. On the other hand, epi-convergence of convex functions entails the latter holding for every such sequence of points (x^ν, y^ν). The convergence of the characteristic manifolds is thus hardly more than a restatement of these facts of convex analysis. □

Our goal in these terms is to describe how the characteristic manifold for V_τ evolves from that of g. We introduce the following extension of the Hamiltonian system (6.1)–(6.2), which we speak of as the *characteristic system* associated with H:

$$(\dot{x}(t), \dot{y}(t), \dot{z}(t)) \in \bar{G}(x(t), y(t)) \text{ for a.e. } t \tag{6.12}$$

for the set-valued mapping $\bar{G}$ defined by

$$\bar{G}(x,y) := \big\{ (v,w,u) \,\big|\, (v,w) \in G(x,y),\ u = \langle v, y\rangle - H(x,y) \big\}. \tag{6.13}$$

The trajectories $(x(\cdot), y(\cdot), z(\cdot))$ of this system will be called *characteristic trajectories*. Like G itself, $\bar{G}$ is nonempty-closed-convex-valued and locally bounded with closed graph, so a characteristic trajectory exists, at least locally, through every point of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$. The corresponding flow mapping for each $\tau \in [0,\infty)$ will be denoted by $\bar{S}_\tau$:

$$\begin{aligned} &\bar{S}_\tau : (\xi_0, \eta_0, \zeta_0) \mapsto \\ &\Big\{ (\xi,\eta,\zeta) \,\Big|\, \exists \text{ characteristic trajectory } (x(\cdot), y(\cdot), z(\cdot)) \text{ over } [0,\tau] \text{ with} \\ &\qquad (x(0), y(0), z(0)) = (\xi_0, \eta_0, \zeta_0),\ (x(\tau), y(\tau), z(\tau)) = (\xi, \eta, \zeta) \Big\}. \end{aligned} \tag{6.14}$$

THEOREM 6.8 (subgradient method of characteristics). *Let M_τ be the characteristic manifold for $V_\tau = V(\tau, \cdot)$, with M_0 the characteristic manifold for $g = V_0$. Then*

$$M_\tau = \bar{S}_\tau(M_0) \quad \text{for all} \quad \tau \geq 0. \tag{6.15}$$

Moreover M_τ, as a closed subset of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$, depends continuously on τ.

Proof. The continuity of the mapping $\tau \mapsto M_\tau$ is immediate from Proposition 6.7 and the epi-continuity in Theorem 2.1. The evolution of ∂V_τ through the drift of its graph in the underlying system (6.1)–(6.2) has already been verified in Theorem

2.4, so the only issue here is what happens when the z component is added as in (6.12)–(6.13). We have

$$\dot{z}(t) = \langle \dot{x}(t), y(t)\rangle - H(x(t), y(t)) = L(x(t), \dot{x}(t)) \tag{6.16}$$

when $(\dot{x}(t), \dot{y}(t)) \in G(x(t), y(t))$, since that relation entails $\dot{x}(t) \in \partial_y H(x(t), y(t))$, which is equivalent to the second equation in (6.16) because the convex functions $H(x(t), \cdot)$ and $L(x(t), \cdot)$ are conjugate to each other. The arc $x(\cdot)$ is optimal for the minimization problem that defines $V(\tau, \xi)$, so that

$$V(\tau, \xi) = g(x(0)) + \int_0^\tau L(x(t), \dot{x}(t))dt = z(0) + \int_0^\tau \dot{z}(t)dt = z(\tau).$$

The trajectory $(x(\cdot), y(\cdot), z(\cdot))$ does, therefore, carry the point $(x(0), y(0), z(0)) \in M_0$ to the point $(x(\tau), y(\tau), z(\tau)) \in M_\tau$. Conversely, of course, (6.16) is essential for that. □

Theorem 6.8 provides a remarkably global version of the method of characteristics, made possible by convexity. It relies on the one-to-one correspondence between lsc, proper, convex functions and their characteristic manifolds in Proposition 6.5 and on the preservation of such function properties over time, as in Theorem 2.1. By transforming the evolution of functions into the evolution of the associated manifolds, one is able to reduce the function evolution to the drift of those manifolds in the characteristic dynamical system associated with the given Hamiltonian H, or Lagrangian L.

In contrast, the classical method of characteristics requires differentiability at every turn and, in adopting the implicit (or inverse) function theorem as the main tool, is ordinarily limited to local validity. The characteristic manifold M_0 associated with g has to be a smooth manifold, and g must therefore be $\mathcal{C}^2$. The Hamiltonian H has to be $\mathcal{C}^2$ as well, so that the mappings $\bar{S}_\tau$ are single-valued and smooth. But even these assumptions are not enough to guarantee that the characteristic dynamics will carry M_0 into *smooth* manifolds M_τ. The trouble is that the functions V_τ are defined by minimization, and that operation, in its inherent failure to preserve differentiability, simply does not fit well in the framework of classical analysis.

A generalized "method of characteristics" for value functions has also been developed by Subbotin [30], [21], but in a different framework from ours, namely one focused on bounded control dynamics and not convexity, and not revolving around the Hamiltonian function H and its dynamical system. This is also the case in [22] and [23].

7. Hamilton–Jacobi equation and regularity. The time has come to move beyond subgradients of convex analysis and establish properties of the subgradient mapping ∂V as a whole.

Proof of Theorem 2.5. Our first goal is to prove the equivalence of the conditions $\eta \in \partial_\xi V(\tau, \xi)$ and $\sigma = -H(\xi, \eta)$ with having $(\sigma, \eta) \in \hat{\partial} V(\tau, \xi)$ when $\tau > 0$. Here $\partial_\xi V(\tau, \xi)$ is the same as $\hat{\partial}_\xi V(\tau, \xi)$, since the function $V(\tau, \cdot) = V_\tau$ is convex.

Let $\bar{\eta} \in \partial_\xi V(\bar{\tau}, \bar{\xi})$ with $\bar{\tau} > 0$. We need to show that $(-H(\bar{\xi}, \bar{\eta}), \bar{\eta}) \in \hat{\partial} V(\bar{\tau}, \bar{\xi})$, or in other words that

$$V(\tau, \xi) - V(\bar{\tau}, \bar{\xi}) + (\tau - \bar{\tau})H(\bar{\xi}, \bar{\eta}) - \langle \xi - \bar{\xi}, \bar{\eta}\rangle \ \geq\ o\big(|(\tau, \xi) - (\bar{\tau}, \bar{\xi})|\big). \tag{7.1}$$

By Theorem 2.4 there is a Hamiltonian trajectory $(x(\cdot), y(\cdot))$ over $[0, \bar{\tau}]$ that starts in $\operatorname{gph} \partial g$ and goes to $(\bar{\xi}, \bar{\eta})$. Through the local existence property of the Hamiltonian system, this trajectory can be extended to a larger interval $[0, \bar{\tau} + \varepsilon]$, in which case $y(\tau) \in \partial_\xi V(\tau, x(\tau))$ for all $\tau \in [0, \bar{\tau} + \varepsilon]$ by Corollary 6.4, so that

$$(7.2)\ \ V(\tau, \xi) \ \geq \ V(\tau, x(\tau)) + \langle \xi - x(\tau), y(\tau) \rangle \quad \text{for all } \xi \in \mathbb{R}^n \text{ when } \tau \in [0, \bar{\tau} + \varepsilon].$$

We have $V(\tau, x(\tau)) \ = \ g(x(0)) + \int_0^\tau \big[\langle \dot{x}(t), y(t) \big\rangle - H(x(t), y(t))\big] dt$ by Theorem 6.8, where $H(x(t), y(t)) \equiv H(x(\bar{\tau}), y(\bar{\tau}))$ because H is constant along Hamiltonian trajectories. Hence

$$(7.3)\ \ V(\tau, x(\tau)) = V(\bar{\tau}, \bar{\xi}) - (\tau - \bar{\tau}) H(\bar{\xi}, \bar{\eta}) + \int_{\bar{\tau}}^{\tau} \langle \dot{x}(t), y(t) \rangle dt \ \text{ when } \tau \in [0, \bar{\tau} + \varepsilon].$$

Also $\int_{\bar{\tau}}^{\tau} \langle \dot{x}(t), y(t) \rangle dt = \langle x(\tau), y(\tau) \rangle - \langle x(\bar{\tau}), y(\bar{\tau}) \rangle - \int_{\bar{\tau}}^{\tau} \langle x(t), \dot{y}(t) \rangle dt$, so in combining (7.3) with (7.2), we see that the left side of (7.1) is bounded below by the expression

$$\begin{aligned} -\langle \xi - \bar{\xi}, \bar{\eta} \rangle + \langle \xi - x(\tau), y(\tau) \rangle + \langle x(\tau), y(\tau) \rangle - \langle x(\bar{\tau}), y(\bar{\tau}) \rangle - \int_{\bar{\tau}}^{\tau} \langle x(t), \dot{y}(t) \rangle dt& \\ = \langle \xi - \bar{\xi},\, y(\tau) - \bar{\eta} \rangle + \langle \bar{\xi},\, y(\tau) - \bar{\eta} \rangle - \int_{\bar{\tau}}^{\tau} \langle x(t), \dot{y}(t) \rangle dt& \\ = \langle \xi - \bar{\xi},\, y(\tau) - y(\bar{\tau}) \rangle - \int_{\bar{\tau}}^{\tau} \langle x(t) - x(\bar{\tau}), \dot{y}(t) \rangle dt.& \end{aligned}$$

This expression is of type $o\big(|(\tau, \xi) - (\bar{\tau}, \bar{\xi})|\big)$ because $x(\cdot)$ and $y(\cdot)$ are continuous and $\dot{y}(\cdot)$ is essentially bounded on $[0, \bar{\tau} + \varepsilon]$. Thus, $(-H(\bar{\xi}, \bar{\eta}), \bar{\eta}) \in \hat{\partial} V(\bar{\tau}, \bar{\xi})$, as claimed.

To argue the converse implication, we consider now any pair $(\bar{\sigma}, \bar{\eta}) \in \hat{\partial} V(\bar{\tau}, \bar{\xi})$. Such a pair satisfies

$$(7.4) \qquad V(\tau, \xi) \ \geq \ V(\bar{\tau}, \bar{\xi}) + (\tau - \bar{\tau}) \bar{\sigma} + \langle \xi - \bar{\xi}, \bar{\eta} \rangle + o\big(|(\tau, \xi) - (\bar{\tau}, \bar{\xi})|\big).$$

In particular, $\bar{\eta} \in \hat{\partial}_\xi V(\bar{\tau}, \bar{\xi}) = \partial_\xi V(\bar{\tau}, \bar{\xi})$, and we therefore have, as just explained, the existence of a Hamiltonian trajectory $(x(\cdot), y(\cdot))$ for which (7.3) holds. Specializing (7.4) to $\xi = x(\tau)$ and using the expression in (7.3) for $V(\tau, x(\tau))$, we obtain

$$\begin{aligned} V(\bar{\tau}, \bar{\xi}) - &(\tau - \bar{\tau}) H(\bar{\xi}, \bar{\eta}) + \int_{\bar{\tau}}^{\tau} \langle \dot{x}(t), y(t) \rangle dt \\ &\geq \ V(\bar{\tau}, \bar{\xi}) + (\tau - \bar{\tau}) \bar{\sigma} + \langle x(\tau) - x(\bar{\tau}), \bar{\eta} \rangle + o\big(|(\tau, x(\tau)) - (\bar{\tau}, x(\bar{\tau}))|\big), \end{aligned}$$

where the final term is of type $o(|\tau - \bar{\tau}|)$ because $x(\cdot)$ is locally Lipschitz continuous. Then

$$(\tau - \bar{\tau})(\bar{\sigma} + H(\bar{\xi}, \bar{\eta})) \ \leq \ \int_{\bar{\tau}}^{\tau} \langle \dot{x}(t),\, y(t) - y(\bar{\tau}) \rangle dt + o(|\tau - \bar{\tau}|),$$

with the integral term likewise being of type $o(|\tau - \bar{\tau}|)$. Necessarily, then, $\bar{\sigma} + H(\bar{\xi}, \bar{\eta}) = 0$.

We turn now to showing that $\partial V(\tau, \xi) = \hat{\partial} V(\tau, \xi)$ for all ξ when $\tau > 0$. Since $\hat{\partial} V(\tau, \xi) \subset \partial V(\tau, \xi)$ in general, only the opposite inclusion has to be checked. Suppose $(\sigma, \eta) \in \partial V(\tau, \xi)$. By definition, there are sequences $(\tau^\nu, \xi^\nu) \to (\tau, \xi)$ and $(\sigma^\nu, \eta^\nu) \to (\sigma, \nu)$ with $V(\tau^\nu, \xi^\nu) \to V(\tau, \xi)$ and $(\sigma^\nu, \eta^\nu) \in \hat{\partial} V(\tau^\nu, \xi^\nu)$. We have seen that the latter means $\sigma^\nu = -H(\xi^\nu, \eta^\nu)$ and $\eta^\nu \in \partial_\xi V(\tau^\nu, \xi^\nu)$. Then $\sigma = -H(\xi, \eta)$ by the continuity of H.

On the other hand, the sets $C^\nu = \operatorname{gph} \partial_\xi V(\tau^\nu, \cdot)$ converge to $C = \operatorname{gph} \partial_\xi V(\tau, \cdot)$ by Corollary 2.2. Hence from having $\eta^\nu \in \partial_\xi V(\tau^\nu, \xi^\nu)$ we get $\eta \in \partial_\xi V(\tau, \xi)$. The pair (σ, η) thus satisfies the conditions we have identified as describing the elements of $\hat{\partial} V(\tau, \xi)$. □

Through the duality in Theorem 5.1, the statements in Theorem 2.5 are valid equally for the dual value function $\tilde{V}$. From this we obtain the following.

THEOREM 7.1 (dual Hamilton–Jacobi equation). *The dual value function $\tilde{V}$ satisfies*

$$\sigma - H(\xi, \eta) = 0 \quad \textit{for all} \quad (\sigma, \xi) \in \partial \tilde{V}(\tau, \eta) \quad \textit{when} \quad \tau > 0. \tag{7.5}$$

Indeed, for $\tau > 0$ one has $(\sigma, \xi) \in \partial \tilde{V}(\tau, \eta)$ if and only if $(-\sigma, \eta) \in \partial V(\tau, \xi)$.

Proof. In translating Theorem 2.5 to the context of $\tilde{V}$, as justified by Theorem 5.1, we bring into the scene the dual Hamiltonian $\tilde{H}(y, x) = -H(x, y)$ corresponding (in Proposition 3.5) to the dual Lagrangian $\tilde{L}$. The vectors $(\sigma, \xi) \in \partial \tilde{V}(\tau, \eta)$ are characterized by $\xi \in \partial_\eta \tilde{V}(\tau, \eta)$ and $\sigma = -\tilde{H}(\eta, \xi) = H(\xi, \eta)$. Invoking the conjugacy between $V(\tau, \cdot)$ and $\tilde{V}(\tau, \cdot)$ in Theorem 5.1, specifically the relation (5.5), we get the subgradient equivalence. Then (7.5) is immediate from the Hamilton–Jacobi equation already in Theorem 2.5. □

We take up next the issue of what additional properties of continuity, differentiability, etc. the value function V is guaranteed to have beyond the convexity and epi-continuity in Theorem 2.1. We begin with a characterization of the interior of the set

$$\operatorname{dom} V = \big\{ (\tau, \xi) \in [0, \infty) \times \mathbb{R}^n \,\big|\, V(\tau, \xi) < \infty \big\}.$$

PROPOSITION 7.2 (domain interiors). *In terms of $V_\tau = V(\tau, \cdot)$, one has that*

$$(\tau, \xi) \in \operatorname{int} \operatorname{dom} V \quad \Longleftrightarrow \quad \tau > 0, \ \xi \in \operatorname{int} \operatorname{dom} V_\tau.$$

Proof. It's evident that "⇒" holds. We focus therefore on "⇐." Consider $\bar{\tau} > 0$ and $\bar{\xi} \in \operatorname{int} \operatorname{dom} V_{\bar{\tau}}$. The epi-convergence of V_τ to $V_{\bar{\tau}}$ as $\tau \to \bar{\tau}$ in Theorem 2.1 entails through the convexity of these functions that V_τ converges pointwise to $V_{\bar{\tau}}$ uniformly on all compact subsets of $\operatorname{int} \operatorname{dom} V_{\bar{\tau}}$; cf. [10, 7.17]. In particular, this convergence holds on some open neighborhood U of $\bar{x}$ in $\operatorname{dom} V_{\bar{\tau}}$, so for some open interval I around $\bar{\tau}$ we have $U \subset \operatorname{dom} V_\tau$ for all $\tau \in I$. Then $I \times O$ is an open subset of $\operatorname{dom} V$ containing $(\bar{\tau}, \bar{\xi})$, and we conclude that $(\bar{\tau}, \bar{\xi}) \in \operatorname{int} \operatorname{dom} V$. □

The argument just given shows further that V is continuous on the interior of $\operatorname{dom} V$, but we're headed toward showing that V is in fact locally Lipschitz continuous there. The agreement between $\partial V(\tau, \xi)$ and $\hat{\partial} V(\tau, \xi)$ in Theorem 2.5 will have a part in this, and it will yield other strong properties besides.

Recall that a locally Lipschitz continuous function is *subdifferentially regular* (in the sense of Clarke regularity of its epigraph) when all its subgradients are regular subgradients, or equivalently, its subderivatives and regular subderivatives coincide everywhere; for background, see [10, Chapters 8 and 9]. The *subderivative* function for V at a point (τ, ξ) is defined in general by

$$dV(\tau, \xi) : (\tau', \xi') \mapsto dV(\tau, \xi)(\tau', \xi') := \liminf_{\substack{\varepsilon \searrow 0 \\ (\tau'', \xi'') \to (\tau', \xi')}} \frac{V(\tau + \varepsilon \tau'', \, \xi + \varepsilon \xi'') - V(\tau, \xi)}{\varepsilon}.$$

To say that V is *semidifferentiable* at (τ,ξ) is to say that, for all (τ',ξ'), this lower limit exists actually as the full limit

$$\lim_{\substack{\varepsilon\searrow 0\\ (\tau'',\xi'')\to(\tau',\xi')}} \frac{V(\tau+\varepsilon\tau'',\,\xi+\varepsilon\xi'')-V(\tau,\xi)}{\varepsilon}.$$

Then $dV(\tau,\xi)(\tau',\xi')$ must be finite and continuous as a function of (τ',ξ'); cf. [10, 7.21].

THEOREM 7.3 (regularity consequences). *On* $\operatorname{int}\operatorname{dom} V$*, the subgradient mapping* ∂V *is nonempty-compact-convex-valued and locally bounded, and* V *itself is locally Lipschitz continuous and subdifferentially regular, moreover, semidifferentiable with*

$$dV(\tau,\xi)(\tau',\xi') \;=\; \max\Big\{\langle\xi',\eta\rangle-\tau' H(\xi,\eta)\,\Big|\,\eta\in\partial_\xi V(\tau,\xi)\Big\}. \tag{7.6}$$

Indeed, V *is strictly differentiable wherever it is differentiable, which is at almost every point of* $\operatorname{int}\operatorname{dom} V$*, and relative to such points the gradient mapping* ∇V *is continuous.*

Proof. The points $(\tau,\xi)\in\operatorname{int}\operatorname{dom} V$ have been identified in Proposition 7.2 as the ones with $\tau>0$ and $\xi\in\operatorname{int}\operatorname{dom} V(\tau,\cdot)$. Because $V(\tau,\cdot)$ is convex, the mapping $\partial_\xi V(\tau,\cdot)$ is nonempty-compact-valued and locally bounded on $\operatorname{int}\operatorname{dom} V(\tau,\cdot)$, as already known through convex analysis; cf. [6, section 24]. These properties carry over to the behavior of $\partial_\xi V$ on $\operatorname{int}\operatorname{dom} V$ because of the epi-continuous dependence of $V(\tau,\cdot)$ on τ in Theorem 2.1; see [6, section 24] again. The local boundedness of $\partial_\xi V$, when joined with the formula $\sigma=-H(\xi,\eta)$ in Theorem 5.1 and the continuity of H, gives us the nonempty-compact-valuedness and local boundedness of ∂V.

The local boundedness of ∂V on $\operatorname{int}\operatorname{dom} V$ implies that V is Lipschitz continuous there locally; cf. [10, 9.13]. Then from having $\hat{\partial} V(\tau,\xi)=\partial V(\tau,\xi)$ in Theorem 2.5 we get the subdifferential regularity of V on $\operatorname{int}\operatorname{dom} V$ and the convexity of $\partial V(\tau,\xi)$ (because $\hat{\partial} V(\tau,\xi)$ is always convex). Local Lipschitz continuity and subdifferential regularity yield semidifferentiability by [10, 9.16]. Formula (7.6) specializes the semiderivative formula in that result to V by way of the description of $\partial V(\tau,\xi)$ in Theorem 2.5.

By virtue of being locally Lipschitz continuous, V is differentiable a.e. on $\operatorname{int}\operatorname{dom} V$. In the presence of subdifferential regularity, the differentiability is strict and the gradient mapping has the stated continuity property; see [10, 9.20]. □

Elementary examples illustrate the possible nondifferentiability of V. The simplest is to let $H=0$, in which case $V(\tau,\xi)=g(\xi)$, and thus any nondifferentiability in g is propagated forward for all time. A similar effect is provided in the one-dimensional case (n=1) by $H(x,y)=-y$, which yields $V(\tau,\xi)=g(\tau+\xi)$. If g is nondifferentiable at some point $\bar{\xi}$, then V is likewise nondifferentiable at every (τ,ξ) on the line $\tau+\xi=\bar{\xi}$.

To see the trouble from another angle, let $H(x,y)=\psi(x)$, where ψ is any finite concave function. Then, no matter what the choice of convex g, one has $V(\tau,xi)=g(\xi)-\tau\psi(\xi)$. When g is finite, so too is V, but even when g is differentiable, V need not be unless ψ is differentiable. This example underscores that singularities may appear in V even with smooth initial data.

As a complement to Theorem 7.3, we develop further information about $\operatorname{int}\operatorname{dom} V$, utilizing Proposition 7.2 to translate the issue into an investigation of when $\operatorname{int}\operatorname{dom} V_\tau \neq\emptyset$. It will be convenient to work with the calculus of relative interiors and the fact that, for a convex set C in a space $\mathbb{R}^d$, one has $\operatorname{int} C\neq\emptyset$ if and only if $\operatorname{aff} C=\mathbb{R}^d$

(i.e., C isn't included in any hyperplane in $\mathbb{R}^d$), in which case $\operatorname{int} C = \operatorname{ri} C$ (cf. [10, Chapter 2]).

Additional motivation for the following result, besides facilitating use of Theorem 7.3, comes from the fact that the set $\operatorname{dom} V_\tau = \{\xi \mid V(\tau,\xi) < \infty\}$ is the *reachable set* at time τ, giving the points $\xi = x(\tau)$ reached by arcs $x(\cdot) \in \mathcal{A}^1_n[0,\tau]$ that start in $\operatorname{dom} g$ and have finite running cost $\int_0^\tau L(x(t),\dot{x}(t))dt$.

PROPOSITION 7.4 (relative interiors of reachable sets). *For every $\tau \in [0,\infty)$ one has*

$$\emptyset \neq \operatorname{ri}\operatorname{dom} V_\tau = \{\xi \mid \operatorname{ri}\operatorname{dom} g \cap \operatorname{ri}\operatorname{dom} E(\tau,\cdot,\xi) \neq \emptyset\}. \tag{7.7}$$

Here $\operatorname{ri}\operatorname{dom} V_\tau$ *reduces to* $\operatorname{int}\operatorname{dom} V_\tau$ *if and only if there exists $\xi \in \operatorname{dom} V_\tau$ such that* $\operatorname{dom} g \cup \operatorname{dom} E(\tau,\cdot,\xi)$ *does not lie in a hyperplane, that being true then for all $\xi \in \operatorname{dom} V_\tau$.*

Proof. Let $D_\tau = \operatorname{dom} V_\tau$ so $D_0 = \operatorname{dom} g$. Clearly D_τ is the image under $(\xi',\xi) \mapsto \xi$ of $C := \operatorname{dom} E(\tau,\cdot,\cdot) \cap [D_0 \times \mathbb{R}^n]$, all these sets being convex and nonempty. Then, under the same projection mapping, $\operatorname{ri} D_\tau$ is the image of $\operatorname{ri} C$; cf. [10, 2.44]. For each ξ the convex set $\operatorname{dom} E(\tau,\cdot,\xi)$ is nonempty by Corollary 4.4; likewise for each ξ' the convex set $\operatorname{dom} E(\tau,\xi',\cdot)$ is nonempty. The rule for relative interiors in product spaces (cf. [10, 2.43]) says then that

$$\operatorname{ri}\operatorname{dom} E(\tau,\cdot,\cdot) = \{(\xi',\xi) \mid \xi' \in \operatorname{ri}\operatorname{dom} E(\tau,\cdot,\xi)\} = \{(\xi',\xi) \mid \xi \in \operatorname{ri}\operatorname{dom} E(\tau,\xi',\cdot)\}. \tag{7.8}$$

This relative interior meets the set $\operatorname{ri}[D_0 \times \mathbb{R}^n] = \operatorname{ri} D_0 \times \mathbb{R}^n$, as seen from the second of the expressions in (7.8) by taking any $\xi' \in \operatorname{ri} D_0$ and then any $\xi \in \operatorname{ri}\operatorname{dom} E(\tau,\xi',\cdot)$. The rule for relative interiors of intersections (cf. [10, 2.42]) then yields

$$\operatorname{ri} C = [\operatorname{ri}\operatorname{dom} E(\tau,\cdot,\cdot)] \cap [\operatorname{ri} D_0 \times \mathbb{R}^n].$$

Returning to the observation that D_τ is the projection of $\operatorname{ri} C$, and utilizing the first of the expressions in (7.8), we get (7.7).

For the claim about interiors, we have to show that the stated condition on a point $\xi \in D_\tau$ is equivalent to the nonexistence of a hyperplane $M \supset D_\tau$. Fix any $\bar{\xi} \in D_\tau$ and any $\bar{\xi}' \in D_0$ with $(\bar{\xi}',\bar{\xi}) \in \operatorname{dom} E(\tau,\cdot,\cdot)$. A vector ζ gives a hyperplane $M = \{\xi \mid \langle \xi,\zeta\rangle = \alpha\}$ that includes D_τ if and only if $\zeta \neq 0$ and $\pm\zeta \in N_{D_\tau}(\bar{\xi})$, this being the normal cone to D_τ at $\bar{\xi}$. Likewise, a vector ζ' gives a hyperplane $M' = \{\xi' \mid \langle \xi',\zeta'\rangle = \alpha'\}$ that includes both D_0 and $\operatorname{dom} E(\tau,\cdot,\bar{\xi})$ if and only if $\zeta' \neq 0$ and both $\pm\zeta' \in N_{D_0}(\bar{\xi}')$ and $\pm\zeta' \in N_{\operatorname{dom} E(\tau,\cdot,\bar{\xi})}(\bar{\xi}')$. (Here we appeal to the fact that $\bar{\xi}'$ belongs to both D_0 and $\operatorname{dom} E(\tau,\cdot,\bar{\xi})$.) From the calculus of normals to convex sets (cf. [9, section 23], [10, Chapter 6]), the cone $N_{\operatorname{dom} E(\tau,\cdot,\bar{\xi})}(\bar{\xi}')$ is the projection of the cone $N_{\operatorname{dom} E(\tau,\cdot,\cdot)}(\bar{\xi}',\bar{\xi})$:

$$\pm\zeta' \in N_{\operatorname{dom} E(\tau,\cdot,\bar{\xi})}(\bar{\xi}') \iff \exists\,\zeta \text{ with } \pm(\zeta',\zeta) \in N_{\operatorname{dom} E(\tau,\cdot,\cdot)}(\bar{\xi}',\bar{\xi});$$

this relies on the nonemptiness of $\operatorname{dom} E(\tau,\cdot,\xi)$ for all $\xi \in \mathbb{R}^n$ (cf. Corollary 4.4), which in turn ensures that ζ' must be nonzero in this formula when $\zeta \neq 0$. Further calculus, utilizing the set relations that were developed above in determining $\operatorname{ri} D_\tau$, reveals that $\pm\zeta \in N_{D_\tau}(\bar{\xi})$ if and only if $(0,\pm\zeta) \in N_C(\bar{\xi}',\bar{\xi})$, and on the other hand that

$$N_C(\bar{\xi}',\bar{\xi}) = N_{\operatorname{dom} E(\tau,\cdot,\cdot)}(\bar{\xi}',\bar{\xi}) + N_{D_0\times\mathbb{R}^n}(\bar{\xi}',\bar{\xi}),$$
$$\text{where } N_{D_0\times\mathbb{R}^n}(\bar{\xi}',\bar{\xi}) = N_{D_0}(\bar{\xi}') \times \{0\}.$$

Thus, having a $\zeta \neq 0$ such that $\pm\zeta \in N_{D_\tau}(\bar{\xi})$ corresponds to having a $\zeta' \neq 0$ such that $\pm\zeta' \in N_{D_0}(\bar{\xi}')$ and $\pm(\zeta',\zeta) \in N_{\operatorname{dom} E(\tau,\cdot,\cdot)}(\bar{\xi}',\bar{\xi})$. This yields the claimed equivalence. □

Corollary 7.5 (interiors of reachable sets). *If* $\operatorname{int}\operatorname{dom} g \neq \emptyset$*, then for every* $\tau \in [0,\infty)$*,*

$$\emptyset \neq \operatorname{int}\operatorname{dom} V_\tau = \big\{\xi \,\big|\, \operatorname{int}\operatorname{dom} g \cap \operatorname{dom} E(\tau,\cdot,\xi) \neq \emptyset\big\}.$$

Proof. For convex sets C_1 and C_2 with $\operatorname{int} C_2 \neq \emptyset$, one has $\operatorname{ri} C_1 \cap \operatorname{ri} C_2 \neq \emptyset$ if and only if $C_1 \cap \operatorname{int} C_2 \neq \emptyset$. Then too, $C_1 \cup C_2$ cannot lie in a hyperplane. □

Corollary 7.6 (propagation of finiteness).

(a) *If* g *is finite on* $\mathbb{R}^n$*, then* V *is finite on* $[0,\infty) \times \mathbb{R}^n$*.*

(b) *If* L *is finite on* $\mathbb{R}^n \times \mathbb{R}^n$*, then* V *is finite on* $(0,\infty) \times \mathbb{R}^n$*.*

Proof. We get (a) immediately from Corollary 7.5 as the case where $\operatorname{int}\operatorname{dom} g = \mathbb{R}^n$. We get (b) by observing that, for $\tau > 0$, $\operatorname{dom} E(\tau,\cdot,\cdot)$ is all of $\mathbb{R}^n \times \mathbb{R}^n$ when L is finite. □

Corollary 7.7 (propagation of coercivity).

(a) *If* g *is coercive, then* V_τ *is coercive for every* $\tau \in [0,\infty)$*.*

(b) *If* L *is coercive, then* V_τ *is coercive for every* $\tau \in (0,\infty)$*.*

Proof. We rely on the fact that a proper convex function is coercive if and only if its conjugate is finite [10, 11.5]. The claims are justified then by the duality between V_τ and $\tilde{V}$ in Theorem 5.1 and that between L and $\tilde{L}$ in (2.15). □

REFERENCES

[1] R.T. Rockafellar, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

[2] R.T. Rockafellar, *Existence theorems for general control problems of Bolza and Lagrange type*, Adv. Math., 15 (1975), pp. 312–333.

[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[4] R.T. Rockafellar, *Hamiltonian trajectories and duality in the optimal control of linear systems with convex costs*, SIAM J. Control Optim., 27 (1989), pp. 1007–1025.

[5] R.T. Rockafellar, *Linear-quadratic programming and optimal control*, SIAM J. Control Optim., 25 (1987), pp. 781–814.

[6] R. T. Rockafellar, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–428.

[7] R. T. Rockafellar, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.

[8] R. T. Rockafellar, *Semigroups of convex bifunctions generated by Lagrange problems in the calculus of variations*, Math. Scand., 36 (1975), pp. 137–158.

[9] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[10] R. T. Rockafellar and R. J-B Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.

[11] H. Frankowska, *Optimal trajectories associated with a solution of the contingent Hamilton-Jacobi equation*, Appl. Math. Optim., 19 (1989), pp. 291–311.

[12] H. Frankowska, *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[13] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Qualitative properties of trajectories of control systems: A survey*, J. Dynam. Control Systems, 1 (1995), pp. 1–48.

[14] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[15] R. T. Rockafellar, *State constraints in convex problems of Bolza*, SIAM J. Control, 10 (1972), pp. 691–715.

[16] R. T. Rockafellar, *Dual problems of Lagrange for arcs of bounded variation*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 155–192.

[17] R. T. Rockafellar, *Optimality conditions for convex control problems with nonnegative states and the possibility of jumps*, in Game Theory and Math. Economics, O. Moeschlin, ed., North-Holland, Amsterdam, 1981, pp. 339–349.

[18] R. T. Rockafellar, *Equivalent subgradient versions of Hamiltonian and Euler-Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1315.

[19] P. D. Loewen and R. T. Rockafellar, *New necessary conditions for the generalized problem of Bolza*, SIAM J. Control Optim., 34 (1996), pp. 1496–1511.

[20] A. D. Ioffe, *Euler-Lagrange and Hamiltonian formalisms in dynamic optimization*, Trans. Amer. Math. Soc., 349 (1997), pp. 2871–2900.

[21] A. I. Subbotin, *Generalized Solutions to First-Order PDEs*, Birkhäuser, Boston, 1995.

[22] A. Melikyan, *Generalized Characteristics of First Order PDEs*, Birkhäuser, Boston, 1998.

[23] D. V. Tran, M. Tsuji, and D. Nguyen, *The Characteristic Method and Its Generalizations for First-Order Nonlinear Partial Differential Equations*, Chapman & Hall/CRC Monogr. Surv. Pure Appl. Math. 101, Chapman & Hall/CRC, Boca Raton, FL, 2000.

[24] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[25] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[26] M. G. Crandall, L. C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 478–502.

[27] R.T. Rockafellar, *Optimal arcs and the minimum value function in problems of Lagrange*, J. Optim. Theory Appl., 12 (1973), pp. 53–83.

[28] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[29] F. H. Clarke, *Hamiltonian analysis of the generalized problem of Bolza*, Trans. Amer. Math. Soc., 301 (1987), pp. 385–400.

[30] A. I. Subbotin, *Generalization of the main equation of differential game theory*, J. Optim. Theory Appl., 43 (1984), pp. 103–133.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1351–1372

CONVEXITY IN HAMILTON–JACOBI THEORY II: ENVELOPE REPRESENTATIONS*

R. TYRRELL ROCKAFELLAR† AND PETER R. WOLENSKI‡

Abstract. Upper and lower envelope representations are developed for value functions associated with problems of optimal control and the calculus of variations that are fully convex, in the sense of exhibiting convexity in both the state and the velocity. Such convexity is used in dualizing the upper envelope representations to get the lower ones, which have advantages not previously perceived in such generality and in some situations can be regarded as furnishing, at least for value functions, extended Hopf–Lax formulas that operate beyond the case of state-independent Hamiltonians.

The derivation of the lower envelope representations centers on a new function called the dualizing kernel, which propagates the Legendre–Fenchel envelope formula of convex analysis through the underlying dynamics. This kernel is shown to be characterized by a kind of double Hamilton–Jacobi equation and, despite overall nonsmoothness, to be smooth with respect to time and concave-convex in the primal and dual states. It furnishes a means whereby, in principle, value functions and their subgradients can be determined through optimization without having to deal with a separate, and typically much less favorable, Hamilton–Jacobi equation for each choice of the initial or terminal cost data.

Key words. convex value functions, Hamilton–Jacobi equations, dualizing kernels, fundamental kernels, envelope formulas, Hopf–Lax formulas, viscosity solutions, optimal control

AMS subject classifications. Primary, 49L25; Secondary, 93C10, 49N15

PII. S0363012998345378

1. Introduction.

A major goal of Hamilton–Jacobi theory is the characterization of value functions that arise from problems of optimal control and the calculus of variations in which endpoints are treated as parameters. The value function $V : [0,\infty) \times \mathbb{R}^n \to \overline{\mathbb{R}} := [-\infty,\infty]$ is defined from a Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ and an function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$\begin{aligned} V(\tau,\xi) &:= \inf\left\{ g\big(x(0)\big) + \int_0^\tau L\big(x(t),\dot{x}(t)\big)dt \,\middle|\, x(\tau) = \xi \right\} \quad \text{for } \tau > 0, \\ V(0,\xi) &:= g(\xi), \end{aligned} \tag{1.1}$$

with the minimization taking place over all the arcs (i.e., absolutely continuous functions) $x(\cdot) : [0,\tau] \to \mathbb{R}^n$ that reach ξ at time τ. Here $V(\tau,\cdot)$ is viewed as an evolving function on $\mathbb{R}^n$ which starts as g and describes how g is propagated forward to any time τ in a manner dictated by L.

Similarly, value functions can be considered that describe how g is propagated backward from a future time T, and such a "cost-to-go" formulation is common in optimal control. From a theoretical perspective, of course, backward models are equivalent to forward models through time reversal and do not require separate treatment in basic theory. The expression of control problems in terms of a Lagrangian L in

*Received by the editors October 2, 1998; accepted for publication (in revised form) July 21, 2000; published electronically December 28, 2000.

http://www.siam.org/journals/sicon/39-5/34537.html

†Department of Mathematics, Box 354350, University of Washington, Seattle, WA 98195-4350 (rtr@math.washington.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9803089.

‡Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918 (wolenski@marais.math.lsu.edu). The research of this author was supported by grant DMS-9623406.

which control parameters do not appear is parallel to the expression of control dynamics in terms of differential inclusions and has generated a substantial literature in nonsmooth optimization, going back to around 1970. More about that can be found in our companion paper [1], which is the springboard for the efforts here.

In the classical context of the calculus of variations, g and L would be smooth (i.e., continuously differentiable). For applications such as in control, however, it is important to allow g and L to be nonsmooth and even to take on ∞ because infinite penalties can systematically be used in incorporating constraints. Under the assumption that g and L are lower semicontinuous (lsc) and proper (i.e., not identically ∞, and nowhere having the value $-\infty$), the integrand $t \mapsto L(x(t), \dot{x}(t))$ is measurable, and the functional $J[x(\cdot)]$ being minimized is well defined. (The usual convention of "inf addition" is followed, in which ∞ dominates in any conflict with $-\infty$.) Then $J[x(\cdot)] = \infty$ unless the arc $x(\cdot)$ satisfies the constraints

$$\begin{aligned} &x(0) \in D, \ \text{ where } D := \big\{ x \,\big|\, g(x) < \infty \big\}, \\ &\dot{x}(t) \in F(x(t)) \ \text{ almost everywhere (a.e.) } t, \ \text{ where } F(x) := \big\{ v \,\big|\, L(x,v) < \infty \big\}. \end{aligned} \tag{1.2}$$

The customary tool for characterizing value functions is the Hamilton–Jacobi PDE in one form or another. It revolves around the Hamiltonian function H associated with L, which is defined through the Legendre–Fenchel transform by

$$H(x,y) := \sup_v \Big\{ \langle v, y \rangle - L(x,v) \Big\}. \tag{1.3}$$

Because V typically lacks smoothness, even when g and L are smooth, various generalizations of the classical PDE have been devised, the foremost being "viscosity" versions. The recent book of Bardi and Capuzzo-Dolcetta [2], with its helpful references, provides broad access to that subject. Viscosity theory is able to characterize V in situations far from classical, and sometimes even when V takes on ∞, but uniqueness results are still lacking in many situations of interest for us here, due to the failure of V to satisfy the continuity, boundedness, or growth conditions that current results demand.

In this paper, instead of working with a generalized Hamilton–Jacobi PDE for V, we develop basic "envelope representations," which characterize V as the pointwise inf or sup of a family of more elementary functions. In cases where a description of V as a unique Hamilton–Jacobi solution of some sort can indeed be furnished, now or in the future, these formulas become PDE solution formulas. For state-independent Hamiltonians, $H(x,y) \equiv H_0(y)$, they reduce to Hopf–Lax formulas. We aim at contributing to Hamilton–Jacobi theory by opening a way for such classical formulas to be extended to state-dependent Hamiltonians, while exploring representations of value functions in their own right, especially as a potential means of determining value functions and their subgradients through optimization without having to deal with a separate Hamilton–Jacobi equation for each choice of the cost function g.

We look at two kinds of envelope formulas: upper and lower. Both kinds have long been known in the Hopf–Lax setting but haven't systematically been sought outside of that. Upper envelope formulas, involving pointwise minimization, are elementary and easy to obtain very generally. However, in order for them to express V on $(0,\infty)$ as the envelope of a family of *finite* functions, not to speak of smooth or subsmooth functions, significant restrictions are necessary. Lower envelope formulas, involving pointwise maximization, arise by dualization and therefore thrive only in the presence of convexity, as with the original Hopf formula itself. In compensation for assumptions of convexity, though, they offer a number of unusual and attractive features.

Our focus will primarily be on lower envelope formulas, because of their special potential, but we will also investigate properties enjoyed by upper envelope formulas under the convexity assumptions we impose.

Convex analysis [3] will be heavily used, but mostly through the results obtained in our preceding paper [1]. To the extent that broader variational analysis is required, we rely on the book [4].

After introducing the duality scheme and deriving the basic envelope formulas in section 2 in terms of the "fundamental kernel" and the "dualizing kernel," we concentrate in section 3 on the dualizing kernel and its characterization by a double Hamilton–Jacobi equation. The lower envelope formula for V in terms of the dualizing kernel and the properties of that kernel developed in section 3 and also in section 4, where connections with subgradients of V are brought out, constitute the paper's main results. To complete the picture, relationships with standard Hopf–Lax formulas are discussed in section 5,

2. Envelopes and convexity. Upper envelope formulas rely on the "double" value function $E : [0,\infty) \times \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ that corresponds to two-endpoint, i.e., Lagrangian, minimization problems for L:

$$\begin{aligned} &E(\tau,\xi',\xi) := \inf\left\{ \int_0^\tau L\big(x(t),\dot{x}(t)\big)dt \,\middle|\, x(0)=\xi',\ x(\tau)=\xi \right\} \quad \text{for } \tau > 0,\\ &E(0,\xi',\xi) := \begin{cases} 0 & \text{if } \xi = \xi',\\ \infty & \text{otherwise,} \end{cases} \end{aligned} \tag{2.1}$$

where the minimization is over all the arcs $x(\cdot)$ that go from ξ' at time 0 to ξ at time τ.

THEOREM 2.1 (upper envelope representation). *The value function V is expressed in terms of E by the formula*

$$V(\tau,\xi) = \inf\nolimits_{\xi'} \Big\{ g(\xi') + E(\tau,\xi',\xi) \Big\} \quad \textit{for } \tau \geq 0. \tag{2.2}$$

Moreover, when $\tau > 0$, an arc $x(\cdot)$ achieves the minimum in the problem defining $V(\tau,\xi)$ in (1.1) *if and only if it achieves the minimum in the problem defining $E(\tau,\xi',\xi)$ in* (2.1) *for some choice of ξ' yielding the minimum in* (2.2).

Proof. Elementary and evident. □

We will call E the *fundamental kernel* associated with L. The "kernel" term comes from the far-reaching analogy between minimizing a sum of functions and integrating a product of functions. Formula (2.2) gives a transform whereby g is converted to $V(\tau,\cdot)$ for $\tau > 0$. It is an "upper envelope" formula because it expresses V as the pointwise infimum of a certain family of functions on $[0,\infty)\times\mathbb{R}^n$, namely the functions $e_{\xi'} : (\tau,\xi) \mapsto g(\xi') + E(\tau,\xi',\xi)$ indexed by $\xi' \in D$, where D is the effective domain of g as in (1.2). In some situations E may be finite or even smooth on $(0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$, and the same then holds for these functions $e_{\xi'}$.

Often E takes on ∞, though, and the upper envelope representation may be difficult to exploit directly. Clearly, $E(\tau,\xi',\xi)$ can't be finite unless there is an arc $x(\cdot)$ that conforms to the differential inclusion in (1.2) and carries ξ' to ξ. Thus, extended-real-valuedness of E is inevitable in applications where the implicit constraints in (1.2) can seriously come into play.

This motivates a search for alternative envelope representations in which troublesome infinite values can be bypassed. Such representations will be generated by way of the function $K : [0,\infty) \times \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ with

$$\begin{aligned} &K(\tau,\xi,\eta) := \inf\left\{ \langle x(0),\eta\rangle + \int_0^\tau L\big(x(t),\dot{x}(t)\big)dt \,\middle|\, x(\tau) = \xi \right\}, \\ &K(0,\xi,\eta) := \langle \xi,\eta\rangle, \end{aligned} \tag{2.3}$$

which we introduce now as the *dualizing kernel* associated with L. The minimization takes place over all arcs $x(\cdot) : [0,\infty) \to \mathbb{R}^n$ that reach ξ at time τ.

For fixed η, $K(\cdot,\cdot,\eta)$ is the value function obtained as in (1.1) but with the linear function $\langle \cdot,\eta\rangle$ in place of g. As a consequence of Theorem 2.1, therefore, we have

$$K(\tau,\xi,\eta) = \inf_{\xi'}\Big\{ \langle \xi',\eta\rangle + E(\tau,\xi',\xi) \Big\}, \tag{2.4}$$

and indeed, this could serve as well as (2.3) in defining K.

Observe that (2.4) dualizes E by employing a variant of the Legendre–Fenchel transform: $-K(\tau,\xi,\eta)$ is calculated by taking the function conjugate to $E(\tau,\cdot,\xi)$ under that transform and evaluating it at $-\eta$. When $E(\tau,\cdot,\xi)$ is lsc, proper, and convex, it can be recovered by the reciprocal formula

$$E(\tau,\xi',\xi) = \sup_\eta \Big\{ K(\tau,\xi,\eta) - \langle \xi',\eta\rangle \Big\}. \tag{2.5}$$

Our strategy is to use such duality between E and K, along with convexity of g, to translate the upper envelope representation in Theorem 2.1 into a lower one involving K and the function g^* conjugate to g. A prerequisite for this, however, is placing assumptions on L that will ensure that E has the properties needed for (2.5) to be valid.

Such assumptions have been identified in our paper [1]. In stating them, we call a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *coercive* if f is bounded from below and has $f(v)/|v| \to \infty$ as $|v| \to \infty$ (where $|\cdot|$ is the Euclidean norm). When applied to a proper, nondecreasing function θ on $[0,\infty)$, coercivity means having $\theta(s)/s \to \infty$ as $s \to \infty$.

Basic Assumptions (A).

(A0) The initial function g is convex, proper, and lsc on $\mathbb{R}^n$.

(A1) The Lagrangian function L is convex, proper, and lsc on $\mathbb{R}^n \times \mathbb{R}^n$.

(A2) The mapping F underlying L in (1.2) is nonempty-valued everywhere, and there is a constant ρ such that $\operatorname{dist}(0,F(x)) \le \rho\big(1+|x|\big)$ for all x.

(A3) There are constants α and β and a coercive, proper, nondecreasing function θ on $[0,\infty)$ such that $L(x,v) \ge \theta\big(\max\{0,\, |v| - \alpha|x|\,\}\big) - \beta|x|$ for all x and v.

The meaning of these assumptions has thoroughly been elucidated in [1], so for present purposes we need only to record some key facts and examples.

An immediate consequence of $L(x,v)$ being, by (A1) and (A2), a convex, proper, lsc function of v for each x is that L can be recovered from H by

$$L(x,v) = \sup_y \Big\{ \langle v,y\rangle - H(x,y) \Big\}. \tag{2.6}$$

The correspondence between Lagrangians and Hamiltonians is thus one-to-one under our conditions. For each H of a certain class, the associated L is uniquely determined by (2.6). The Hamiltonian class is described as follows.

PROPOSITION 2.2 (Hamiltonian conditions). *The Hamiltonians for the Lagrangians L satisfying* (A1), (A2), *and* (A3) *are the functions $H : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ such that*

(a) *$H(x,y)$ is finite, concave in x, and convex in y (hence locally Lipschitz continuous).*

(b) *There are constants α and β and a finite, convex function φ such that*

$$H(x,y) \;\le\; \varphi(y) + (\alpha|y| + \beta)|x| \quad \textit{for all } x,\ y.$$

(c) *There are constants γ and δ and a finite, concave function ψ such that*

$$H(x,y) \;\ge\; \psi(x) - (\gamma|x| + \delta)|y| \quad \textit{for all } x,\ y.$$

Proof. This comes from Theorem 2.3 of [1]. Finite concave-convex functions are locally Lipschitz continuous by [3, 35.1]. □

Example 2.3 (subseparable Lagrangians). *Let the Lagrangian have the form*

$$L(x,v) = G(x) + L_0(v - Ax) \tag{2.7}$$

for $A \in \mathbb{R}^{n\times n}$, a finite convex function G on $\mathbb{R}^n$, and a proper convex function L_0 on $\mathbb{R}^n$ that is lsc and coercive. Then L satisfies (A1), (A2), *and* (A3)*, and its Hamiltonian is*

$$H(x,y) = \langle Ax, y\rangle - G(x) + H_0(y), \tag{2.8}$$

where H_0 is a finite convex function on $\mathbb{R}^n$, namely $H_0 = L_0^*$. Conversely, if H has the form (2.8) *for finite convex functions G and H_0, then L has the form* (2.7) *with $L_0 = H_0^*$ and falls in the category described.*

Detail. This is evident from Proposition 2.2 and the conjugacy between finite convex functions (always continuous) and proper convex functions that are lsc and coercive. □

Subseparable Lagrangians illustrate also, in a relatively simple case, the way that our framework of Lagrangians and Hamiltonians connects with control theory. An optimal control problem with linear dynamics $\dot{x}(t) = Ax(t) + Bu(t)$ and running cost integral

$$\int_0^T \big\{G(x(t)) + F(u(t))\big\}dt,$$

with F convex, proper, lsc, and coercive (but possibly taking on ∞) corresponds to the Lagrangian L in (2.7) for

$$L_0(z) = \min\big\{F(u)\,\big|\, Bu = z\big\},$$

and then the Hamiltonian H in (2.8) has $H_0(y) = F^*(B^*y)$, where B^* is the transpose of B and F^* is the convex function conjugate to F, this function being finite because of the coercivity of F. Control constraints are incorporated here through the specification of the set where F is finite. Control formats much more general than this, yet still fully convex and (as may be shown) still fitting with our assumptions, can be found in [5], [6].

Note that if the coercivity condition in (A3) were replaced by a simpler condition like $L(x,v) \ge \theta(|v|)$, Lagrangians of the type in Example 2.3 would have to have $A = 0$, and G would have to be bounded from below.

Of course, there are many more Lagrangians satisfying (A1), (A2), and (A3) than the ones in Example 2.3. An illustration is $L(x,v) = \frac{1}{p}\max\{|x|^p, |v|^p\}$ with $p \in (1,\infty)$, which for q determined by $(1/p) + (1/q) = 1$ has

$$H(x,y) = \begin{cases} \frac{1}{q}|y|^q & \text{when } |y| \geq |x|^{p-1}, \\ |x||y| - \frac{1}{p}|x|^p & \text{when } |y| \leq |x|^{p-1}. \end{cases} \tag{2.9}$$

Proposition 2.4 ([1] convexity of the fundamental kernel). *Under* (A1), (A2), *and* (A3), *$E(\tau,\xi',\xi)$ is a convex, proper, lsc function of (ξ',ξ) for each $\tau \geq 0$. In fact, $E(\tau,\cdot,\xi)$ is proper and coercive for every ξ, and $E(\tau,\xi',\cdot)$ is proper and coercive for every ξ'.*

Proof. This is extracted from Proposition 4.2 and Corollary 4.4 of [1]. □

On the basis of this result we do have the reciprocal formula in (2.5) along with the one in (2.4), and E and K are entirely dual to each other. We are able then to convert the envelope formula in (1.1) into one for functions that are likely to be better behaved. The technique is to apply Fenchel's duality theorem to the minimization problem in (1.1) in order to recast it as a maximization problem.

In the next theorem, and henceforth in this paper, we take assumptions (A) for granted, unless otherwise mentioned.

Theorem 2.5 (lower envelope representation). *The dualizing kernel $K(\tau,\xi,\eta)$ is everywhere finite, convex in ξ, and concave in η. The value function V is expressed in terms of K by the formula*

$$V(\tau,\xi) = \sup_\eta \Big\{K(\tau,\xi,\eta) - g^*(\eta)\Big\}. \tag{2.10}$$

Proof. For any $(\tau,\xi) \in [0,\infty) \times \mathbb{R}^n$, the function $f = E(\tau,\cdot,\xi)$ is lsc, proper, convex, and coercive by Proposition 2.4, so its conjugate f^* is finite. We have $-f^*(-\eta) = K(\tau,\xi,\eta)$ by (2.4), hence $K(\tau,\xi,\eta)$ is finite and concave in η. On the other hand, the convexity of $E(\tau,\xi',\xi)$ in (ξ',ξ) in Proposition 2.4 implies the convexity of $K(\tau,\cdot,\eta)$ by the general principle that when the Legendre–Fenchel transform is applied to one argument of a convex function of two arguments, the result is concave in the residual argument; see [3, 33.3] or [4, 11.48]. (The concavity becomes convexity under the changes of sign.)

To obtain the lower envelope representation, we fix ξ along with τ and view the upper envelope representation in (2.2) as expressing $V(\tau,\xi)$ as the optimal value in the problem of minimizing $g(\xi') + f(\xi')$ for $f = E(\tau,\cdot,\xi)$ as above. By Fenchel's duality theorem (cf. [3, 31.1] or [4, 11.41]), one has

$$\inf_{\xi'}\Big\{g(\xi') + f(\xi')\Big\} = \sup_\eta\Big\{-f^*(-\eta) - g^*(\eta)\Big\} \tag{2.11}$$

if the relative interiors of the convex sets $\{\eta \,|\, -f^*(-\eta) > -\infty\}$ and $\{\eta \,|\, g^*(\eta) < +\infty\}$ have a point in common. That criterion is met through the finiteness of f^*, which makes the first set be all of $\mathbb{R}^n$. Since the inf in (2.11) gives the left side of (2.10) and the sup in (2.11) gives the right side, the equation in (2.10) is confirmed. □

For $\tau = 0$, the lower envelope representation in (2.10) reduces to the Legendre–Fenchel envelope formula

$$g(\xi) = \sup_\eta\Big\{\langle\xi,\eta\rangle - g^*(\eta)\Big\}, \tag{2.12}$$

which expresses the proper, lsc, convex function g as the pointwise supremum of all the affine functions majorized by g. For $\tau > 0$, it can be viewed as extending this formula forward in time through a Hamilton–Jacobi propagation of those affine functions into a different family of functions.

In employing the Fenchel duality rule (2.11) as the tool for passing between upper and lower envelope representations, we are in effect invoking a "minimax principle" in a manner reminiscent in Hamilton–Jacobi theory of the duality seen in classical Hopf–Lax formulas (which will be taken up in section 5). Indeed, our formulas can be recast as follows.

THEOREM 2.6 (envelope formulas in minimax mode). *In terms of the dualizing kernel K, the value function V always has the representation*

$$V(\tau,\xi) = \inf_{\xi'} \sup_{\eta} \Big\{ g(\xi') - \langle \xi', \eta \rangle + K(\tau,\xi,\eta) \Big\}, \tag{2.13}$$

and this even holds for an arbitrary choice of $g : \mathbb{R}^n \to \overline{\mathbb{R}}$. When g is convex, proper, and lsc, however, V also has the representation

$$V(\tau,\xi) = \sup_{\eta} \inf_{\xi'} \Big\{ g(\xi') - \langle \xi', \eta \rangle + K(\tau,\xi,\eta) \Big\}. \tag{2.14}$$

Proof. We get (2.13) by combining the elementary formula (2.2) for V in terms of E with the reciprocal formula (2.5) for E in terms of K, which is valid by Proposition 2.4 under our assumptions. We get (2.14) by combining the representation (2.10) of V in terms of K and g^* with the definition of g^* in terms of g. □

All of duality theory in convex optimization, a very highly developed subject, has the character of a "minimax principle" of course, but there is no single minimax theorem to invoke that would fit all cases. Everything revolves around the precise conditions under which "inf" and "sup" can legitimately be interchanged when the simplest compactness and continuity properties may be absent, as here. Duality of a much deeper kind than in the proof of Theorem 2.5 will be crucial later, for instance, in ascertaining the circumstances in which the supremum in (2.10) is attained and how this can be used in determining the subgradients of V from those of K (cf. Theorem 4.2 and Corollary 4.4 below). Observe that this brings out an important advantage of expressing lower envelope representations as in (2.10) instead of as in (2.14).

The appearance of g^* instead of g in (2.10) shouldn't be regarded as much of a drawback. In many situations g^* can explicitly be determined from g (see [3] and [4, Chapter 11] for the calculus of conjugates), but even if not, there is much that might be made of this formula. Depending on the particular structure of g (in terms of operations like addition and composition), it's common for $g^*(\eta)$ to be expressible as the optimal value in a minimization problem with respect to some other vector, let's call it ζ, in which η is a parameter. When such an expression is substituted into (2.10), one gets a representation of $V(\tau,\xi)$ as the optimal value in a maximization problem involving both η and ζ.

Anyway, as a practical matter, optimization formulas for $V(\tau,\xi)$, whether directly as in (2.10) or with some expansion of the $g^*(\eta)$ term, are generally more favorable for computation than integration formulas, which become intractable numerically in more than a few dimensions. Furthermore, for applications such as to feedback in optimal control the subgradients of V are at least as important as its values. The lower representation in (2.10) affords a much better grip on those than does the upper representation in (2.2), because K is typically far better behaved than E, as will emerge from the results that follow. These better properties suggest that K may be easier to generate than V in a Hamilton–Jacobi context, after which the lower envelope representation in Theorem 2.5 might be used to compute aspects of V as needed, for instance in feedback. Moreover, the same K would be able to serve for every V that relies on the Lagrangian L, no matter what the choice of the initial cost function g.

3. Characterization of the dualizing kernel.

Turning now to the development of properties of K that underpin the lower envelope representation in Theorem 2.5, we begin with a special kind of Hamilton–Jacobi characterization. Only subgradients in the sense of convex analysis are needed in this characterization, but other subgradients will soon enter the discussion too, so we go straight to a review of the full definitions. For background, see [4].

Consider any function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and let x be any point at which $f(x)$ is finite. A vector $y \in \mathbb{R}^n$ is a *regular subgradient* of f at x, written $y \in \hat{\partial} f(x)$, if

$$f(x') \;\geq\; f(x) + \langle y,\, x' - x\rangle + o(|x' - x|). \tag{3.1}$$

It is a (*general*) *subgradient* of f at x, written $y \in \partial f(x)$, if there is a sequence of points $x^\nu \to x$ with $f(x^\nu) \to f(x)$ for which regular subgradients $y^\nu \in \hat{\partial} f(x^\nu)$ exist with $y^\nu \to y$. (We consistently use superscript ν for sequences; $\nu \to \infty$.) When f is convex, the sets $\hat{\partial} f(x)$ and $\partial f(x)$ are the same and agree with the subgradient set of convex analysis, defined by (3.1) without the "o" term.

These of course are "lower" subgradients, the corresponding regular and general "upper" subgradient sets, defined with the opposite inequality in (3.1) and will be denoted here by $\tilde{\hat{\partial}} f(x)$ and $\tilde{\partial} f(x)$; thus

$$\tilde{\partial} f = -\partial[-f]. \tag{3.2}$$

This notation is expedient because most situations can be couched in terms of lower subgradients alone, cf. [4], although just now we'll have something of an exception.

Regular subgradients have been the mainstay in viscosity theory, but general subgradients are the vehicle for many of the strongest results in variational analysis [4].

In the following theorem, $\partial_\xi K(\tau, \xi, \eta)$ refers to subgradients of the convex function $K(\tau, \cdot, \eta)$, whereas $\tilde{\partial}_\eta K(\tau, \xi, \eta)$ refers to subgradients of the concave function $K(\tau, \xi, \cdot)$.

THEOREM 3.1 (double Hamilton–Jacobi equation). *The kernel $K(\tau, \xi, \eta)$ is continuously differentiable with respect to τ and satisfies, for $\tau \geq 0$,*

$$\begin{aligned} -\frac{\partial K}{\partial \tau}(\tau, \xi, \eta) &= \begin{cases} H(\xi, \eta') & \textit{for all} \quad \eta' \in \partial_\xi K(\tau, \xi, \eta), \\ H(\xi', \eta) & \textit{for all} \quad \xi' \in \tilde{\partial}_\eta K(\tau, \xi, \eta), \end{cases} \\ K(0, \xi, \eta) &= \langle \xi, \eta \rangle, \end{aligned} \tag{3.3}$$

where $\partial K / \partial \tau$ is interpreted as the right partial derivative when $\tau = 0$. Moreover, K is the only such function with $K(\tau, \xi, \eta)$ convex in ξ and concave in η.

The proof of Theorem 3.1 will be furnished later in this section, after some additional developments. The continuous differentiability refers to $(\partial K / \partial \tau)(\tau, \xi, \eta)$ depending continuously on $(\tau, \xi, \eta) \in [0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n$.

The double Hamilton–Jacobi equation in (3.3) has been placed in the elementary picture of subgradients of convex and concave functions and partial derivatives in time, because that seems most conducive to possible uses of the result. What comparison, though, can be made with viscosity versions of Hamilton–Jacobi equations? And why use two equations instead of one?

The double aspect of the characterization comes from the fact that, through duality, K has an alternative expression in which the roles of ξ and η are interchanged.

PROPOSITION 3.2 (alternative formula for the dualizing kernel). *In minimizing over arcs $y(\cdot):[0,\tau]\to\mathbb{R}^n$, one has*

$$-K(\tau,-\xi,\eta)=\inf\Big\{\langle\xi,y(0)\rangle+\int_0^\tau L_*\big(y(t),\dot y(t)\big)dt\,\Big|\,y(\tau)=\eta\Big\},\tag{3.4}$$

where $L_*(y,w)=L^*(-w,y)$. Moreover L_*, like L, satisfies (A1), (A2), *and* (A3), *and its Hamiltonian H_* is given by*

$$H_*(y,x)=\sup\nolimits_w\Big\{\langle w,x\rangle-L_*(y,w)\Big\}=-H(-x,y).\tag{3.5}$$

Proof. The duality theory for convex problems of Bolza [7] will be applied in the form distilled in [1, section 4]. The minimization problem that defines $K(\tau,\xi,\eta)$ in (2.3) is

$$\text{minimize}\ \int_0^\tau L(x(t),\dot x(t))dt+l(x(0),x(\tau))\ \text{ over arcs }\ x(\cdot):[0,\tau]\to\mathbb{R}^n,\tag{$\mathcal{P}$}$$

where $l(a,b)=\langle a,\eta\rangle$ if $b=\xi$ but $l(a,b)=\infty$ if $b\neq\xi$. The duality theory pairs this with

$$\text{minimize}\ \int_0^\tau \tilde L(y(t),\dot y(t))dt+\tilde l(y(0),y(\tau))\ \text{ over arcs }\ y(\cdot):[0,\tau]\to\mathbb{R}^n,\tag{$\tilde{\mathcal{P}}$}$$

where $\tilde L(y,w)=L^*(w,y)$ and $\tilde l(c,d)=l^*(c,-d)$; the latter comes out here as $\tilde l(c,d)=-\langle\xi,d\rangle$ if $c=\eta$ but $\tilde l(c,d)=\infty$ if $c\neq\eta$. Because $l(\cdot,b)$ is finite on $\mathbb{R}^n$ for a certain b, and $\tilde l(c,\cdot)$ is finite on $\mathbb{R}^n$ for a certain c, the optimal values in the two problems are related by $\inf(\mathcal{P})=-\inf(\tilde{\mathcal{P}})$; this holds by [1, Corollary 4.6]. Thus,

$$-K(\tau,\xi,\eta)=\inf\bigg\{\int_0^\tau\tilde L\big(y(t),\dot y(t)\big)dt-\langle\xi,y(\tau)\rangle\,\Big|\,y(0)=\eta\bigg\}.$$

By rewriting in terms of $z(t)=y(\tau-t)$, we can convert this to

$$-K(\tau,\xi,\eta)=\inf\bigg\{\int_0^\tau\tilde L\big(z(t),-\dot z(t)\big)dt-\langle\xi,z(0)\rangle\,\Big|\,z(\tau)=\eta\bigg\}.$$

It remains only to replace ξ by $-\xi$ and the z notation by y again to obtain (3.4).

The fact that L_* again satisfies (A1), (A2), and (A3) comes from the fact that $\tilde L$ inherits these properties from L, as demonstrated in [1, Proposition 3.5]. The expression for the Hamiltonian H_* in terms of H arises similarly from that result, which asserts that the Hamiltonian $\tilde H$ for $\tilde L$ has $\tilde H(y,x)=-H(x,y)$. □

Through results in [1], the value function formulas for K in (2.3) and (3.4) lead to major conclusions about the subgradients of K and in particular to a viscosity version of the double Hamilton–Jacobi equation in Theorem 3.1. This time we use $\partial_{\tau,\xi}K(\tau,\xi,\eta)$ to denote subgradients of the function $K(\cdot,\cdot,\xi)$ on $[0,\infty)\times\mathbb{R}^n$, and so forth.

THEOREM 3.3 (subgradients of the dualizing kernel). *For $\tau>0$, one has*

$$\begin{aligned}(\sigma,\eta')\in\partial_{\tau,\xi}K(\tau,\xi,\eta)\quad&\Longleftrightarrow\quad(\sigma,\eta')\in\hat\partial_{\tau,\xi}K(\tau,\xi,\eta)\\&\Longleftrightarrow\quad\eta'\in\partial_\xi K(\tau,\xi,\eta),\ \ \sigma=-H(\xi,\eta'),\end{aligned}\tag{3.6}$$

and on the other hand

$$\begin{aligned}(\sigma,\xi') \in \tilde{\partial}_{\tau,\eta}K(\tau,\xi,\eta) \quad &\Longleftrightarrow \quad (\sigma,\xi') \in \tilde{\hat{\partial}}_{\tau,\eta}K(\tau,\xi,\eta)\\ &\Longleftrightarrow \quad \xi' \in \tilde{\partial}_{\eta}K(\tau,\xi,\eta),\ \ \sigma = -H(\xi',\eta).\end{aligned} \tag{3.7}$$

Proof. We simply apply [1, Theorem 2.5] first to $K(\cdot,\cdot,\eta)$, which is the value function that propagates $\langle\cdot,\eta\rangle$ under L as in (2.3), and second to $-K(\cdot,-\xi,\cdot)$, which by Proposition 3.2 is the value function that propagates $\langle\xi,\cdot\rangle$ under L_*. □

Corollary 3.4 (double viscosity equation). *For* $\tau > 0$*, one has*

$$\begin{cases}\sigma + H(\xi,\eta') = 0 & \textit{for all} \quad (\sigma,\eta') \in \hat{\partial}_{\tau,\xi}K(\tau,\xi,\eta),\\ \sigma + H(\xi',\eta) = 0 & \textit{for all} \quad (\sigma,\xi') \in \tilde{\hat{\partial}}_{\tau,\eta}K(\tau,\xi,\eta).\end{cases} \tag{3.8}$$

It will be established in the next theorem that K is locally Lipschitz continuous. In view of this, the first of the subgradient equations in (3.8) is equivalent, as shown by Frankowska [8], to $K(\cdot,\cdot,\eta)$ being a Hamilton–Jacobi viscosity solution in the sense of satisfying the upper and lower inequalities of Crandall, Evans, and Lions [9], with initial $K(0,\cdot,\eta) = \langle\cdot,\eta\rangle$. The second equation has a similar viscosity interpretation relative to a switch in the roles of the ξ and η arguments.

It might be hoped that either of these subgradient equations, by itself, would be enough to determine K uniquely. That could be true, but unfortunately the existing results on uniqueness of viscosity solutions are not fully up to the task. The trouble is that H and K need not satisfy the kinds of growth or boundedness conditions assumed in such results. Because of the initial condition $K(\tau,\xi,\eta)$ is certainly neither globally bounded from above nor globally bounded from below, even for fixed ξ or η. One or the other kind of boundedness would be needed to apply the latest uniqueness theorem of Ishii [10], for example. Anyway, the Hamiltonian can grow at rates like those in (2.9), and this can be problematical as well.

Theorem 3.5 (Lipschitz continuity of the dualizing kernel). *The function* K *is locally Lipschitz continuous on* $[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$.

Proof. By [1, Theorem 2.1] the functions $K(\cdot,\cdot,\eta)$ are lsc on $[0,\infty)\times\mathbb{R}^n$ as value functions in the mode of (2.3). Similarly by this result, as applied in the context of Proposition 3.2, the functions $-K(\cdot,-\xi,\cdot)$ are lsc on $[0,\infty)\times\mathbb{R}^n$. Hence the functions $K(\cdot,\xi,\cdot)$ are usc on $[0,\infty)\times\mathbb{R}^n$, and it follows in particular that $K(\tau,\xi,\eta)$ is continuous in $\tau\in[0,\infty)$ for each (ξ,η). Thus, $K(\tau,\cdot,\cdot)$ converges pointwise to $K(\bar{\tau},\cdot,\cdot)$ whenever $\tau\to\bar{\tau}$ in $[0,\infty)$. The functions $K(\tau,\cdot,\cdot)$ are finite and convex-concave by Theorem 2.5, and pointwise convergence of such functions on $\mathbb{R}^n\times\mathbb{R}^n$ implies uniform convergence on bounded sets (see [3, 35.4]). In consequence, K is continuous on $[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$. Furthermore, the convergence implies that the mapping

$$S : (\tau,\xi,\eta) \mapsto \big\{(\eta',\xi') \,\big|\, \eta' \in \partial_{\xi}K(\tau,\xi,\eta),\ \ \xi' \in \tilde{\partial}_{\eta}K(\tau,\xi,\eta)\,\big\} \tag{3.9}$$

is locally bounded on $[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$ and has closed graph (see [3, 35.7]).

This yields through the continuity of H (in Proposition 2.2) the closed graph property and local boundedness on $[0,\infty)\times\mathbb{R}^n$ of the mappings

$$\begin{aligned}(\tau,\xi) &\mapsto \big\{(\sigma,\eta')\,\big|\,\eta' \in \partial_{\xi}K(\tau,\xi,\eta),\ \ \sigma = -H(\xi,\eta')\big\},\\ (\tau,\eta) &\mapsto \big\{(\sigma,\eta')\,\big|\,\eta' \in \tilde{\partial}_{\eta}K(\tau,\xi,\eta),\ \ \sigma = -H(\xi',\eta)\big\}.\end{aligned} \tag{3.10}$$

In general, a function f that is finite and lsc on an open set O in a space $\mathbb{R}^d$ is Lipschitz continuous with constant κ on any set $X \subset O$ such that $|y| \leq \kappa$ for all $y \in \partial f(x)$ when $x \in X$; this holds by [4, 9.2, 9.13]. We invoke this now for $K(\cdot,\cdot,\eta)$ on $(0,\infty)\times\mathbb{R}^n$. From the subgradient characterization in (3.6) of Theorem 3.3 and the local boundedness of the first mapping in (3.10), on $[0,\infty)\times\mathbb{R}^n$ rather than just $(0,\infty)\times\mathbb{R}^n$, we get that $K(\cdot,\cdot,\eta)$ is locally Lipschitz continuous on $(0,\infty)\times\mathbb{R}^n$, and moreover that the Lipschitz constants don't blow up as $\tau \searrow 0$. Since $K(\cdot,\cdot,\eta)$ is anyway continuous on $[0,\infty)\times\mathbb{R}^n$, we conclude it must actually be Lipschitz continuous on $[0,\infty)\times\mathbb{R}^n$.

A parallel argument utilizing the dual formula in Proposition 3.2 shows that the functions $K(\cdot,\xi,\cdot)$ are Lipschitz continuous on $[0,\infty)\times\mathbb{R}^n$. The two properties of Lipschitz continuity combine to give the Lipschitz continuity of K itself on $[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$. □

The subgradient result in Theorem 3.3 will be complemented now by one about subderivatives. These are defined as follows; see [4] for background. For $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point x where $f(x)$ is finite, the *subderivative* of f at x for a vector w is

$$df(x)(w) := \liminf_{\substack{\varepsilon \searrow 0 \\ w' \to w}} \frac{f(x+\varepsilon w') - f(x)}{\varepsilon}. \tag{3.11}$$

If this "liminf" coincides with the associated "limsup" and thus exists as a full limit, f is said to be *semidifferentiable at x for w*, or simply *semidifferentiable at x* if true for all $w \in \mathbb{R}^n$. Semidifferentiability at x corresponds to the difference quotient functions $\Delta_\varepsilon f(x) : w \mapsto [f(x+\varepsilon w) - f(x)]/\varepsilon$ converging uniformly on bounded subsets of $\mathbb{R}^n$, as $\varepsilon \searrow 0$, to a continuous function of w [4, 7.21]. Differentiability is the case where, in addition, the limit function $df(x)$ is linear.

THEOREM 3.6 (subderivatives of the dualizing kernel). *On $(0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$, K is semidifferentiable everywhere, and its subderivative formula is as follows. For any (τ,ξ,η), the quantities $H(\xi,\eta')$ for $\eta' \in \partial_\xi K(\tau,\xi,\eta)$ and $H(\xi',\eta)$ for $\xi' \in \tilde{\partial}_\eta K(\tau,\xi,\eta)$ all have the same value, and in denoting it by $k(\tau,\xi,\eta)$ one has*

$$\begin{aligned} dK(\tau,\xi,\eta)(\theta,\omega,\zeta) = {} & -\theta k(\tau,\xi,\eta) \\ & + \max\bigl\{\langle\eta',\omega\rangle \bigm| \eta' \in \partial_\xi K(\tau,\xi,\eta)\bigr\} \\ & + \min\bigl\{\langle\xi',\zeta\rangle \bigm| \xi' \in \partial_\eta K(\tau,\xi,\eta)\bigr\}. \end{aligned} \tag{3.12}$$

Proof. The first part of the proof, devoted to the existence of the common value $k(\tau,\xi,\eta)$, will be the basis later for knowing that K is continuously differentiable in τ as claimed in Theorem 3.1 but not yet justified.

Let $K_\eta = K(\cdot,\cdot,\eta)$. Since K_η is the value function that propagates a finite convex function under L, namely $\langle\cdot,\eta\rangle$, it is semidifferentiable on $(0,\infty)\times\mathbb{R}^n$ by [1, Theorem 7.3] with the formula

$$dK_\eta(\tau,\xi)(\theta,\omega) = \max\bigl\{\langle\omega,\eta'\rangle - \theta H(\xi,\eta') \bigm| \eta' \in \partial_\xi K_\eta(\tau,\xi)\bigr\}. \tag{3.13}$$

Likewise for $K^\xi = -K(\cdot,-\xi,\cdot)$ in the context of Proposition 3.2, K^ξ is the value function that propagates $\langle\xi,\cdot\rangle$ under L_* and thus is semidifferentiable on $(0,\infty)\times\mathbb{R}^n$ with formula

$$dK^\xi(\tau,\eta)(\theta,\zeta) = \max\bigl\{\langle\xi',\zeta\rangle - \theta H_*(\eta,\xi') \bigm| \xi' \in \partial_\eta K^\xi(\tau,\eta)\bigr\},$$

where $H_*(\eta,\xi') = -H(-\xi',\eta)$. In terms of $\tilde{K}_\xi = K(\cdot,\xi,\cdot)$ the latter can be rewritten as

$$d\tilde{K}_\xi(\tau,\eta)(\theta,\zeta) = \min\bigl\{\langle\xi',\zeta\rangle - \theta H(\xi',\eta) \,\big|\, \xi' \in \tilde{\partial}_\eta \tilde{K}_\xi(\tau,\eta)\bigr\}. \tag{3.14}$$

In particular, K has right and left partial derivatives in τ,

$$\begin{aligned}(\partial^+K/\partial\tau)(\tau,\xi,\eta) &= dK_\eta(\tau,\xi)(1,0) = d\tilde{K}_\xi(\tau,\eta)(1,0),\\ (\partial^-K/\partial\tau)(\tau,\xi,\eta) &= -dK_\eta(\tau,\xi)(-1,0) = -d\tilde{K}_\xi(\tau,\eta)(-1,0),\end{aligned}$$

which by (3.13) must satisfy

$$\begin{aligned}(\partial^-K/\partial\tau)(\tau,\xi,\eta) &= \min\bigl\{-H(\xi,\eta') \,\big|\, \eta' \in \partial_\xi K(\tau,\xi,\eta)\bigr\},\\ (\partial^+K/\partial\tau)(\tau,\xi,\eta) &= \max\bigl\{-H(\xi,\eta') \,\big|\, \eta' \in \partial_\xi K(\tau,\xi,\eta)\bigr\},\end{aligned} \tag{3.15}$$

and on the other hand, by (3.14), must satisfy

$$\begin{aligned}(\partial^-K/\partial\tau)(\tau,\xi,\eta) &= \max\bigl\{-H(\xi',\eta) \,\big|\, \xi' \in \tilde{\partial}_\eta K(\tau,\xi,\eta)\bigr\},\\ (\partial^+K/\partial\tau)(\tau,\xi,\eta) &= \min\bigl\{-H(\xi',\eta) \,\big|\, \xi' \in \tilde{\partial}_\eta K(\tau,\xi,\eta)\bigr\}.\end{aligned} \tag{3.16}$$

We get $(\partial^-K/\partial\tau)(\tau,\xi,\eta) \le (\partial^+K/\partial\tau)(\tau,\xi,\eta)$ from (3.15) but the opposite inequality from (3.16). The partial derivative $(\partial K/\partial\tau)(\tau,\xi,\eta)$ therefore exists and is given by all four expressions on the right in (3.15) and (3.16). In particular, the quantities $H(\xi,\eta')$ and $H(\xi',\eta)$ involved in these expressions must have the same value.

In denoting this common value by $k(\tau,\xi,\eta)$, we have a function that is continuous not only on $(0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$ but has a continuous extension to $[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$. That follows from the closed graph property and local boundedness on $[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$ of the mappings in (3.10), as demonstrated in the proof of Theorem 3.5. Hence $(\partial K/\partial\tau)(\tau,\xi,\eta)$ exists even at $\tau = 0$, when interpreted there as the right partial derivative, and it depends continuously on $(\tau,\xi,\eta)\in[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$.

Henceforth in proceeding with the proof of Theorem 3.6, we argue solely on the basis of $K(\tau,\xi,\eta)$ being continuously differentiable in τ while convex in ξ and concave in η. This will help with something needed eventually in the proof of Theorem 3.1, although a price must be paid in overlaps with arguments already furnished for Theorem 3.5.

Each of the functions $K(\tau,\cdot,\cdot)$, being finite and convex-concave, is locally Lipschitz continuous on $\mathbb{R}^n\times\mathbb{R}^n$ by [3, 35.1]. The differentiability of $K(\tau,\xi,\eta)$ with respect to τ entails continuity in τ. Therefore, whenever $\tau\to\bar{\tau}$ in $[0,\infty)$ the functions $K(\tau,\cdot,\cdot)$ converge pointwise on $\mathbb{R}^n\times\mathbb{R}^n$ to $K(\bar{\tau},\cdot,\cdot)$. We have already seen in the proof of Theorem 3.6 how this convergence implies that the mapping S in (3.9) is locally bounded on $[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$ with closed graph. This guarantees that the local Lipschitz continuity of the functions $K(\tau,\cdot,\cdot)$ is uniform locally with respect to τ (by virtue of [3, 24.7] as applied in the convex and concave arguments separately). In taking this together with the continuity of $\partial K/\partial\tau$, which ensures the local Lipschitz continuity of $K(\tau,\xi,\eta)$ in τ, we deduce that K is locally Lipschitz continuous as a function of $(\tau,\xi,\eta)\in[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$.

We work next with the difference quotient functions concerned in generating the subderivatives of K:

$$\begin{aligned}\Delta_\varepsilon K(\tau,\xi,\eta)(\theta,\omega,\zeta) = {}&\frac{K(\tau+\varepsilon\theta,\,\xi+\varepsilon\omega,\,\eta+\varepsilon\zeta) - K(\tau,\,\xi+\varepsilon\omega,\eta+\varepsilon\zeta)}{\varepsilon}\\ &+\frac{K(\tau,\,\xi+\varepsilon\omega,\,\eta+\varepsilon\zeta) - K(\tau,\xi,\eta)}{\varepsilon}.\end{aligned} \tag{3.17}$$

When $\varepsilon \searrow 0$, the first expression in the sum in (3.17), as a function of (θ, ω, ζ), converges uniformly over bounded sets to the function

$$(\theta, \omega, \zeta) \mapsto (\partial K/\partial \tau)(\tau, \xi, \eta)\theta$$

because of the continuity of $\partial K/\partial \tau$ (through a classical argument using the mean value theorem). The second expression in the sum in (3.17), as a function of (ω, ζ) that is convex-concave, is known from convex analysis [3, 35.6] to converge pointwise to the function

$$(\omega, \zeta) \;\mapsto\; \max_{\eta' \in \partial_\xi K(\tau,\xi,\eta)} \langle \eta', \omega \rangle + \min_{\xi' \in \partial_\eta K(\tau,\xi,\eta)} \langle \xi', \zeta \rangle.$$

The convergence must then be uniform over bounded subsets of $\mathbb{R}^n \times \mathbb{R}^n$ (by [3, 35.4]). Thus, as $\varepsilon \searrow 0$, the functions $\Delta_\varepsilon K(\tau, \xi, \eta)$ do converge uniformly on bounded sets to the function described by the right side of (3.12) with $k = (\partial K/\partial \tau)$. Hence K is semidifferentiable with this as its formula. □

Proof of Theorem 3.1. The continuous differentiability of $K(\tau, \xi, \eta)$ has been demonstrated in the first part of the proof of Theorem 3.6 along with the double formula for $(\partial K/\partial \tau)$ in (3.3), the common value on the right side of (3.3) being the expression $k(\tau, \xi, \eta)$ introduced in the statement of Theorem 3.6. The remaining task is to show the uniqueness in this characterization. Let $J(\tau, \xi, \eta)$ on $[0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n$ be convex in ξ, concave in η, and continuously differentiable in τ, satisfying (3.3). We have to prove that $J = K$.

As a tool in this endeavor, we can use the fact that J, like K, has the subderivative properties in Theorem 3.6, since those properties depend only on the facts now being assumed; see the remark in the middle of the proof of Theorem 3.6 (in the paragraph starting with "Henceforth"). Thus

$$\begin{aligned} dJ(\tau, \xi, \eta)(\theta, \omega, \zeta) = {}& \theta\, (\partial J/\partial \tau)(\tau, \xi, \eta) \\ & + \max\big\{ \langle \eta', \omega \rangle \,\big|\, \eta' \in \partial_\xi J(\tau, \xi, \eta) \big\} \\ & + \min\big\{ \langle \xi', \zeta \rangle \,\big|\, \xi' \in \partial_\eta J(\tau, \xi, \eta) \big\}. \end{aligned} \tag{3.18}$$

In addition we can take J to be locally Lipschitz continuous, because that property was likewise seen there to be a consequence of the current assumptions.

Fix (τ, ξ, η). Certainly $J(\tau, \xi, \eta) = K(\tau, \xi, \eta)$ when $\tau = 0$, so suppose $\tau > 0$. The infimum in the definition (2.3) of $K(\cdot, \cdot, \eta)$ as the value function propagating $\langle \xi, \cdot \rangle$ is attained by an arc $x(\cdot)$ on $[0, \tau]$ which moreover is Lipschitz continuous; this holds by [1, Theorem 5.2], which under our assumptions (A) applies to value functions at interior points (τ, ξ) of their domains. Then too, for any $\tau' \in (0, \tau)$ and the point $\xi' = x(\tau')$, the restriction of $x(\cdot)$ to $[0, \tau']$ is optimal for the minimization problem defining $K(\tau', \xi', \eta)$ (by the "principle of optimality"). Thus

$$K(\tau', x(\tau'), \eta) = \langle x(0), \eta \rangle + \int_0^{\tau'} L\big(x(s), \dot{x}(s)\big) dt \quad \text{for } 0 \le \tau' \le \tau. \tag{3.19}$$

In terms of the functions $\varphi : [0, \tau] \to \mathbb{R}$ and $\psi : [0, \tau] \to \mathbb{R}$ defined by

$$\varphi(t) := K(t, x(t), \eta), \qquad \psi(t) := J(t, x(t), \eta),$$

we have $\varphi(0) = K(0, x(0), \eta) = J(0, x(0), \eta) = \psi(0)$, whereas $\varphi(\tau) = K(\tau, \xi, \eta)$ and $\psi(\tau) = J(\tau, \xi, \eta)$. Furthermore, φ is Lipschitz continuous on $[0, \tau]$, because $x(\cdot)$ has this property and K is locally Lipschitz continuous on $[0,\infty)\times\mathbb{R}^n\times\mathbb{R}^n$. Likewise ψ is Lipschitz continuous on $[0,\tau]$. It follows that φ and ψ are the integrals of their derivatives, which exist a.e. Hence

$$K(\tau,\xi,\eta)-J(\tau,\xi,\eta)=\int_0^\tau \varphi'(t)dt-\int_0^\tau \psi'(t)dt. \tag{3.20}$$

On the basis of (3.19), we have

$$\varphi'(t)=L\big(x(t),\dot{x}(t)\big) \text{ for a.e. } t. \tag{3.21}$$

On the other hand, the semidifferentiability of J in (3.18) yields

$$\psi'(t) \;=\; (\partial J/\partial\tau)(t,x(t),\eta)+\max\big\{\langle\eta',\dot{x}(t)\rangle \,\big|\, \eta'\in\partial_\xi J(t,x(t),\eta)\big\}.$$

For each t let $y(t)$ be a vector ξ' attaining this maximum. Because J satisfies the Hamilton–Jacobi equations in (3.3), we have $(\partial J/\partial\tau)(t,x(t),\eta)=-H(x(t),y(t))$, so that

$$\psi'(t)=-H(x(t),y(t))+\langle y(t),\dot{x}(t)\rangle. \tag{3.22}$$

Since $L(x(t),\cdot)$ and $H(x(t),\cdot)$ are conjugate convex functions, we know from the reciprocal Legendre–Fenchel formula in (2.6) that $\langle y(t),\dot{x}(t)\rangle-H(x(t),y(t))\le L(x(t),\dot{x}(t))$. Therefore $\psi'(t)\le\varphi'(t)$ by (3.22) and (3.21). When this inequality is combined with (3.20), we arrive at the conclusion that $J(\tau,\xi,\eta)\le K(\tau,\xi,\eta)$.

So far, we have established that $J\le K$. To get the opposite inequality, it suffices to show that $-J(\tau,-\xi,\eta)\le -K(\tau,-\xi,\eta)$ for all (τ,ξ,η). But for this we need only to appeal to the alternative value function formula for K in Proposition 3.2 and in such terms reapply the argument just given. □

4. Additional kernel properties and subgradient formulas.

Other facts about the kernels K and E will now be developed, with emphasis on subgradients and regularity. Connections between subgradients of the value function V and those of the dualizing kernel K are featured because of their possible use in applications to feedback in optimal control.

An important role in bringing out such connections is played by the generalized Hamiltonian dynamical system associated with H, which has the form

$$\dot{x}(t)\in\partial_y H(x(t),y(t)), \qquad -\dot{y}(t)\in\tilde{\partial}_x H(x(t),y(t)). \tag{4.1}$$

This dynamical system is the key to characterizing optimality in the theory of generalized problems of Bolza for the Lagrangian L, where it originated in [11]. More on its properties and history can be found in [1] and its references. A Hamiltonian *trajectory* over $[0,\tau]$ is a pair of arcs $x(\cdot)$ and $y(\cdot)$ satisfying (4.1) for almost every t.

THEOREM 4.1 (kernel subgradients and Hamiltonian dynamics). *The following properties are equivalent for any* $\tau\ge 0$:

(a) $\eta'\in\partial_\xi K(\tau,\xi,\eta)$ *and* $\xi'\in\tilde{\partial}_\eta K(\tau,\xi,\eta)$;

(b) $(-\eta,\eta')\in\partial_{\xi',\xi}E(\tau,\xi',\xi)$;

(c) *there is a Hamiltonian trajectory* $(x(\cdot),y(\cdot))$ *over* $[0,\tau]$ *from* (ξ',η) *to* (ξ,η').

Proof. The equivalence between (a) and (b) reflects a general principle about how subgradients behave when partial conjugates are taken, as in the passage between E and K in (2.4) and (2.5); cf. [4, 11.48].

The equivalence between (a) and (c) will come out of a result in [1, Theorem 2.4] about the subgradients of value functions V; more generally, one has $\eta' \in \partial_\xi V(\tau,\xi)$ if and only if there is a Hamiltonian trajectory $(x(\cdot),y(\cdot))$ over $[0,\tau]$ that starts with $y(0) \in \partial g(x(0))$ and ends at (ξ,η'). Since $K(\cdot,\cdot,\eta)$ is the value function that propagates $\langle\cdot,\eta\rangle$, a function with constant subgradient (gradient) η, we deduce that $\eta' \in \partial_\xi K(\tau,\xi,\eta)$ if and only if there is a Hamiltonian trajectory $(x(\cdot),y(\cdot))$ over $[0,\tau]$ that starts with $y(0)=\eta$ (any $x(0)$) and ends at (ξ,η').

For the remainder, we argue in terms of the dual expression in Proposition 3.2, where $-K(\cdot,-\xi,\cdot)$ is the value function that propagates $\langle\xi,\cdot\rangle$ under L_*, a Lagrangian with Hamiltonian H_* given by (3.5). Invoking the same theorem from [1] in this setting, we obtain, after the ± signs settle down and the trajectories are reversed in time, the fact that $\xi' \in \tilde{\partial}_\eta K(\tau,\xi,\eta)$ if and only if there is a Hamiltonian trajectory over $[0,\tau]$ that starts at (ξ',η) and ends with $x(\tau)=\xi$ (any $y(\tau)$). In putting this together with the earlier statement, we arrive at the description in (c). $\square$

THEOREM 4.2 (determination of value function subgradients). *For every $\tau>0$, one has*

$$\text{(4.2)}\qquad \begin{aligned} \partial V(\tau,\xi) &= \bigcup\Big\{\partial_{\tau,\xi}K(\tau,\xi,\eta)\,\Big|\,\eta\in M(\tau,\xi)\Big\},\\ &\textit{where}\ \ M(\tau,\xi) := \operatorname{argmax}_\eta\Big\{K(\tau,\xi,\eta)-g^*(\eta)\Big\}. \end{aligned}$$

Therefore, subgradients of V can be determined from those of K by carrying out the maximization in the lower envelope formula with

$$\text{(4.3)}\qquad (\sigma,\eta')\in\partial V(\tau,\xi) \iff \exists\,\eta\in M(\tau,\xi)\ \textit{with}\ \begin{cases}\eta'\in\partial_\xi K(\tau,\xi,\eta),\\ \sigma=-H(\xi,\eta').\end{cases}$$

Proof. Recall from Theorem 3.3 that the subgradients in $\partial_{\tau,\xi}K(\tau,\xi,\eta)$ are of the form $(-H(\xi,\eta'),\eta')$ for $\eta'\in\partial_\xi K(\tau,\xi,\eta)$. A similar result was obtained in [1, Theorem 2.5] for V; its subgradients have the form $(-H(\xi,\eta'),\eta')$ for $\eta'\in\partial_\xi V(\tau,\xi)$. Further, as already noted in the proof of Theorem 4.1, it was demonstrated in [1, Theorem 2.4] that $\eta'\in\partial_\xi V(\tau,\xi)$ if and only if there is a Hamiltonian trajectory $(x(\cdot),y(\cdot))$ that starts with $y(0)\in\partial g(x(0))$ and ends at (ξ,η').

On the other hand, the condition for η to belong to $M(\tau,\xi)$, i.e., to maximize $K(\tau,\xi,\eta)-g^*(\eta)$, can be expressed in subgradient terms as $0\in\tilde{\partial}_\eta K(\tau,\xi,\eta)-\partial g^*(\eta)$. (This is both necessary and sufficient for optimality because g^* is a convex function while $K(\tau,\xi,\cdot)$ is a finite concave function; see [3, section 31].) Equivalently, there exists some $\xi'\in\tilde{\partial}_\eta K(\tau,\xi,\eta)\cap\partial g^*(\eta)$. But for conjugate convex functions we have $\xi'\in\partial g^*(\eta)$ if and only if $\eta\in\partial g(\xi')$. In view of Theorem 4.1, then, we have $\eta\in M(\tau,\xi)$ if and only if there is a Hamiltonian trajectory $(x(\cdot),y(\cdot))$ that starts with $y(0)\in\partial g(x(0))$ and ends at (ξ,η'). This is the same as the condition derived in terms of V, so we conclude that the subgradient formula in the theorem is correct. $\square$

Theorem 4.2 puts the spotlight on the maximizing set $M(\tau,\xi)$ in the lower envelope formula (1.5) and raises questions about the nature of this subproblem of maximization, in particular whether the maximum is actually attained. We address these questions next.

THEOREM 4.3 (compactness and attainment in the lower envelope formula). *For any $\tau>0$ and ξ, the following properties in the lower envelope formula are equivalent:*

(a) *the set $M(\tau,\xi)=\operatorname{argmax}_\eta\{K(\tau,\xi,\eta)-g^*(\eta)\}$ is nonempty and compact;*

(b) *for every $\beta\in\mathbb{R}$, the upper level set $\{\eta\,|\,K(\tau,\xi,\eta)-g^*(\eta)\geq\beta\}$ is compact;*

(c) $\xi \in \operatorname{int} D(\tau)$ *for the set* $D(\tau) = \big\{\xi \,\big|\, V(\tau,\xi) < \infty\big\}$.

Proof. We return to the proof of Theorem 2.5 and the framework of Fenchel duality in which it was placed, with $f = E(\tau,\cdot,\xi)$ and $-f^*(-\eta) = K(\tau,\xi,\eta)$. It is well known in that theory, in terms of the convex sets $\operatorname{dom} f$ and $\operatorname{dom} g$ (where f and g are finite), that $\operatorname{argmin}_\eta \big\{f^*(-\eta) + g^*(\eta)\big\}$ is nonempty and bounded if and only if $0 \in \operatorname{int}(\operatorname{dom} f - \operatorname{dom} g)$ and the infimum is finite (see, for instance, [4, 11.41, 11.39(b)].) That is in turn equivalent to having $\operatorname{ri} \operatorname{dom} g \cap \operatorname{ri} \operatorname{dom} f \neq \emptyset$ with $\operatorname{dom} g \cup \operatorname{dom} f$ not lying in a hyperplane (cf. [4, 2.45]). In [1, Proposition 7.4] this property has been identified with (c). Thus, (a) is equivalent to (c).

The equivalence between (a) and (b), on the other hand, results from the fact that the function being maximized is concave and upper semicontinuous; cf. [4, 3.27]. □

COROLLARY 4.4 (finite value functions). *When V is finite on $(0,\infty) \times \mathbb{R}^n$, the maximizing set $M(\tau,\xi)$ is nonempty and compact for every $(\tau,\xi) \in (0,\infty) \times \mathbb{R}^n$. In particular this is the case when g is finite on $\mathbb{R}^n$ or on the other hand when L is finite on $\mathbb{R}^n \times \mathbb{R}^n$.*

Proof. The first assertion is justified through condition (c) in Theorem 4.3. The rest cites elementary circumstances in which V is know from [1, Corollary 7.6] to be finite. □

We look further now at the fundamental kernel E, first demonstrating a property of epi-continuity. Epi-continuity, which refers to epigraphs depending continuously on a parameter in the sense of Painléve–Kuratowski set convergence, was established in [1, Theorem 2.1] for the dependence of $V(\tau,\cdot)$ on $\tau \in [0,\infty)$. We'll apply that result to the functions $E(\tau,\cdot,\cdot)$ by way of a reformulation trick.

PROPOSITION 4.5 (fundamental epi-continuity). *The function $E(\tau,\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ depends epi-continuously on $\tau \in [0,\infty)$: whenever $\tau^\nu \to \tau$ with $\tau^\nu \geq 0$ one has*

$$\begin{cases} \liminf_\nu E(\tau^\nu,\xi'^\nu,\xi^\nu) \geq E(\tau,\xi',\xi) & \textit{for every sequence } (\xi'^\nu,\xi^\nu) \to (\xi',\xi), \\ \limsup_\nu E(\tau^\nu,\xi'^\nu,\xi^\nu) \leq E(\tau,\xi',\xi) & \textit{for some sequence } (\xi'^\nu,\xi^\nu) \to (\xi',\xi). \end{cases}$$

Proof. Although E seems to fit a different pattern than V, in being generated as a value function in terms of a variable pair (ξ',ξ) of initial and terminal points, instead of an initial function g and a terminal point ξ, we can nonetheless obtain results about E from those for V by an adaptation. The trick is to view E as the value function $V_E : [0,\infty) \times \mathbb{R}^{2n} \to \overline{\mathbb{R}}$ that is generated from the Lagrangian L_E and initial function g_E defined as follows:

$$L_E(x',x,v',v) := \begin{cases} L(x,v) & \text{if } v' = 0, \\ \infty & \text{if } v' \neq 0, \end{cases}$$
$$g_E(x',x) := \begin{cases} 0 & \text{if } x' = x, \\ \infty & \text{if } x' \neq x. \end{cases} \tag{4.4}$$

Indeed, under these definitions $V_E(\tau,\xi',\xi)$ is the infimum of $\int_0^\tau L(x(t),\dot{x}(t))dt$ over all arcs $(x'(\cdot),x(\cdot)) \in \mathcal{A}^1_{2n}[0,\tau]$ such that $x'(0) = x(0)$, $\dot{x}'(t) = 0$ a.e., and $x'(\tau) = \xi'$. The latter conditions obviously force $x(0)$ to be ξ'. Note that g_E and L_E satisfy our blanket assumptions (A) (in higher-dimensional interpretation) because L satisfies (A1)–(A3). By this route, we get justification of our claims through epi-continuity results for value functions in [1, Theorem 2.1]. □

In our next result, we record a basic relationship between certain effective domains, which although convex, could in general have empty interior, namely

$$\begin{aligned} \operatorname{dom} E(\tau,\cdot,\cdot) &= \big\{(\xi',\xi)\,\big|\,E(\tau,\xi',\xi)<\infty\big\},\\ \operatorname{dom} E &= \big\{(\tau,\xi',\xi)\,\big|\,\tau>0,\ E(\tau,\xi',\xi)<\infty\big\}. \end{aligned} \tag{4.5}$$

Proposition 4.6 (domain interior). *The following properties are equivalent:*

(a) $\tau > 0$ *and* $(\xi',\xi)\in \operatorname{int}\operatorname{dom} E(\tau,\cdot,\cdot)$;

(b) $(\tau,\xi',\xi)\in\operatorname{int}\operatorname{dom} E$.

Proof. This is [1, Proposition 7.2] as applied to the value function V_E in the reformulation in the proof of Proposition 4.5. □

In the following theorem, subdifferential regularity is a property that a function has when its epigraph is closed and Clarke regular; see [4].

Theorem 4.7 (regularity of the fundamental kernel). *On* $\operatorname{int}\operatorname{dom} E$, *the subgradient mapping* ∂E *is nonempty-compact-convex-valued and locally bounded, and* E *itself is locally Lipschitz continuous and subdifferentially regular, moreover semidifferentiable with*

$$dE(\tau,\xi',\xi)(\tau',\omega',\omega) \;=\; \max\Big\{\langle\omega,\eta'\rangle-\langle\omega',\eta\rangle-\tau' H(\xi,\eta')\,\Big|\,(-\eta,\eta')\in\partial_{\xi',\xi}E(\tau,\xi',\xi)\Big\},$$

where $H(\xi,\eta')$ *could be replaced by* $H(\xi',\eta)$*. Indeed,* E *is strictly differentiable wherever it is differentiable, which is at almost every point of* $\operatorname{int}\operatorname{dom} E$*, and with respect to such points the gradient mapping* ∇E *is continuous.*

Proof. We apply [1, Theorem 7.3], a result for value functions V in general under our assumptions, to V_E in the pattern of the proof of Proposition 4.6 above. □

Theorem 4.8 (Hamilton–Jacobi equations for the fundamental kernel). *The subgradients of* E *on* $(0,\tau)\times\mathbb{R}^n\times\mathbb{R}^n$ *have the property that*

$$\begin{aligned} (\sigma,-\eta,\eta')\in\partial E(\tau,\xi',\xi) \quad &\Longleftrightarrow \quad (\sigma,-\eta,\eta')\in\hat\partial E(\tau,\xi',\xi)\\ &\Longleftrightarrow \quad (-\eta,\eta')\in\partial_{\xi',\xi}E(\tau,\xi',\xi),\ \ \sigma=-H(\xi,\eta'),\\ &\Longleftrightarrow \quad (-\eta,\eta')\in\partial_{\xi',\xi}E(\tau,\xi',\xi),\ \ \sigma=-H(\xi',\eta). \end{aligned} \tag{4.6}$$

In particular, E *is a solution to the generalized double Hamilton–Jacobi equation:*

$$\left.\begin{aligned} \sigma+H(\xi,\eta')=0\\ \sigma+H(\xi',\eta)=0\end{aligned}\right\}\ \textit{for all}\ \ (\sigma,-\eta,\eta')\in\partial E(\tau,\xi',\xi)\ \ \textit{when}\ \tau>0. \tag{4.7}$$

Proof. We get the equivalence of the first three conditions by applying [1, Theorem 2.5] to V_E, once again following the pattern of reformulation in the proof of Proposition 4.5, but for that purpose it is necessary to know the Hamiltonian H_E for the Lagrangian L_E in (4.4). This calculates out simply to $H_E(x',x,y',y)=H(x,y)$. To add the fourth condition in (4.6), we utilize the subgradient description in Theorem 4.1. Along any Hamiltonian trajectory, H is constant (as proved in [11]), so if the trajectory goes from (ξ',η) to (ξ,η') we must have $H(\xi,\eta')=H(\xi',\eta)$. □

The double Hamilton–Jacobi equation for E isn't surprising in view of the one for K in Theorem 3.1. Indeed, each double equation is essentially equivalent to the other by virtue of the relations in Theorem 4.1. It follows that E is uniquely determined by (4.7) and the initial condition in its definition (2.1). An earlier viscosity version of the double equation for E in simpler cases where E is finite can be seen in the book of Lions [12]. In general cases where E can be discontinuous and take on ∞, however, the Hamilton–Jacobi characterization of K has a major advantage over the one for E in Theorem 4.8, due to the assured finiteness and local Lipschitz continuity of $K(\tau,\xi,\eta)$ (Theorem 3.5), its smoothness in τ (Theorem 3.1), and its semidifferentiability everywhere with respect to all arguments jointly (Theorem 3.6).

5. Application to Hopf–Lax formulas and their generalization. Upper and lower envelope representations of value functions as solutions to Hamilton–Jacobi equations first appeared in works of Hopf [13] and Lax [14] in very particular situations where the Hamiltonian $H(x,y)$ is independent actually of x. We inspect the state-independent case as an example within our framework and then go on to describe how our results cover an extension of the Hopf–Lax formulas beyond that case. The aim is to provide further perspective on how our formulas for value functions tie in with Hamilton–Jacobi theory.

EXAMPLE 5.1 (formulas of classical Hopf–Lax type). *Suppose that $L(x,v) = L_0(v)$ for a coercive, proper, lsc, convex function $L_0 : \mathbb{R}^n \to \overline{\mathbb{R}}$, or that $H(x,y) = H_0(y)$ for a finite convex function $H_0 : \mathbb{R}^n \to \mathbb{R}$, these assumptions being equivalent through the conjugacy relations $H_0 = L_0^*$, $L_0 = H_0^*$. Then the dualizing kernel is given by*

$$K(\tau,\xi,\eta) = \langle \xi,\eta\rangle - \tau H_0(\eta), \tag{5.1}$$

whereas the fundamental kernel is given by

$$E(\tau,\xi',\xi) = \begin{cases} \tau L_0\big(\tau^{-1}[\xi-\xi']\big) & \textit{if } \tau > 0, \\ 0 & \textit{if } \tau = 0,\ \xi-\xi' = 0, \\ \infty & \textit{if } \tau = 0,\ \xi-\xi' \neq 0. \end{cases} \tag{5.2}$$

Thus, for any initial function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ one has the upper envelope formula

$$V(\tau,\xi) = \inf_{\xi'}\Big\{ g(\xi') + \tau L_0\big(\tau^{-1}[\xi-\xi']\big)\Big\}, \tag{5.3}$$

while if g is convex, proper, and lsc, one also has the lower envelope formula

$$V(\tau,\xi) = \sup_{\eta}\Big\{ \langle \xi,\eta\rangle - \tau H_0(\eta) - g^*(\eta)\Big\}. \tag{5.4}$$

Proof. Conditions (A1), (A2), and (A3) are fulfilled, since this amounts to Example 2.3 with $A = 0$ and $G \equiv 0$. The formula for K in (5.1) follows at once from the second half of the double Hamilton–Jacobi equation in Theorem 3.1, according to which $(\partial K/\partial\tau)(\tau,\xi,\eta) = -H_0(\eta)$. The formula for E in (5.2) then follows from the general one for E in terms of K in (2.5). Finally, we get the upper envelope representation from Theorem 2.1 and the lower envelope representation from Theorem 2.5. □

The duality between the upper and lower envelope representations in this example can also be seen from the angle that (5.4) can be written as

$$V(\tau,\cdot) = (g^* + \tau H_0)^*, \tag{5.5}$$

whereas the right side of (5.3) gives the well known formula of convex analysis for such a conjugate function in terms of the functions $g^{**} = g$ and $H_0^* = L_0$ (see [4, section 11], for instance). In the traditions of Hamilton–Jacobi theory going back to Hopf

[13], g^* and H_0^* don't appear and the formulas for these functions are substituted instead. The upper representation comes out then as

$$V(\tau,\xi) = \inf_{\xi'} \sup_{\eta} \Big\{ g(\xi') + \langle \xi - \xi', \eta \rangle - \tau H_0(\eta) \Big\}, \tag{5.6}$$

while the lower representation becomes

$$V(\tau,\xi) = \sup_{\eta} \inf_{\xi'} \Big\{ g(\xi') + \langle \xi - \xi', \eta \rangle - \tau H_0(\eta) \Big\}. \tag{5.7}$$

Nowadays, though, with the Legendre–Fenchel transform so well understood, there's no reason not to simplify these expressions by writing them with conjugate functions. The equality between the "inf sup" in (5.6) and the "sup inf" in (5.7) falls into the pattern of minimax representations of primal and dual optimization problems of convex type for which there is, by now, an enormous literature; see [3] and [4, Chapter 11]. Generally speaking, such an equality is deeply involved with convexity and requires other qualifications besides. Such qualifications are met here because of our assumptions (A).

Although both (5.6) and (5.7) were proposed by Hopf [13] as possible formulas for solutions to a generalized Hamilton–Jacobi PDE in the mode of

$$u_t(t,x) + H_0(u_x(t,x)) = 0, \qquad u(0,x) = g(x),$$

the first of these is often called the Lax formula because of its appearance in a special case in the earlier paper of Lax [14] on hyperbolic conservation laws.

In work since Hopf, the lecture notes of Lions [12] and Evans [15] have provided further treatment of Hopf–Lax formulas. The paper of Bardi and Evans [16] deserves particular mention. Those authors proved that the upper formula in (5.6), or equivalently (5.3), gives the unique viscosity solution to the Hamilton–Jacobi equation in the case of a finite convex function H_0 and a possibly nonconvex function g that is globally Lipschitz continuous; alternatively by Evans [15], g can be merely continuous if H_0 is coercive. (Coercivity of H_0 corresponds in convex analysis to finiteness of L_0.) In Example 5.1, this formula has been seen to give the value function V regardless of such extra conditions on g or H_0.

Bardi and Evans [16] also showed that the lower formula (5.4), or equivalently (5.7), gives the unique viscosity solution as long as g is convex and globally Lipschitz continuous (which is known in convex analysis to correspond to the effective domain of g^* being bounded). Recently Alvarez, Barron, and Ishii [17] have removed these restrictions: the assertion holds true for all lsc, proper, convex functions g. In the context of Example 5.1, therefore, it follows that the value function V is the unique viscosity solution—in the sense of Barron and Jensen [18] or Frankowska [8], [19] (who employs a subgradient equation in place of a pair of inequalities involving upper as well as lower subgradients).

In the case of the lower envelope formula, Bardi and Evans [16] don't actually assume that H_0 is convex but just that it is continuous, and they still are able then to identify the unique viscosity solution under their strong assumptions on g. Our framework doesn't cover that feature, because the case is not one of optimization and there is no value function V of type (1.1) as a solution candidate.

We demonstrate now that the formulas in Example 5.1 can be extended to a significantly larger class of situations connected with optimal control (in the manner explained after Example 2.3), where the Lagrangian and Hamiltonian *aren't* state-independent, while maintaining their relatively explicit character. Again we emphasize that in the absence of a uniqueness theorem in Hamilton–Jacobi theory capable of handling all the Hamiltonians and value functions in our framework, these formulas, although they uniquely describe value functions, can't yet be claimed to give unique Hamilton–Jacobi solutions.

EXAMPLE 5.2 (extended Hopf–Lax formulas with linear state dependence). *Suppose that $L(x, v) = L_0(v - Ax)$ for a coercive, proper, lsc, convex function $L_0 : \mathbb{R}^n \to \overline{\mathbb{R}}$, or that $H(x, y) = \langle Ax, y\rangle + H_0(y)$ for a finite convex function $H_0 : \mathbb{R}^n \to \mathbb{R}$, these assumptions being equivalent through the conjugacy relations $H_0 = L_0^*$, $L_0 = H_0^*$. Here A is any $n \times n$ matrix. Let A^* be the transpose of A and define $\Psi : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}$ by*

$$\Psi(\tau, \eta) := \int_0^\tau H_0\big(e^{-tA^*}\eta\big)\, dt, \tag{5.8}$$

this expression being finite and convex in η. Then the dualizing kernel is given by

$$K(\tau, \xi, \eta) = \langle e^{-\tau A}\xi, \eta\rangle \tag{5.9}$$

and the fundamental kernel is given by

$$E(\tau, \xi', \xi) = \Phi(\tau, e^{-\tau A}\xi - \xi'), \tag{5.10}$$

where $\Phi(\tau, \zeta) = \sup_\eta \big\{\langle\zeta, \eta\rangle - \Psi(\tau, \eta)\big\}$, or in other words, $\Phi(\tau, \cdot)$ is the convex function conjugate to $\Psi(\tau, \cdot)$. Thus, for any initial function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ one has the upper envelope representation

$$\begin{aligned} V(\tau, \xi) &= \inf\nolimits_{\xi'} \Big\{g(\xi') + \Phi(\tau, e^{-\tau A}\xi - \xi')\Big\} \\ &= \inf\nolimits_{\xi'} \sup\nolimits_\eta \Big\{g(\xi') + \langle e^{-\tau A}\xi - \xi', \eta\rangle - \Psi(\tau, \eta)\Big\}, \end{aligned} \tag{5.11}$$

while if g is convex, proper, and lsc, one also has the lower envelope representation

$$\begin{aligned} V(\tau, \xi) &= \sup\nolimits_\eta \Big\{\langle e^{-\tau A}\xi, \eta\rangle - \Psi(\tau, \eta) - g^*(\eta)\Big\} \\ &= \sup\nolimits_\eta \inf\nolimits_{\xi'} \Big\{g(\xi') + \langle e^{-\tau A}\xi - \xi', \eta\rangle - \Psi(\tau, \eta)\Big\}. \end{aligned} \tag{5.12}$$

Proof. Fix (τ, ξ, η) and let $k(t) := K(t, \xi, y(t))$ for $y(t) := e^{(t-\tau)A^*}\eta$. From Theorem 3.5, k is Lipschitz continuous on $[0, \tau]$. We have $k(\tau) = K(\tau, \xi, \eta)$ and

$$k(0) = K(0, \xi, y(0)) = \langle\xi, y(0)\rangle = \langle\xi, e^{-\tau A^*}\eta\rangle = \langle e^{-\tau A}\xi, \eta\rangle. \tag{5.13}$$

Furthermore, from the semidifferentiability of K in Theorem 3.6 and its differentiability with respect to τ we have (a.e.)

$$\dot k(t) = (\partial K/\partial\tau)(\tau, \xi, y(t)) + \min\big\{\langle\xi', \dot y(t)\rangle \,\big|\, \xi' \in \tilde\partial_\eta K(t, \xi, y(t))\big\}, \tag{5.14}$$

where $\dot y(t) = A^* y(t)$. For each $t \in [0, \tau]$ let $x(t)$ denote some vector ξ' for which the minimum in (5.14) is attained. Then

$$\begin{aligned} \dot k(t) =& (\partial K/\partial\tau)(\tau, \xi, y(t)) + \langle x(t), \dot y(t)\rangle, \\ &\text{where } \langle x(t), \dot y(t)\rangle = \langle x(t), A^* y(t)\rangle = \langle Ax(t), y(t)\rangle. \end{aligned} \tag{5.15}$$

The second of the Hamilton–Jacobi equations in Theorem 3.1 gives us

$$(5.16)\qquad (\partial K/\partial\tau)(\tau,\xi,y(t)) = -H(x(t),y(t)) = -\langle Ax(t),y(t)\rangle - H_0(y(t)).$$

In combining (5.15) and (5.16) we get $\dot{k}(t) = -H_0(y(t)) = -H_0(e^{(t-\tau)A^*}\eta)$, hence

$$k(\tau) = k(0) - \int_0^\tau H_0\big(e^{(t-\tau)A^*}\eta\big)dt,$$

with the integral equaling $\Psi(\tau,\eta)$ (as seen through time reversal). The desired formula for K in (5.9) comes out now from (5.13) and the fact that $k(\tau) = K(\tau,\xi,\eta)$.

The corresponding formula for E in (5.10) is immediate then from (2.5), and the envelope representations are valid on the basis of Theorems 2.1 and 2.5. □

Example 5.2 may be compared to a recent result of Arisawa and Tourin in [20], extending the upper envelope formula (5.3) to a very special case of state-dependent Hamiltonians of concave-convex type. Those authors take $\mathbb{R}^n$ to be $\mathbb{R}^m \times \mathbb{R}^m$ and treat

$$H(x,y) = H(x_1,x_2;y_1,y_2) = \langle x_2,y_1\rangle + h(y_2)$$

with h a finite *coercive* convex function on $\mathbb{R}^m$ having $\min h = h(0) = 0$. For that case they work out a more detailed expression for the fundamental kernel than the one in (5.10).

The formulas in Example 5.2 reduce to the familiar ones in Example 5.1 when $A = 0$, of course. The big difference is that with $A \neq 0$ they can be applied to optimal control problems with dynamics $\dot{x} = Ax + Bu$ through the connection laid out in section 2 after Example 2.3. Then $H_0(y) = F^*(B^*y)$ for a finite convex conjugate function F^*, hence

$$\Psi(\tau,\eta) := \int_0^\tau F^*(B^*e^{-tA^*}\eta)\,dt.$$

In many situations it could well be possible to generate the values and even subgradients of Ψ numerically. The lower envelope representation of V in (4.12) could then, in light of Theorems 4.2 and 4.3, furnish an effective way of generating subgradients (or approximate subgradients) of V for potential use in feedback rules, through solving real-time optimization subproblems in $\mathbb{R}^n$.

The case in Example 5.2 is still relatively special within our framework. What might be said about value functions that come from state-dependent Hamiltonians more generally under our basic assumptions, as translated through Proposition 2.2? Everything really goes back to Theorem 2.6. The extent to which the basic formulas (2.13) and (2.14) in Theorem 2.6 can be regarded as “explicit” analogs of the classical Hopf–Lax formulas (5.6) and (5.7) hinges on how far one can go in obtaining an “explicit” formula for the dualizing kernel K that we have introduced. This requires an exploration of favorable cases in which the Hamilton–Jacobi characterization of K in Theorem 3.1 can be made to yield an “explicit” expression for K.

Here we have shown that the classical case in Example 5.1, where K is given by (5.1), can be extended with essentially no loss to the case in Example 5.2, where K is given by (5.9). Further research might yield other attractive cases. In the end, though, it must be borne in mind that the notion of what is an explicit expression for a function has evolved considerably in mathematics, and now is more a matter of whether a formula supports insightful analysis tied to modern computational methodology.

REFERENCES

[1] R.T. Rockafellar and P.R. Wolenski, *Convexity in Hamilton-Jacobi theory* I: *Dynamics and duality*, SIAM J. Control Optim., 39 (2000), pp. 1323–1350.

[2] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[3] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[4] R.T. Rockafellar and R.J-B Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.

[5] R.T. Rockafellar, *Linear-quadratic programming and optimal control*, SIAM J. Control Optim., 25 (1987), pp. 781–814.

[6] R.T. Rockafellar, *Hamiltonian trajectories and duality in the optimal control of linear systems with convex costs*, SIAM J. Control Optim., 27 (1989), pp. 1007–1025.

[7] R.T. Rockafellar, *Existence theorems for general control problems of Bolza and Lagrange*, Adv. Math., 15 (1975), pp. 315–333.

[8] H. Frankowska, *Hamilton-Jacobi equations: Viscosity solutions and generalized gradients*, J. Math. Anal. Appl., 141 (1989), pp. 21–26.

[9] M.G. Crandall, L.C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 478–502.

[10] H. Ishii, *A Comparison Result for Hamilton-Jacobi Equations Without Growth Condition on Solutions from Above*, preprint, 1998.

[11] R.T. Rockafellar, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–428.

[12] P.-L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Res. Notes Math. 69, Pitman, Boston, 1982.

[13] E. Hopf, *Generalized solutions of non-linear equations of first order*, J. Math. Mech., 14 (1965), pp. 201–230.

[14] P.D. Lax, *Hyperbolic Systems of Conservation Laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[15] L.C. Evans, *Partial Differential Equations*, Berkeley Mathematics Lecture Notes, Vols. 3A and 3B, University of California at Berkeley, 1993.

[16] M. Bardi and L.C. Evans, *On Hopf's formulas for solutions of Hamilton-Jacobi equations*, Nonlinear Anal., 8 (1984), pp. 1373–1381.

[17] O. Alvarez, E.N. Barron, and H. Ishii, *Hopf-Lax formulas for semicontinuous data*, Indiana Univ. Math. J., 48 (1999), pp. 993–1035.

[18] E.N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[19] H. Frankowska, *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control. Optim., 31 (1993), pp. 257–272.

[20] M. Arisawa and A. Tourin, *Regularizing effects for a class of first-order Hamilton-Jacobi equations*, Nonlinear Anal., 29 (1997), pp. 1405–1419.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1373–1390

QUADRATIC CONTROLLABILITY, STRONG CONTROLLABILITY, AND A RELATED OUTPUT FEEDBACK PROPERTY*

SHAN-MIN SWEI[†], TETSUYA IWASAKI[‡], AND MARTIN CORLESS[§]

Abstract. Quadratic controllability and strong controllability are the two system properties to be discussed for a special class of norm bounded uncertain linear systems. The main contribution of this paper is twofold. First we show that it is possible to reduce the problem of checking quadratic controllability or strong controllability of a given uncertain system to the same problem for a reduced order subsystem, which we call the *essential subsystem*. Then we use this result to examine a related output feedback property. Specifically, for a given uncertain system, we will answer the following question: What condition is necessary and sufficient for the existence of a dynamic output feedback controller such that the states of the closed-loop system converge to the origin with an arbitrarily fast rate? Our aim is to provide an insight and better understanding of uncertain systems through these results.

Key words. quadratic controllability, strong controllability, dynamic output feedback, quadratic feedback minimality

AMS subject classifications. 93B05, 15A03, 15A39

PII. S0363012999365108

1. Introduction.

The problem of quadratic stabilization of uncertain linear systems has attracted a number of researchers in the past and the results on this subject are abundant; see, for example, [1, 4, 6, 12, 13, 14, 15, 16, 19, 28], just to name a few. In the references [13, 14, 15, 28], the attention was paid to a special class of uncertain linear systems, namely, the systems with (unstructured) norm bounded uncertainties. One important feature of this class of uncertain systems is that the quadratic stabilization problem can be posed as the $\mathcal{H}_\infty$ control problem [8, 17]. Consequently, any standard tool for $\mathcal{H}_\infty$ control (e.g. [2, 3, 7, 22, 23]) can be directly applied to find a quadratically stabilizing controller. Thus, the quadratic stabilization problem has been solved completely for the case of systems with unstructured norm bounded uncertainties.

While the $\mathcal{H}_\infty$ control theory provides a tool for quadratic stabilization, one may lose the insight of the problem by blindly applying such a tool for control design. In this regard, the reference [25] has thoroughly studied the quadratic stabilization problem from a very different perspective in that it exploits the structure of the underlying system. An order reduction process which is similar to the back-stepping [9] was introduced for singular systems and was shown to preserve the quadratic stabilizability property. By repeating the reduction process, the original quadratic stabilization

*Received by the editors December 1, 1999; accepted for publication (in revised form) August 8, 2000; published electronically December 28, 2000.
http://www.siam.org/journals/sicon/39-5/36510.html

[†]School of Mechanical & Production Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798 (msmswei@ntu.edu.sg). The work of this author was sponsored by Nanyang Technological University and the Ministry of Education of Singapore under an AcRF research grant.

[‡]Department of Mechanical and Aerospace Engineering, University of Virginia, Charlottesville, VA 22904-4746, (iwasaki@virginia.edu). Formerly with Department of Control Systems Engineering, Tokyo Institute of Technology, Tokyo, Japan. The research of this author was supported by The Grant-in Aid for COE Research Project of Super Mechano-Systems by The Ministry of Education, Science, Sport, and Culture in Japan.

[§]School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN 47907 (corless @ecn.purdue.edu).

problem is equivalently transformed to that for a reduced order subsystem, which is much simpler.

In light of this equivalence between the original system and its reduced order subsystem, the notion of quadratic controllability was then introduced in [18]. Quadratic controllability was defined as quadratic stabilizability with an arbitrary rate of convergence. The conditions under which the norm bounded uncertain systems possess such a property were first presented in [24] for a class of scalar input systems and in [18] for more general multi-input systems. Moreover, the equivalence was established [18] in the sense of quadratic controllability between the original system and its reduced order subsystem.

The strong controllability is another important system property which was studied in relation to the (almost) disturbance decoupling problem for linear time-invariant systems, see [5, 10, 11, 20, 21, 22, 23, 26]. This property was defined geometrically in terms of invariant subspaces. The notion of strongly controllable subspace was used in [22, 23] to derive a solution to the singular $\mathcal{H}_\infty$ control problem, by reducing it to the almost disturbance decoupling problem. Roughly speaking, the almost disturbance decoupling problem is solvable if the image of the matrix, through which the disturbance enters the system, is contained in the strongly controllable subspace.

In this paper, we study the quadratic controllability and strong controllability properties through a system reduction process which is similar to but more concise than the one discussed in [18, 25]. Specifically, we give a necessary and sufficient condition for an uncertain linear system to be quadratically controllable or strongly controllable in terms of a reduced order subsystem, and show that strong controllability is stronger than quadratic controllability in general. In the latter half of the paper, we extend the notion of quadratic controllability to the dynamic output feedback case and provide a complete characterization in terms of reduced order subsystems. The matrix inequality solution to the $\mathcal{H}_\infty$ control problem and its connection to the quadratic stabilization will be frequently used in developing many of our results. The essence is briefly outlined as follows.

Let the control input matrix be denoted by B and the direct feedthrough matrix by D. A system order reduction process will be defined when the matrix $B(I-D^+D)$ neither has full row rank nor equals zero, where superscript $+$ denotes the Moore–Penrose inverse. The reduction process is repeated until the matrix "$B(I-D^+D)$" in the reduced order system becomes either a full rank matrix or a zero matrix. We call this resulting reduced order system the *essential subsystem*. Our contribution is to show that the original system is quadratically controllable or strongly controllable if and only if its essential subsystem is so.

A related output feedback property is then discussed for a class of uncertain linear systems. We first define the notion of quadratic feedback minimality of an uncertain system by the existence of a dynamic output feedback controller which quadratically stabilizes the system with any given exponential convergent rate. For systems without uncertainty, this notion can be translated into the property of simultaneous controllability and observability, i.e., the minimality. From this observation, it is tempting to conjecture that an uncertain system is quadratically feedback minimal if both the state feedback problem and its dual problem are quadratically controllable. We show that this conjecture is false. However, it will be shown that an additional orthogonality condition which relates these two problems is needed to make the statement valid.

This paper is organized as follows. In section 2 we review the notions of quadratic controllability and strong controllability, and some important preliminary results are presented for the case where the matrix $B(I - D^+D)$ either has full row rank or is a zero matrix. When the matrix $B(I - D^+D)$ is neither, an order reduction procedure is proposed in section 3 to obtain the reduced order essential subsystem for which the matrix $B(I - D^+D)$ either has full row rank or is a zero matrix. The equivalence in the sense of quadratic controllability or strong controllability is then established between the original system and its essential subsystem. In section 4 our main result shows a necessary and sufficient condition under which the uncertain system possesses the aforementioned quadratic feedback minimality property. We conclude this paper with some remarks in section 5.

We use the following standard notation in this paper. For a matrix A, A^t denotes its (complex conjugate) transpose. A left annihilator of A, denoted by $A^\perp$, is a matrix whose rows form a basis for the null space of A^t. The image and the spectral norm of A are respectively denoted by $\mathrm{Im}(A)$ and $\|A\|$, and $\mathrm{Re}(\cdot)$ means the real part of the argument.

2. Quadratic controllability and strong controllability.

Consider an uncertain system described by

$$\dot{x}(t) = A_\Delta x(t) + B_\Delta u(t), \tag{2.1}$$

where $x(t) \in \boldsymbol{R}^n$ is the state, $u(t) \in \boldsymbol{R}^m$ is the control input. The matrices A_Δ and B_Δ are defined by

$$A_\Delta := A + G\Delta C, \qquad B_\Delta := B + G\Delta D,$$

where $\Delta = \Delta(t, x) \in \boldsymbol{R}^{k\times\ell}$ is a matrix-valued function which may depend on the time and/or the states, and the matrices A, B, G, C, and D are real and known with compatible dimensions. Throughout the paper, we assume that the value of Δ is uncertain but is Lebesgue measurable with respect to the time and continuous with respect to the states. Furthermore, it is assumed that Δ belongs to a *norm bounded set* $\boldsymbol{\Delta}$ defined by

$$\boldsymbol{\Delta} := \{ \ \Delta \in \boldsymbol{R}^{k\times\ell} : \ \|\Delta\| \le 1 \ \}. \tag{2.2}$$

We call the uncertain system described by (2.1) and (2.2) with $\Delta \in \boldsymbol{\Delta}$ a norm bounded uncertain system.

DEFINITION 2.1. *The uncertain system given in* (2.1) *is called* regular *if the matrix* $B(I - D^+D) = 0$ *and is called* simple *if the matrix* $B(I - D^+D)$ *has full row rank.*

The physical implication of Definition 2.1 is that if the control input u is orthogonally decoupled into two parts, namely,

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} := \begin{bmatrix} I - D^+D \\ D^+D \end{bmatrix} u,$$

then with this input description, we will have

$$B_\Delta u = Bu_1 + (B + G\Delta D)u_2, \tag{2.3}$$

which indicates that u_1 is the part of the control input that affects the system without direct influence of the uncertainty Δ, whereas u_2 enters the system through the uncertainty-dependent matrix B_Δ. If a system is regular, then only u_2 is effective, while a simple system receives full control authority through u_1.

Another important observation from Definition 2.1 is that the uncertain system (2.1) is regular if and only if $\mathrm{Im}(B^t) \subseteq \mathrm{Im}(D^t)$, since $D(I - D^+D) = 0$. On the other hand, (2.1) is simple if and only if B has full row rank and $\mathrm{Im}(D^t) \cap \mathrm{Im}(B^t) = \{0\}$.

Definition 2.2. *The uncertain system described in* (2.1) *is said to be* quadratically controllable *(QC) if, for each $\alpha > 0$, there exist a positive definite symmetric matrix $P \in \boldsymbol{R}^{n\times n}$ and a state feedback gain $K \in \boldsymbol{R}^{m\times n}$ such that*

$$P(A_\Delta + B_\Delta K + \alpha I) + (A_\Delta + B_\Delta K + \alpha I)^t P < 0\,,$$

for all $\Delta \in \boldsymbol{\Delta}$.

The notion of quadratic controllability was first introduced in [18]. Basically, it defines that an uncertain system is QC if, for each $\alpha > 0$, it is quadratically stabilizable with the convergent rate α. In other words, it is QC if there exists a state feedback gain K such that the resulting closed-loop system admits a *single* quadratic Lyapunov function (independent of Δ) which proves stability with rate of convergence α against all possible $\Delta \in \boldsymbol{\Delta}$. For systems without uncertainty, i.e., $\dot{x}(t) = Ax(t) + Bu(t)$, the notion of quadratic controllability means the assignability of the closed-loop eigenvalues with arbitrarily large negative real parts, i.e., the controllability of the pair (A,B).

It is important to note that the uncertain system given in (2.1) can be alternatively described as the feedback connection of a nominal system with an uncertain block, which is given by

$$\Sigma: \quad \begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Gw(t), \\ z(t) = Cx(t) + Du(t), \\ w(t) = \Delta z(t), \end{cases} \tag{2.4}$$

where the two new variables, $w(t) \in \boldsymbol{R}^k$ and $z(t) \in \boldsymbol{R}^\ell$, denote a fictitious disturbance input and a fictitious controlled output, respectively.

The following lemma characterizes QC systems in terms of matrix inequalities.

Lemma 2.3. *Consider the uncertain system Σ described in* (2.4). *The following statements are equivalent.*

(i) *The uncertain system Σ is QC.*

(ii) *For each $\alpha > 0$, there exists $X^t = X > 0$ such that*

$$\begin{bmatrix} B \\ D \end{bmatrix}^\perp \begin{bmatrix} (A+\alpha I)X + X(A+\alpha I)^t + GG^t & XC^t \\ CX & -I \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^{\perp t} < 0. \tag{2.5}$$

(iii) *For each $\alpha > 0$, there exists $X^t = X > 0$ such that*

$$E^\perp[(\tilde{A} + \alpha I)X + X(\tilde{A} + \alpha I)^t + X\tilde{C}X - \tilde{B}]E^{\perp t} < 0, \tag{2.6}$$

where

$$\begin{aligned} \tilde{A} &:= A - BD^+C, & \tilde{B} &:= B(D^tD)^+B^t - GG^t, \\ \tilde{C} &:= C^t(I - DD^+)C, & E &:= B(I - D^+D). \end{aligned} \tag{2.7}$$

Proof. The equivalence of (i) and (ii) simply follows from the quadratic stabilizability result [8] and the state feedback $\mathcal{H}_\infty$ synthesis result [7]. The equivalence of (ii) and (iii) can be verified by noting that

$$\begin{bmatrix} B \\ D \end{bmatrix}^\perp = \begin{bmatrix} E \\ D \end{bmatrix}^\perp \begin{bmatrix} I & -BD^+ \\ 0 & I \end{bmatrix}$$

and

$$\begin{bmatrix} E \\ D \end{bmatrix}^{\perp} = \left(\begin{bmatrix} E \\ D \end{bmatrix} \begin{bmatrix} E \\ D \end{bmatrix}^t \right)^{\perp} = \begin{bmatrix} EE^t & 0 \\ 0 & DD^t \end{bmatrix}^{\perp} = \begin{bmatrix} E^{\perp} & 0 \\ 0 & D^{\perp} \end{bmatrix}$$

and using the Schur complement. □

Recall that the matrix E in Lemma 2.3 is the uncertainty-free input matrix through which the control input u_1 enters the system as shown in (2.3). If Σ is regular, then $E^{\perp} = I$ and (2.6) becomes a standard Riccati inequality. On the other hand, if Σ is simple, then $E^{\perp}$ becomes trivial; hence (2.6) is trivially satisfied. An immediate consequence of this is that the system Σ in (2.4) is QC if it is simple.

The following lemma shows when an uncertain system is QC. A similar result was obtained in [18, 24]. For completeness, a proof is included.

LEMMA 2.4. *Consider the uncertain system Σ described in* (2.4). *The following statements hold true.*

(a) *If Σ is simple, then it is QC.*

(b) *If Σ is regular, then it is QC if and only if*

$$\tilde{B} \geq 0 \quad \textit{and} \quad (\tilde{A}, \tilde{B}) \ \textit{controllable}, \tag{2.8}$$

where $\tilde{A}$ and $\tilde{B}$ are defined in (2.7).

Proof. Statement (a) is obvious as noted above, and we need only to prove (b).

By the regularity assumption, Lemma 2.3 implies that the uncertain system is QC if and only if, for each $\alpha > 0$, there exists $X^t = X > 0$ such that

$$(\tilde{A} + \alpha I)X + X(\tilde{A} + \alpha I)^t + X\tilde{C}X - \tilde{B} < 0 \tag{2.9}$$

where $\tilde{A}$, $\tilde{B}$, and $\tilde{C} \geq 0$ are defined in (2.7).

Suppose Σ is QC. Then, from Lemma 5.1 in the appendix, we have $\tilde{B} \geq 0$. Suppose $(\tilde{A}$,$\tilde{B})$ is not controllable, then there exist a complex number λ and a nonzero vector v such that

$$v^t(\tilde{A} - \lambda I) = 0, \quad v^t\tilde{B} = 0.$$

Fix such λ and v, and pre- and postmultiply (2.9) by v^t and v, respectively, and we obtain that for each $\alpha > 0$, there exists $X^t = X > 0$ satisfying

$$2[\alpha + \mathrm{Re}(\lambda)]v^tXv + v^tX\tilde{C}Xv < 0\,. \tag{2.10}$$

If we choose $\alpha > 0$ to be sufficiently large such that $\alpha + \mathrm{Re}(\lambda) \geq 0$, then (2.10) implies that $v^tX\tilde{C}Xv < 0$, which contradicts that $\tilde{C} \geq 0$. Hence, $(\tilde{A}$,$\tilde{B})$ must be controllable.

Conversely, suppose $\tilde{B} \geq 0$ and $(\tilde{A}$,$\tilde{B})$ is controllable. Then, for each $\alpha > 0$, there exist a gain matrix K such that $\tilde{A} + \tilde{B}K + \alpha I$ is Hurwitz and a $P^t = P > 0$ satisfying the following Lyapunov inequality:

$$P(\tilde{A} + \tilde{B}K + \alpha I) + (\tilde{A} + \tilde{B}K + \alpha I)^tP + K^t\tilde{B}K + \tilde{C} < 0.$$

Completing the square with respect to K, we obtain

$$P(\tilde{A} + \alpha I) + (\tilde{A} + \alpha I)^tP + \tilde{C} - P\tilde{B}P < -(P+K)^t\tilde{B}(P+K) \leq 0.$$

Now, it is straightforward to verify that $X := P^{-1}$ satisfies (2.9). Thus the system is QC. □

It is easy to derive, under the regularity assumption on Σ, that the first condition in (2.8) is equivalent to the existence of a matrix M, such that

$$G = BM \quad \text{and} \quad \|DM\| \leq 1,$$

which is precisely the *matching condition* obtained in [18]. Intuitively, this condition implies that it is possible to completely compensate the effect of the uncertainty Δ entering the system through G by using the control input through B.

Next we introduce the notion of strong controllability.

DEFINITION 2.5. The strongly controllable subspace *of the uncertain system* (2.1) *is the smallest subspace* $\mathbf{S}$ *of* $\boldsymbol{R}^n$ *for which there exists a matrix* K *such that*

$$(A + KC)\mathbf{S} \subset \mathbf{S}, \qquad \mathrm{Im}(B + KD) \subset \mathbf{S}.$$

The system is said to be strongly controllable *(SC) if its strongly controllable subspace is equal to the whole state space.*

The notion of strong controllability was introduced and studied in the references [5, 10, 11, 20]. It was considered in the context of invariant properties of linear systems, rather than in the robustness analysis context which is our main focus in this paper. It was shown in [26] for the case where $D = 0$ that if Σ is SC, then the problem of almost disturbance decoupling with an arbitrary degree of stability (or rate of convergence) for Σ is solvable. Therefore, the notion of strong controllability is stronger than that of quadratic controllability when $D = 0$.

The following lemma [5, 20] characterizes the strong controllability property in terms of a rank condition.

LEMMA 2.6. *The uncertain system* Σ *is SC if and only if*

$$\mathrm{rank} \begin{bmatrix} A - sI_n & B \\ C & D \end{bmatrix} = n + \mathrm{rank} \begin{bmatrix} C & D \end{bmatrix} \tag{2.11}$$

holds for all $s \in \boldsymbol{C}$.

The characterization of strong controllability given in Lemma 2.6 implies that the system Σ has no finite invariant zeros. If there is no uncertainty in the system, i.e., when $C = 0$ and $D = 0$, then the condition (2.11) reduces to

$$\mathrm{rank} \begin{bmatrix} A - sI_n & B \end{bmatrix} = n \; \forall \, s \in \boldsymbol{C}, \tag{2.12}$$

which is the Popov–Belevitch–Hautus (PBH) rank test for the controllability of the pair (A, B). In this case, strong controllability is equivalent to quadratic controllability and the SC subspace given in Definition 2.5 reduces to the smallest A-invariant subspace containing $\mathrm{Im}(B)$, which is precisely the controllability subspace defined in [27].

The following lemma shows some implications of the uncertain system Σ being simple or regular in relation to the notion of strong controllability.

LEMMA 2.7. *Consider the uncertain system* Σ *described in* (2.4). *The following statements hold true.*

(a) *If* Σ *is simple, then it is SC.*

(b) *If* Σ *is regular, then it is not SC.*

Proof. First we prove (a). Since $B(I - D^+D)$ has full row rank n,

$$\mathrm{rank} \begin{bmatrix} A - sI_n & B \\ C & D \end{bmatrix} = \mathrm{rank} \left(\begin{bmatrix} A - sI_n & B \\ C & D \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & D^+D & I - D^+D \end{bmatrix} \right)$$

$$= \operatorname{rank} \begin{bmatrix} A - sI_n & BD^+D & B(I - D^+D) \\ C & D & 0 \end{bmatrix}$$
$$= n + \operatorname{rank} \begin{bmatrix} C & D \end{bmatrix}.$$

To prove (b), we note that since $B = BD^+D$, we have

$$\begin{aligned} \operatorname{rank} \begin{bmatrix} A - sI_n & B \\ C & D \end{bmatrix} &= \operatorname{rank} \left(\begin{bmatrix} I & -BD^+ \\ 0 & I \end{bmatrix} \begin{bmatrix} A - sI_n & BD^+D \\ C & D \end{bmatrix} \right) \\ &= \operatorname{rank} \begin{bmatrix} A - BD^+C - sI_n & 0 \\ C & D \end{bmatrix} \\ &\leq \operatorname{rank}(A - BD^+C - sI_n) + \operatorname{rank} \begin{bmatrix} C & D \end{bmatrix} \\ &< n + \operatorname{rank} \begin{bmatrix} C & D \end{bmatrix}, \end{aligned}$$

where the last strict inequality holds when $s \in \boldsymbol{C}$ is chosen to be an eigenvalue of $A - BD^+C$. □

Note that Lemmas 2.4 and 2.7 completely characterize the QC and SC properties of an uncertain system when it is either simple or regular. However, these results shed little light on uncertain systems which are neither simple nor regular. In the next section, we will introduce the notion of *essential subsystem*, which plays a central role in considering such general cases.

3. Characterization of quadratic controllability and strong controllability via the essential subsystem.

Suppose the uncertain system Σ in (2.4) is neither simple nor regular. In what follows, we will define a reduced order subsystem of Σ which is obtained through state and control input transformations, followed by state truncations. The main features of this subsystem are as follows: (a) it has smaller dimension than the original system, (b) it is either simple or regular, and (c) quadratic controllability or strong controllability of the original system can be verified by checking quadratic controllability or strong controllability of this reduced order subsystem. We call this subsystem of Σ the *essential subsystem*.

3.1. The essential subsystem.

Suppose the system Σ is neither simple nor regular. Let N_0 and R_0 be matrices whose columns form orthonormal bases for the left null space and the range space of $B(I - D^+D)$, respectively. Using the following state and input transformations,

$$\begin{bmatrix} \xi_n \\ \xi_r \end{bmatrix} := \begin{bmatrix} N_0^t \\ R_0^t \end{bmatrix} x, \qquad \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} := \begin{bmatrix} I - D^+D \\ D^+D \end{bmatrix} u,$$

the nominal part of Σ in (2.4) will have the following form:

$$\left\{ \begin{array}{lcl} \dot{\xi}_n &=& A_{nn}\xi_n + A_{nr}\xi_r \qquad\qquad + B_n u_2 + G_n w, \\ \dot{\xi}_r &=& A_{rn}\xi_n + A_{rr}\xi_r + \hat{B}_r u_1 + B_r u_2 + G_r w, \\ z &=& C_n\xi_n + C_r\xi_r \qquad\qquad + D\, u_2, \end{array} \right.$$

where

$$\begin{bmatrix} A_{nn} & A_{nr} \\ A_{rn} & A_{rr} \end{bmatrix} := \begin{bmatrix} N_0^t \\ R_0^t \end{bmatrix} A \begin{bmatrix} N_0 & R_0 \end{bmatrix},$$

$$\begin{bmatrix} 0 & B_n \\ \hat{B}_r & B_r \end{bmatrix} := \begin{bmatrix} N_0^t \\ R_0^t \end{bmatrix} B \begin{bmatrix} I - D^+D & D^+D \end{bmatrix},$$

$$\begin{bmatrix} G_n \\ G_r \end{bmatrix} := \begin{bmatrix} N_0^t \\ R_0^t \end{bmatrix} G, \quad \begin{bmatrix} C_n & C_r \end{bmatrix} := C \begin{bmatrix} N_0 & R_0 \end{bmatrix}.$$

Moreover, $\hat{B}_r$ is a full row rank matrix. It should be noted from the second equation that ξ_r can be fully controlled as desired through u_1, and therefore it can be regarded as part of control input in the first equation. This motivates us to consider the following reduced order subsystem of Σ:

$$\Sigma_1 : \left\{ \begin{array}{rcl} \dot{\zeta}(t) & = & A_1\zeta(t) + B_1 v(t) + G_1 w(t), \\ z(t) & = & C_1\zeta(t) + D_1 v(t), \\ w(t) & = & \Delta z(t), \end{array} \right. \tag{3.1}$$

where $\zeta := \xi_n$ is the state and $v := \begin{bmatrix} \xi_r \\ u_2 \end{bmatrix}$ is the control input, and

$$\begin{aligned} \begin{bmatrix} A_1 & B_1 & G_1 \\ C_1 & D_1 & 0 \end{bmatrix} &:= \left[\begin{array}{c|cc|c} A_{nn} & A_{nr} & B_n & G_n \\ \hline C_n & C_r & D & 0 \end{array} \right] \\ &= \begin{bmatrix} N_0^t & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B & G \\ C & D & 0 \end{bmatrix} \begin{bmatrix} N_0 & R_0 & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}. \end{aligned}$$

It is worth noting that $B_n = N_0^t B D^+ D = N_0^t B$, since $N_0^t B(I - D^+ D) = 0$. Clearly, the order of Σ_1 is lower than that of Σ, and it could be either simple or regular. If Σ_1 is neither, the reduction procedure can be applied again to obtain a reduced order subsystem of Σ_1. We may repeat this procedure to generate a sequence of subsystems $\Sigma_1, \ldots, \Sigma_q$ until Σ_q becomes either simple or regular. We call Σ_q the *essential subsystem* of Σ. In the subsequent sections, we show that the quadratic controllability and strong controllability properties are preserved in the reduction process and hence Σ is QC and/or SC if and only if Σ_q is also.

The procedure to obtain Σ_q can be formalized as follows.

Algorithm 1

Given the system Σ, define its essential subsystem Σ_q from the following iteration steps.

Step 0. Let $i = 0$ and set the initial system $\Sigma_0 := \Sigma$, i.e.,

$$\begin{bmatrix} A_0 & B_0 & G_0 \\ C_0 & D_0 & 0 \end{bmatrix} := \begin{bmatrix} A & B & G \\ C & D & 0 \end{bmatrix}.$$

Step 1. If Σ_i is either simple or regular, then go to Step 3; otherwise define the subsystem Σ_{i+1} by

$$\begin{bmatrix} A_{i+1} & B_{i+1} & G_{i+1} \\ C_{i+1} & D_{i+1} & 0 \end{bmatrix} := \begin{bmatrix} N_i^t & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_i & B_i & G_i \\ C_i & D_i & 0 \end{bmatrix} \left[\begin{array}{c|cc|c} N_i & R_i & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{array} \right],$$

where N_i and R_i are matrices whose columns form orthonormal bases for the left null space and the range space of $B_i(I - D_i^+ D_i)$, respectively.

Step 2. Set $i = i + 1$ and go to Step 1.

Step 3. Let $q := i$ and $\Sigma_q := \Sigma_i$.

Writing down the above iterative process in the closed form, we see that the essential subsystem can be characterized as

$$\begin{bmatrix} A_q & B_q & G_q \\ C_q & D_q & 0 \end{bmatrix} := \begin{bmatrix} N^t & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B & G \\ C & D & 0 \end{bmatrix} \begin{bmatrix} N & * & 0 \\ 0 & * & 0 \\ 0 & 0 & I \end{bmatrix},$$

where $*$ denotes a known entry and

$$N := \begin{cases} I & \text{if } q = 0, \\ N_0 N_1 \cdots N_{q-1} & \text{if } q \geq 1. \end{cases} \tag{3.2}$$

Thus, it can be seen that the essential subsystem Σ_q is obtained through the projection of the original system Σ onto $\mathrm{Im}(N)$. Define a matrix $\mathcal{N}$ by

$$\mathcal{N} := \begin{cases} N & \text{if } \Sigma_q \text{ is regular,} \\ 0 & \text{if } \Sigma_q \text{ is simple.} \end{cases} \tag{3.3}$$

The importance of matrix $\mathcal{N}$ will become eminent when we discuss an output feedback property in section 4.

3.2. Characterization of QC.

The following is the main result of this subsection.

THEOREM 3.1. *Consider the uncertain system Σ described in* (2.4)*. Define the essential subsystem Σ_q by Algorithm* 1*. Then Σ is QC if and only if Σ_q is QC.*

Before we proceed further, some remarks are in order. If Σ_q is simple, then it follows from Lemma 2.4 that Σ_q is QC. On the other hand, if Σ_q is regular, then the matching condition in Lemma 2.4 is necessary (and sufficient) for Σ_q to be QC.

The following lemma is instrumental to the proof of Theorem 3.1, and it shows that the proposed order reduction process preserves the QC property.

LEMMA 3.2. *Consider the uncertain system Σ described in* (2.4) *and its reduced order subsystem Σ_1 defined in* (3.1)*. The system Σ is QC if and only if its subsystem Σ_1 is QC. Moreover, if for each $\alpha > 0$, there exists $X^t = X > 0$ satisfying inequality* (2.6) *for the original system Σ, then $X_1 := N_0^t X N_0$ satisfies the same inequality for the subsystem Σ_1.*

Proof. Note that $U_0 := \begin{bmatrix} N_0 & R_0 \end{bmatrix}$ is an orthogonal matrix and $E^\perp = N_0^t$. Define

$$\begin{bmatrix} \tilde{A}_{nn} & \tilde{A}_{nr} \\ \tilde{A}_{rn} & \tilde{A}_{rr} \end{bmatrix} := U_0^t \tilde{A} U_0, \quad \begin{bmatrix} X_{nn} & X_{nr} \\ X_{nr}^t & X_{rr} \end{bmatrix} := U_0^t X U_0, \tag{3.4a}$$

$$\begin{bmatrix} B_n \\ B_r \end{bmatrix} := U_0^t B, \quad \begin{bmatrix} G_n \\ G_r \end{bmatrix} := U_0^t G, \quad \begin{bmatrix} C_n & C_r \end{bmatrix} := C U_0. \tag{3.4b}$$

Then, after some algebraic manipulations, (2.6) can be rewritten as

$$\begin{aligned} \Phi &:= (\tilde{A}_{nn} X_{nn} + \tilde{A}_{nr} X_{nr}^t) + (\tilde{A}_{nn} X_{nn} + \tilde{A}_{nr} X_{nr}^t)^t + G_n G_n^t - B_n (D^t D)^+ B_n^t \\ &\quad + 2\alpha X_{nn} + (C_n X_{nn} + C_r X_{nr}^t)^t (I - DD^+)(C_n X_{nn} + C_r X_{nr}^t) < 0. \end{aligned} \tag{3.5}$$

Moreover, (3.5) holds if and only if there exists a K such that

$$\Phi + (B_n D^+ + KDD^+)(B_n D^+ + KDD^+)^t < 0.$$

It is straightforward to verify that the above inequality can be equivalently written as

$$\begin{aligned} &\left(A_{nn} X_{nn} + \begin{bmatrix} A_{nr} & B_n D^+ \end{bmatrix} \begin{bmatrix} X_{nr}^t \\ \tilde{K}^t \end{bmatrix}\right) + \left(A_{nn} X_{nn} + \begin{bmatrix} A_{nr} & B_n D^+ \end{bmatrix} \begin{bmatrix} X_{nr}^t \\ \tilde{K}^t \end{bmatrix}\right)^t + G_n G_n^t \\ &+ 2\alpha X_{nn} + \left(C_n X_{nn} + \begin{bmatrix} C_r & DD^+ \end{bmatrix} \begin{bmatrix} X_{nr}^t \\ \tilde{K}^t \end{bmatrix}\right)^t \left(C_n X_{nn} + \begin{bmatrix} C_r & DD^+ \end{bmatrix} \begin{bmatrix} X_{nr}^t \\ \tilde{K}^t \end{bmatrix}\right) < 0, \end{aligned} \tag{3.6}$$

where A_{nn} and A_{nr} are defined similarly to $\tilde{A}_{nn}$ and $\tilde{A}_{nr}$ in that A replaces $\tilde{A}$, and

$$\tilde{K} := K - X_{nn}C_n^t - X_{nr}C_r^t.$$

Taking the Schur complement of (3.6), we obtain

$$\begin{bmatrix} A_{nn}X_{nn} + X_{nn}A_{nn}^t + 2\alpha X_{nn} + G_nG_n^t & X_{nn}C_n^t \\ C_nX_{nn} & -I \end{bmatrix} + \begin{bmatrix} A_{nr} & B_nD^+ \\ C_r & DD^+ \end{bmatrix} \begin{bmatrix} X_{nr}^t \\ \tilde{K}^t \end{bmatrix} \begin{bmatrix} I & 0 \end{bmatrix}$$

$$+ \begin{bmatrix} I \\ 0 \end{bmatrix} \begin{bmatrix} X_{nr} & \tilde{K} \end{bmatrix} \begin{bmatrix} A_{nr} & B_nD^+ \\ C_r & DD^+ \end{bmatrix}^t < 0. \tag{3.7}$$

It then follows from the projection lemma presented in [3, 7] that there exist X_{nr} and $\tilde{K}$ which satisfy the above inequality if and only if

$$\begin{bmatrix} A_{nr} & B_nD^+ \\ C_r & DD^+ \end{bmatrix}^{\perp} \begin{bmatrix} A_{nn}X_{nn} + X_{nn}A_{nn}^t + 2\alpha X_{nn} + G_nG_n^t & X_{nn}C_n^t \\ C_nX_{nn} & -I \end{bmatrix} \begin{bmatrix} A_{nr} & B_nD^+ \\ C_r & DD^+ \end{bmatrix}^{\perp t} < 0.$$

Finally, since

$$\operatorname{Im} \begin{bmatrix} B_n \\ D \end{bmatrix} = \operatorname{Im} \begin{bmatrix} B_nD^+D \\ DD^+D \end{bmatrix} = \operatorname{Im} \begin{bmatrix} B_nD^+ \\ DD^+ \end{bmatrix},$$

we obtain that

$$\begin{bmatrix} A_{nr} & B_nD^+ \\ C_r & DD^+ \end{bmatrix}^{\perp} = \begin{bmatrix} A_{nr} & B_n \\ C_r & D \end{bmatrix}^{\perp}.$$

Hence the above inequality is precisely the condition for the subsystem Σ_1 to be QC. □

We now prove Theorem 3.1.

Proof. In view of Lemma 3.2, the quadratic controllability property is invariant under the proposed order reduction process. Hence the uncertain system Σ is QC if and only if its essential subsystem Σ_q is QC. □

3.3. Characterization of strong controllability. Consider the uncertain system Σ given in (2.4) and its essential subsystem Σ_q. In this subsection, we will establish the equivalence between Σ and Σ_q in the sense of strong controllability. To show this, we rely on the characterization of strong controllability which was given by the rank condition in Lemma 2.6. First, we will examine the strong controllability property of Σ in terms of Σ_1.

LEMMA 3.3. *Consider the uncertain system Σ described in* (2.4) *and its reduced order subsystem Σ_1 defined in* (3.1)*. The rank condition*

$$\operatorname{rank} \begin{bmatrix} A - sI_n & B \\ C & D \end{bmatrix} - n = \operatorname{rank} \begin{bmatrix} A_1 - sI_{n_1} & B_1 \\ C_1 & D_1 \end{bmatrix} - n_1 \tag{3.8}$$

holds for all $s \in \boldsymbol{C}$, where n_1 denotes the dimension of Σ_1.

Proof. Let r be the rank of $B(I - D^+D)$. Since $n = n_1 + r$, we obtain

$$\operatorname{rank} \begin{bmatrix} A - sI_n & B \\ C & D \end{bmatrix} = \operatorname{rank} \left(\begin{bmatrix} N_0^t & 0 \\ 0 & I \\ R_0^t & 0 \end{bmatrix} \begin{bmatrix} A - sI_n & B \\ C & D \end{bmatrix} \left[\begin{array}{c|cc|c} N_0 & R_0 & 0 & 0 \\ 0 & 0 & I & I - D^+D \end{array} \right] \right)$$

$$= \operatorname{rank} \begin{bmatrix} A_1 - sI_{n_1} & B_1 & 0 \\ C_1 & D_1 & 0 \\ * & * & R_0^t B(I - D^+ D) \end{bmatrix}$$

$$= \operatorname{rank} \begin{bmatrix} A_1 - sI_{n_1} & B_1 \\ C_1 & D_1 \end{bmatrix} + n - n_1,$$

where the last equality follows from the fact that the matrix $R_0^t B(I - D^+ D)$ has full row rank $(n - n_1)$. □

It follows readily from the above lemma that $s \in \boldsymbol{C}$ is an invariant zero for the original system Σ if and only if it is an invariant zero for the reduced order subsystem Σ_1. Thus, the invariant zeros of the original system are preserved under the order reduction process and they are the invariant zeros of the essential subsystem Σ_q. Furthermore, by definition, for Σ_q to be SC it must not have finite invariant zeros. This observation is elaborated in the next theorem, which is the main result of this subsection.

THEOREM 3.4. *Consider the uncertain system* Σ *in* (2.4) *and its essential subsystem* Σ_q *defined by Algorithm* 1. *Then* Σ *is SC if and only if* Σ_q *is SC.*

Proof. If Σ is SC, then it follows from Lemma 2.6 that the rank condition (2.11) holds for all $s \in \boldsymbol{C}$. Moreover, from Lemma 3.3, we note that the order reduction process preserves the quantities

$$\operatorname{rank} \begin{bmatrix} A_i - sI_{n_i} & B_i \\ C_i & D_i \end{bmatrix} - n_i \qquad \text{and} \qquad \operatorname{rank} \begin{bmatrix} C_i & D_i \end{bmatrix}, \tag{3.9}$$

where n_i denotes the dimension of subsystem Σ_i. Hence

$$\operatorname{rank} \begin{bmatrix} A_i - sI_{n_i} & B_i \\ C_i & D_i \end{bmatrix} = n_i + \operatorname{rank} \begin{bmatrix} C_i & D_i \end{bmatrix} \ \forall s \in \boldsymbol{C}, \tag{3.10}$$

holds for all $i = 1, \ldots, q$. Thus, from Lemma 2.6, we see that the essential subsystem Σ_q is SC. Conversely, if Σ_q is SC, then it satisfies the rank condition in (3.10) with $i = q$. Again, the quantities in (3.9) are preserved for each $i < q$ and hence Σ must be SC. □

In light of Lemma 2.7 and the fact that Σ_q is either simple or regular, we note that Σ_q is SC if and only if it is simple. It can be inferred from Theorems 3.1 and 3.4 that strong controllability generally implies quadratic controllability, but not vice versa, since an uncertain system can be QC even if its essential subsystem is not simple.

4. A related output feedback property.

We have considered the quadratic controllability property for a class of uncertain linear systems described in (2.1), provided that all the states are available for linear feedback design. This section, however, is concerned with the uncertain systems in which only the measured states are available for feedback, and in addition, they are contaminated by the uncertainties. For this class of uncertain systems, our objective is to determine a necessary and sufficient condition under which there exists a dynamic output feedback controller that will render the closed-loop system which is quadratically stable with any prescribed rate of convergence. We call an uncertain system with such property *quadratically feedback minimal.*

4.1. Notion of quadratic feedback minimality.

Consider the uncertain system described in (2.1) with measured output

$$\begin{cases} \dot{x}(t) &= A_\Delta x(t) + B_\Delta u(t), \\ y(t) &= M_\Delta x(t) + J_\Delta u(t), \end{cases} \tag{4.1}$$

where $y(t) \in \boldsymbol{R}^p$ denotes the measured output, and M_Δ and J_Δ are defined by

$$M_\Delta := M + J\Delta C\,,\ J_\Delta := J\Delta D\,,$$

where the matrices M and J are real and known with compatible dimensions, and Δ belongs to $\boldsymbol{\Delta}$. This uncertain system is to be controlled by a linear, time-invariant, dynamic, output feedback controller of the form

$$\left\{ \begin{array}{rcl} \dot{x}_c(t) & = & A_c x_c(t) + B_c y(t), \\ u(t) & = & C_c x_c(t), \end{array} \right. \tag{4.2}$$

where $x_c(t) \in \boldsymbol{R}^{n_c}$ is the state of the controller, and A_c, B_c, and C_c are real matrices with compatible dimensions. The closed-loop system, which consists of the plant (4.1) and the controller (4.2), is given by $\dot{x}_{c\ell} = \mathcal{A}_\Delta x_{c\ell}$ with

$$x_{c\ell} := \left[\begin{array}{c} x \\ x_c \end{array} \right], \quad \mathcal{A}_\Delta := \left[\begin{array}{cc} A_\Delta & B_\Delta C_c \\ B_c M_\Delta & A_c + B_c J_\Delta C_c \end{array} \right].$$

DEFINITION 4.1. *The uncertain system described in* (4.1) *is said to be* quadratically feedback minimal *(QFM) if for any given $\alpha > 0$, there exist a dynamic controller of the form* (4.2) *and a positive-definite symmetric matrix $P \in \boldsymbol{R}^{(n+n_c)\times(n+n_c)}$ such that*

$$P(\mathcal{A}_\Delta + \alpha I) + (\mathcal{A}_\Delta + \alpha I)^t P < 0 \tag{4.3}$$

for all $\Delta \in \boldsymbol{\Delta}$.

For linear, time-invariant, certain systems (i.e., when Δ is fixed and known), the notion given in Definition 4.1 is equivalent to the assignability (using dynamic output feedback controllers) of the closed-loop eigenvalues with arbitrarily large negative real parts. As is well known, the class of certain systems having such a property coincides with the class of systems that are controllable and observable, i.e., minimal; hence the term QFM. In view of this fact, it is tempting to conjecture that the class of *uncertain* systems having such a property can also be characterized as the class of systems whose state feedback part is QC and its dual part is also QC. However, our result will show that an additional orthogonality condition has to be satisfied.

In view of the uncertain system Σ given in (2.4), we note that the uncertain system (4.1) can be equivalently described by

$$\Sigma_{of}\ :\ \left\{ \begin{array}{rcll} \dot{x}(t) & = & Ax(t) + Bu(t) & +\ Gw(t), \\ y(t) & = & Mx(t) & +\ Jw(t), \\ z(t) & = & Cx(t) + Du(t), & \\ w(t) & = & \Delta z(t). & \end{array} \right. \tag{4.4}$$

Given the system Σ_{of}, we define two auxiliary systems as follows:

$$\Sigma\ :\ \left\{ \begin{array}{rcl} \dot{x}(t) & = & Ax(t) + Bu(t) + Gw(t), \\ z(t) & = & Cx(t) + Du(t), \\ w(t) & = & \Delta z(t) \end{array} \right. \tag{4.5}$$

and

$$\hat{\Sigma}\ :\ \left\{ \begin{array}{rcl} \dot{\hat{x}}(t) & = & \hat{A}\hat{x}(t) + \hat{B}\hat{u}(t) + \hat{G}\hat{w}(t), \\ \hat{z}(t) & = & \hat{C}\hat{x}(t) + \hat{D}\hat{u}(t), \\ \hat{w}(t) & = & \hat{\Delta}\hat{z}(t), \end{array} \right. \tag{4.6}$$

where $(\hat{A}, \hat{B}, \hat{G}, \hat{C}, \hat{D}) := (A^t, M^t, C^t, G^t, J^t)$ and $\hat{\Delta} := \Delta^t$. The system Σ in (4.5) is precisely the system considered in the previous sections, which defines a state feedback problem, whereas the system $\hat{\Sigma}$ defines another state feedback problem which is the dual of the state estimation problem for Σ_{of}. These two systems will play essential roles in developing the main result of this section. In order to facilitate our presentation, we make use of the following notions. For system $\hat{\Sigma}$, we can also apply Algorithm 1 to attain its essential subsystem. Assume that the essential subsystem $\hat{\Sigma}_{\hat{q}}$ is obtained at $i = \hat{q}$.

4.2. QFM systems. The main result of this section is contained in the following theorem.

THEOREM 4.2. *Consider the uncertain system* Σ_{of} *described in* (4.4). *Let the auxiliary systems* Σ *and* $\hat{\Sigma}$ *be given by* (4.5) *and* (4.6). *Define a matrix* $\mathcal{N}$ *for* Σ *as described in* (3.3), *and similarly a matrix* $\hat{\mathcal{N}}$ *for* $\hat{\Sigma}$. *Then,* Σ_{of} *is quadratically feedback minimal if and only if the following conditions hold.*

(a) Σ *is QC*,

(b) $\hat{\Sigma}$ *is QC*, *and*

(c) $\hat{\mathcal{N}}^t\mathcal{N} = 0$.

We will prove this theorem in the next subsection. For now, let us elaborate on the result. For systems without uncertainties, it is well known that Σ_{of} is QFM if and only if both (A,B) and (A^t,C^t) are controllable. For uncertain systems, in view of Theorem 4.2, these conditions are replaced by the quadratic controllability properties of Σ and $\hat{\Sigma}$ as in (a) and (b), and in addition, we have the orthogonality condition (c).

In order to verify the conditions in Theorem 4.2, we need to go through the order reduction process to obtain the essential subsystems for both Σ and $\hat{\Sigma}$, namely, Σ_q and $\hat{\Sigma}_{\hat{q}}$. Depending on whether each of these essential subsystems is simple or regular, we may have to examine four different cases, as shown in Table 4.1, to determine if the system Σ_{of} is QFM.

TABLE 4.1
Four cases resulted from the reduction process.

$\Sigma_q \setminus \hat{\Sigma}_{\hat{q}}$	Simple	Regular
simple	CASE 1	CASE 2
regular	CASE 3	CASE 4

Case 1: Both essential subsystems are simple, and therefore both Σ and $\hat{\Sigma}$ are QC. Moreover, it was defined previously that $\mathcal{N} = 0$ and $\hat{\mathcal{N}} = 0$, and hence condition (c) in Theorem 4.2 is trivially satisfied. Thus, in this case the system Σ_{of} is always QFM. Furthermore, it follows from Theorem 3.4 and Lemma 2.7(a) that both Σ and $\hat{\Sigma}$ are SC, and from Lemma 2.6, for all $s \in \boldsymbol{C}$, we have

$$\begin{aligned} \operatorname{rank}\begin{bmatrix} sI-A & B \\ C & D \end{bmatrix} &= n + \operatorname{rank}\begin{bmatrix} C & D \end{bmatrix}, \\ \operatorname{rank}\begin{bmatrix} sI-A & G \\ M & J \end{bmatrix} &= n + \operatorname{rank}\begin{bmatrix} G \\ J \end{bmatrix}. \end{aligned} \tag{4.7}$$

Cases 2 *and* 3: We consider *Case* 2 only; *Case* 3 can be treated similarly. Since the essential subsystem Σ_q is simple, Σ is QC and $\mathcal{N} = 0$. Hence conditions (a) and

(c) always hold, and Σ_{of} is QFM if and only if condition (b) holds. Since the essential subsystem $\hat{\Sigma}_{\hat{q}}$ is regular, $\hat{\Sigma}$ is QC if and only if $\hat{\Sigma}_{\hat{q}}$ satisfies the matching condition described in Lemma 2.4. In this case, the second equality in (4.7) is replaced with strict inequality ($<$).

Case 4: Both essential subsystems are regular. Conditions (a) and (b) can be checked by utilizing Lemma 2.4. Condition (c) can be verified by direct calculation. In this case, both equalities in (4.7) are replaced with strict inequalities. Finally, it is important to note that Σ_{of} *cannot* be QFM if Σ and $\hat{\Sigma}$ are both regular, since in this case $\mathcal{N} = \hat{\mathcal{N}} = I$ and thus condition (c) can never hold.

4.3. Proof of the main theorem. First, we define two matrix-valued mappings as follows:

$$\begin{aligned} F_\alpha(\Sigma, X) \quad &:= \quad \begin{bmatrix} B \\ D \end{bmatrix}^{\perp} \begin{bmatrix} (A+\alpha I)X + X(A+\alpha I)^t + GG^t & XC^t \\ CX & -I \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix}^{\perp t}, \\ S(X,Y,Z) \quad &:= \quad \begin{bmatrix} X & Z^t \\ Z & Y \end{bmatrix}, \end{aligned}$$

where Σ is the system described in (4.5).

The next lemma is essential for proving Theorem 4.2.

LEMMA 4.3. *Consider the uncertain system Σ_{of} described in* (4.4). *The following statements are equivalent.*

(i) *The uncertain system Σ_{of} is QFM.*

(ii) *For each $\alpha > 0$, there exist positive definite symmetric matrices X and Y, such that*

$$F_\alpha(\Sigma, X) < 0, \quad F_\alpha(\hat{\Sigma}, Y) < 0, \quad S(X,Y,I) > 0.$$

Proof. The equivalence follows from the quadratic stabilizability result [8] and the $\mathcal{H}_\infty$ synthesis result [3, 7]. □

In view of Lemma 3.2, next we present a condition which is similar to statement (ii) of Lemma 4.3, but is given in terms of $\hat{\Sigma}$ and a reduced order subsystem of Σ.

LEMMA 4.4. *Consider the uncertain system Σ_{of} in* (4.4) *and define Σ and $\hat{\Sigma}$ as in* (4.5) *and* (4.6). *Suppose that Σ is neither simple nor regular. Then statement* (ii) *of Lemma* 4.3 *holds if and only if the following condition is satisfied.*

(i) *For each $\alpha > 0$, there exist symmetric matrices X_1 and Y such that*

$$F_\alpha(\Sigma_1, X_1) < 0, \quad F_\alpha(\hat{\Sigma}, Y) < 0, \quad S(X_1, Y, N_0) > 0,$$

where Σ_1 is the reduced order subsystem of Σ defined in (3.1) *and N_0 is an orthonormal basis for the left null space of $B(I - D^+D)$.*

Proof. Suppose statement (ii) of Lemma 4.3 holds. From Lemma 3.2, we see that $F_\alpha(\Sigma, X) < 0$ implies that $F_\alpha(\Sigma_1, X_1) < 0$, where $X_1 := N_0^t X N_0$. Note that $S(X,Y,I) > 0$ is equivalent to

$$\begin{bmatrix} U_0^t & 0 \\ 0 & I \end{bmatrix} S(X,Y,I) \begin{bmatrix} U_0 & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} X_{nn} & X_{nr} & N_0^t \\ X_{nr}^t & X_{rr} & R_0^t \\ N_0 & R_0 & Y \end{bmatrix} > 0,$$

where U_0 and other matrices are defined in the proof of Lemma 3.2. This implies $S(X_1, Y, N_0) > 0$ and thus the necessity is proved.

Conversely, suppose statement (i) of Lemma 4.4 holds. With $X_{nn} := X_1$, there exists X_{nr} satisfying (3.7) for some $\tilde{K}$ (its existence is guaranteed by $F_\alpha(\Sigma_1, X_1) < 0$). Then X defined by

$$X := U_0 \begin{bmatrix} X_{nn} & X_{nr} \\ X_{nr}^t & X_{rr} \end{bmatrix} U_0^t$$

satisfies $F_\alpha(\Sigma, X) < 0$ for any choice of X_{rr}. Moreover, by choosing $X_{rr} := \gamma I$ with sufficiently large $\gamma > 0$, condition $S(X, Y, I) > 0$ can be met. Thus statement (ii) of Lemma 4.3 holds. □

Applying Lemma 4.4 recursively, a necessary and sufficient condition for the uncertain system Σ_{of} to be QFM can be given in terms of the essential subsystems of Σ and $\hat{\Sigma}$.

LEMMA 4.5. *Consider the uncertain system Σ_{of} and its auxiliary systems Σ and $\hat{\Sigma}$ as given in* (4.5) *and* (4.6). *Let Σ_q and $\hat{\Sigma}_{\hat{q}}$ be the essential subsystems of Σ and $\hat{\Sigma}$, respectively. Then, the following statements are equivalent.*

(i) *The uncertain system Σ_{of} is QFM.*

(ii) *For each $\alpha > 0$, there exist $X_q^t = X_q > 0$ and $Y_{\hat{q}}^t = Y_{\hat{q}} > 0$ such that*

$$F_\alpha(\Sigma_q, X_q) < 0, \quad F_\alpha(\hat{\Sigma}_{\hat{q}}, Y_{\hat{q}}) < 0, \quad S(X_q, Y_{\hat{q}}, \hat{N}^t N) > 0\ ,$$

where N is defined in (3.2) *for Σ and $\hat{N}$ is similarly defined for $\hat{\Sigma}$.*

Proof. Recursively applying Lemma 4.4, we see that statement (i) holds if and only if, for each $\alpha > 0$, there exist $X_q^t = X_q > 0$ and $Y^t = Y > 0$ such that

$$F_\alpha(\Sigma_q, X_q) < 0, \quad F_\alpha(\hat{\Sigma}, Y) < 0, \quad S(X_q, Y, N) > 0.$$

The same argument as described in Lemma 4.4 can be applied for the dual part $\hat{\Sigma}$, leading to the condition in statement (ii). □

Now, we can prove Theorem 4.2.

Proof. We consider the conditions given in Lemma 4.5 for the four cases defined in Table 4.1.

Case 1: Since Σ_q and $\hat{\Sigma}_{\hat{q}}$ are both simple, then it follows from Lemma 2.4(a) that $F_\alpha(\Sigma_q, X_q) < 0$ and $F_\alpha(\hat{\Sigma}_{\hat{q}}, Y_{\hat{q}}) < 0$ are trivially satisfied for any positive definite symmetric X_q and $Y_{\hat{q}}$. Choosing $X_q := \gamma I$ and $Y_{\hat{q}} := \gamma I$ with sufficiently large $\gamma > 0$, the third condition in statement (ii) of Lemma 4.5 can be satisfied. Thus Σ_{of} is QFM.

Cases 2 *and* 3: Consider *Case* 2; *Case* 3 can be shown similarly. Since Σ_q is simple, $F_\alpha(\Sigma_q, X_q) < 0$ holds for any choice of $X_q^t = X_q > 0$. Choosing sufficiently large X_q, the condition $S(X_q, Y_{\hat{q}}, \hat{N}^t N) > 0$ can be met. Thus the system Σ_{of} is QFM if and only if, for each $\alpha > 0$, there exists $Y_{\hat{q}} > 0$ such that $F_\alpha(\hat{\Sigma}_{\hat{q}}, Y_{\hat{q}}) < 0$. This is exactly the condition for $\hat{\Sigma}$ to be QC.

Case 4: Since both Σ_q and $\hat{\Sigma}_{\hat{q}}$ are regular, we have $\hat{\mathcal{N}}^t \mathcal{N} = \hat{N}^t N$. Moreover, it follows from Lemma 5.1 in the appendix that the solutions $X_q(\alpha)$ and $Y_{\hat{q}}(\alpha)$ to $F_\alpha(\Sigma_q, X_q) < 0$ and $F_\alpha(\hat{\Sigma}_{\hat{q}}, Y_{\hat{q}}) < 0$ approach zero when α approaches infinity. Hence, $S(X_q, Y_{\hat{q}}, \hat{N}^t N) > 0$ holds for sufficiently large $\alpha > 0$ only if $\hat{N}^t N = 0$. Conversely, if this is true, condition $S(X_q, Y_{\hat{q}}, 0) > 0$ holds for any positive definite X_q and $Y_{\hat{q}}$, and hence we are left with requirements for Σ_q and $\hat{\Sigma}_{\hat{q}}$ to be QC. □

5. Conclusion. In this paper, we have given complete characterizations of quadratic controllability and strong controllability for a class of norm bounded uncertain linear systems in terms of the essential subsystem. Although the problem of quadratic controllability was already discussed in [18], our approach presented here is straightforward and aimed to include the notion of strong controllability and the case with dynamic output feedback, which are entirely new. The equivalence from the view point of quadratic controllability and strong controllability for the original system and its essential subsystem was derived. A closely related output feedback property was also discussed. Specifically, the necessary and sufficient condition was presented for the existence of a dynamic output feedback controller that quadratically stabilizes the closed-loop system with an arbitrary rate of convergence. However, the geometrical and physical implication of the orthogonal condition given in Theorem 4.2 still needs further investigation.

Appendix: A useful lemma. The following algebraic result is instrumental for proving the main theorems of the paper. A similar result was first used in [24].

LEMMA 5.1. *Let matrices A, $B = B^t$, and $C = C^t \geq 0$ be given. Suppose, for each $\alpha > 0$, there exists $X^t(\alpha) = X(\alpha) > 0$ such that*

$$(A + \alpha I)X(\alpha) + X(\alpha)(A + \alpha I)^t + X(\alpha)CX(\alpha) - B < 0. \tag{5.1}$$

Then,

$$\lim_{\alpha \to \infty} X(\alpha) = 0 \ \textit{ and } \ B \geq 0\,.$$

Proof. It follows readily from (5.1) that, since $C = C^t \geq 0$, we have

$$(A + \alpha I)X(\alpha) + X(\alpha)(A + \alpha I)^t - B < 0\,. \tag{5.2}$$

Suppose $\alpha^* > 0$ is chosen such that $-(A + \alpha I)$ is Hurwitz for all $\alpha \geq \alpha^*$. Then, for each $\alpha \geq \alpha^*$, there exists a unique symmetric solution $Y(\alpha)$ to the following Lyapunov equation:

$$-(A + \alpha I)Y(\alpha) - Y(\alpha)(A + \alpha I)^t + B = 0\,. \tag{5.3}$$

Substituting (5.3) into (5.2), we obtain

$$-(A + \alpha I)[Y(\alpha) - X(\alpha)] - [Y(\alpha) - X(\alpha)](A + \alpha I)^t < 0$$

for all $\alpha \geq \alpha^*$. Since the matrix $-(A + \alpha I)$ is Hurwitz for all $\alpha \geq \alpha^*$, it follows from the Lyapunov stability theory that

$$Y(\alpha) - X(\alpha) > 0 \ \forall \, \alpha \geq \alpha^*\,, \tag{5.4}$$

and this also implies that $Y(\alpha) > 0$ for all $\alpha \geq \alpha^*$. Moreover, we note that $Y(\alpha)$ is monotonically decreasing as α increases. This can be shown by utilizing the fact that $Y(\alpha)$ is analytic in the interval $[\alpha^*, \infty)$ and by taking the derivative of (5.3) with respect to α to obtain

$$-(A + \alpha I)Y^{'}(\alpha) - Y^{'}(\alpha)(A + \alpha I)^t - 2Y(\alpha) = 0\,. \tag{5.5}$$

Since $Y(\alpha) > 0$, (5.5) implies that $Y^{'}(\alpha) := \frac{dY(\alpha)}{d\alpha} < 0$ for $\alpha \geq \alpha^*$. This shows that $Y(\alpha)$ is a monotonically decreasing function; hence $\lim_{\alpha \to \infty} Y(\alpha)$ exists. Moreover, if we rewrite (5.3) as

$$-2\alpha Y(\alpha) = AY(\alpha) + Y(\alpha)A^t - B\,, \tag{5.6}$$

then we observe that since the right-hand side of the above equation has a limit, so must the term on the left-hand side. That is,

$$\Pi_\infty := \lim_{\alpha \to \infty} \alpha Y(\alpha)$$

exists, and this in turn implies that

$$\lim_{\alpha \to \infty} Y(\alpha) = 0 \quad \text{and} \quad \Pi_\infty \geq 0\,. \tag{5.7}$$

Now, it can be readily deduced from (5.4) that

$$\lim_{\alpha \to \infty} X(\alpha) = 0.$$

Furthermore, by taking the limit on both sides of (5.6) and utilizing (5.7), we obtain that $-2\Pi_\infty = -B$, which implies $B \geq 0$. This completes the proof. □

Acknowledgments. Tetsuya Iwasaki gratefully acknowledges helpful discussions with T. Mita and X. Xin at Tokyo Institute of Technology.

REFERENCES

[1] B. R. Barmish, *Necessary and sufficient conditions for quadratic stabilizability of an uncertain system*, J. Optim. Theory Appl., 46 (1985), pp. 399–408.

[2] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[3] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to $\mathcal{H}_\infty$ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[4] K. Q. Gu, Y. H. Chen, M. A. Zohdy, and N. K. Loh, *Quadratic stabilizability of uncertain systems: A two level optimization setup*, Automatica, 27 (1991), pp. 161–165.

[5] M. L. J. Hautus, *Strong detectability and observers*, Linear Algebra Appl., 50 (1983), pp. 353–368.

[6] C. V. Hollot, *Bounded invariant Lyapunov functions: A means for enlarging the class of stabilizing uncertain systems*, Internat. J. Control, 46 (1987), pp. 161–184.

[7] T. Iwasaki and R. E. Skelton, *All controllers for the general H_∞ control problem: LMI existence conditions and state space formulas*, Automatica, 30 (1994), pp. 1307–1317.

[8] P. P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and H_∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.

[9] M. Krstic, I. Kanellakopoulos, and P. Kokotovic, *Nonlinear and Adaptive Control Design*, John Wiley & Sons, New York, 1995.

[10] B. P. Molinari, *Structural invariants of linear multivariable systems*, Internat. J. Control, 28 (1978), pp. 493–510.

[11] A. S. Morse, *Structural invariants of linear multivariable systems*, SIAM J. Control, 11 (1973), pp. 446–465.

[12] I. R. Petersen, *Structural stabilization of uncertain systems: Necessity of the matching condition*, SIAM J. Control Optim., 23 (1985), pp. 286–296.

[13] I. R. Petersen, *Notions of stabilizability and controllability for a class of uncertain linear systems*, Internat. J. Control, 46 (1987), pp. 409–422.

[14] I. R. Petersen, *A stabilization algorithm for a class of uncertain linear systems*, Systems Control Lett., 8 (1987), pp. 351–357.

[15] I. R. Petersen, *Stabilization of an uncertain linear system in which uncertain parameters enter into the input matrix*, SIAM J. Control Optim., 26 (1988), pp. 1257–1264.

[16] I. R. Petersen, *Quadratic stabilizability of uncertain linear systems containing both constant and time-varying uncertain parameters*, J. Optim. Theory Appl., 57 (1988), pp. 439–461.

[17] M. A. Rotea and P. P. Khargonekar, *Stabilization of uncertain systems with norm bounded uncertainty—A control Lyapunov function approach*, SIAM J. Control Optim., 27 (1989), pp. 1462–1476.

[18] M. A. Rotea, M. Corless, S.M. Swei, *Necessary and sufficient conditions for quadratic controllability of a class of uncertain systems*, Systems Control Lett., 26 (1995), pp. 195–201.

[19] W. E. Schmitendorf, *Designing stabilizing controllers for uncertain systems using the Riccati equation approach*, IEEE Trans. Automat. Control, 33 (1988), pp. 376–379.

[20] J. M. Schumacher, *On the structure of strongly controllable systems*, Internat. J. Control, 38 (1983), pp. 525–545.

[21] L. M. Silverman and H. J. Payne, *Input-output structure of linear systems with application to the decoupling problem*, SIAM J. Control, 9 (1971), pp. 199–233.

[22] A. A. Stoorvogel and H. L. Trentelman, *The quadratic matrix inequality in singular $\mathcal{H}_\infty$ control with state feedback*, SIAM J. Control Optim., 28 (1990), pp. 1190–1208.

[23] A. A. Stoorvogel, *The singular H_∞ control problem with dynamic measurement feedback*, SIAM J. Control Optim., 29 (1991), pp. 160–184.

[24] S. M. Swei and M. Corless, *On the necessity of the matching condition in robust stabilization*, in Proceedings of the 30th Conference on Decision & Control, Brighton, U.K., 1991, pp. 2611–2614.

[25] S. M. Swei, M. Rotea, and M. Corless, *System order reduction in robust stabilization problems*, Internat. J. Control, 60 (1994), pp. 223–241.

[26] Jan C. Willems, *Almost invariant subspaces: An approach to high gain feedback design.* I*: Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26 (1981), pp. 235–252.

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

[28] K. Zhou and P. P. Khargonekar, *Robust stabilization of linear systems with norm-bounded time-varying uncertainty*, Systems Control Lett., 10 (1988), pp. 17–20.

SIAM J. Control Optim.
Vol. 39, No. 5, pp. 1391–1407

OPTIMAL CONTROL PROBLEMS WITH MIXED CONTROL-STATE CONSTRAINTS*

N. ARADA† AND J.-P. RAYMOND†

Abstract. We consider control problems governed by semilinear parabolic equations in the presence of pointwise mixed control-state constraints. We obtain optimality conditions with finitely additive measures as multipliers associated to the mixed constraints. We study the regularity of the multipliers for different problems. In particular, when a monotonicity or a separation condition is satisfied, we prove that multipliers are bounded measurable functions.

AMS subject classifications. 49K20, 93C20

Key words. optimal control, mixed control-state constraints, pointwise constraints, regular multipliers

PII. S0363012999357926

1. Introduction.

This paper deals with control problems governed by parabolic equations in the presence of pointwise constraints on the control and the state variables. We are interested in optimality conditions for these problems, and in regularity properties of Lagrange multipliers associated with mixed control-state constraints. For optimization problems in Banach spaces the regularity of the multipliers is, in general, deduced from the following heuristic statement.

> (HS) The Lagrange multiplier associated with a constraint belongs to the dual space of the space in which the constraint set has a nonempty interior (or a finite codimension).

For bounded controls, pointwise mixed control-state constraints are well posed in an L^∞-space. Thus, according to (HS), the corresponding Lagrange multiplier must belong to a dual space of the form $(L^\infty)'$.

In some situations this heuristic statement does not give the best regularity result for the Lagrange multipliers. Let us give an example. The multiplier corresponding to pointwise state constraints is, in general, a bounded Radon measure [7], [10], [15], [12]. Indeed, pointwise state constraints are well posed in the space of continuous functions. For problems considered in [7], [15], the multiplier associated with a state constraint of the form $y \geq 0$ is a Radon measure. If the previous state constraint is slightly perturbed and is replaced by the control-state constraint $y + \varepsilon v \geq 0$ (with $\varepsilon > 0$ and $v \geq 0$), then the multiplier may be a bounded measurable function (see Corollary 5.6). This is clearly a better regularity result than the one derived from (HS).

This kind of result is known since 1962 for the control of ordinary differential equations, and is stated in [14]. The case of ordinary differential equations is very specific since it is often assumed that optimal controls are piecewise continuous [14], [11]. Even under the assumption of piecewise continuous controls, the "informal" Theorem 4.1 in [11] has not, to the authors' knowledge, been proved fully in the literature (see [11, p. 185]).

*Received by the editors June 18, 1999; accepted for publication (in revised from) June 17, 2000; published electronically December 28, 2000.

http://www.siam.org/journals/sicon/39-5/35792.html

†UMR-CNRS 5640, UFR MIG, Université Paul Sabatier, 31062 Toulouse Cedex 4, France (arada@mip.ups-tlse.fr, raymond@mip.ups-tlse.fr).

The situation is more complicated for partial differential equations. Indeed, in this case, it is not realistic to assume that optimal controls are piecewise continuous. Very recently, using a duality method, Bergounioux and Tröltzsch [5] have proved the existence of regular Lagrange multipliers for a linear control problem with constraints of bottleneck type. An extension to the case of semilinear equations is obtained in [6]. The case of integral control-state constraints is studied in [8].

To our knowledge, there are no general optimality conditions for control problems governed by partial differential equations in the presence of pointwise mixed control-state constraints. In this paper we first prove optimality conditions for control problems with constraints of the form

$$g(y,v) \in \mathcal{C} \subset (L^\infty)^\ell \qquad \text{(mixed control-state constraints)}$$

(in this setting $g = (g_1, \dots, g_\ell)$ is a vector valued function). In section 3, we do not make any particular assumption on mixed state-control constraints, and the corresponding multiplier belongs to a space of the type $((L^\infty)^\ell)'$ (Theorem 3.1). In sections 4 and 5, we prove the existence of regular Lagrange multipliers in the following cases:

- $\ell = 1$ and $(g'_v(y,v))^{-1}$ belongs to L^∞ (Theorem 4.1),
- mixed control-state constraints of the form

$$g_i(y,v) \le 0 \quad \text{for } i = 1, \dots, \ell \qquad \text{(Theorems 5.2, 5.5, 5.7).}$$

Let us denote by $\bar{\zeta}_i$ the multiplier of the constraint $g_i(y,v) \le 0$ corresponding to a solution $(\bar{y},\bar{v})$ of problem $(\mathcal{P})$. (The control problem $(\mathcal{P})$ is defined in section 1.1.) With optimality conditions and the Hahn–Banach extension theorem, we first establish that the sum $\Sigma_{i=1}^{\ell} g'_{iv}(\bar{y},\bar{v})\bar{\zeta}_i$ has the same regularity as the adjoint state.

In section 5.1 we study the case when $\ell = 2$ and when (g_1, g_2) satisfies a separation condition of the form $g_1(\bar{y},\bar{v}) + g_2(\bar{y},\bar{v}) \le -\varepsilon < 0$. We are able to prove that the supports of $\bar{\zeta}_1$ and $\bar{\zeta}_2$ are disjoints. We deduce that each multiplier has the same regularity as the adjoint state, and we use a bootstrap argument to obtain the best regularity result (Theorem 5.2).

In section 5.2 we suppose that g satisfies a monotonicity condition of the form $g'_{iv}(y,v) \le 0$ for $i = 1, \dots, \ell$. We prove that, for $i = 1, \dots, \ell$, the additive measures $g'_{iv}(\bar{y},\bar{v})\bar{\zeta}_i$ are nonnegative. With a decomposition theorem for nonnegative additive measures, we prove that each term $g'_{iv}(\bar{y},\bar{v})\bar{\zeta}_i$ has the same regularity as the adjoint state, and we can conclude with a bootstrap argument (Theorem 5.5).

In section 5.3 we study a problem in which the separation and monotonicity conditions are coupled (Theorem 5.7). These results are applied to examples for which the regularity of multipliers cannot be deduced from results known in the literature.

The separation condition seems to be new (see Remark 5.1). The monotonicity condition is similar to regularity conditions stated for the control of ordinary differential equations (condition (b) in [4]).

For clarity, we consider only the case of a boundary control, but our method can be extended to problems with distributed and boundary controls. Our method is general and may be extended to other problems.

1.1. Setting of the control problem.

Consider the semilinear parabolic equation

$$(1.1)\ \frac{\partial y}{\partial t} + Ay + \Phi(\cdot,y) = 0 \ \text{ in } Q, \quad \frac{\partial y}{\partial n_A} + \Psi(\cdot,y,v) = 0 \ \text{ on } \Sigma, \quad y(0) = y_o \ \text{ in } \Omega,$$

where Ω is a bounded domain in $\mathbb{R}^N$, $Q = \Omega \times]0, T[$, $\Sigma = \Gamma \times]0, T[$, Γ is the boundary of Ω, $T > 0$, $v \in L^\infty(\Sigma)$ is a boundary control, A is a second order elliptic operator, Φ and Ψ are Carathéodory functions (assumptions are specified in section 1.2), and y_o is a fixed function in $C(\overline{\Omega})$. Constraints of the form

$$g(\cdot, y(\cdot), v(\cdot)) \in \mathcal{C} \tag{1.2}$$

are imposed on the pair (y, v), where g is a continuous mapping from $\overline{\Sigma} \times \mathbb{R}^2$ into $\mathbb{R}^\ell$, and $\mathcal{C}$ is a closed convex subset of $(L^\infty(\Sigma))^\ell$ with a nonempty interior in $(L^\infty(\Sigma))^\ell$. The paper is concerned with the control problem

$$(\mathcal{P}) \qquad \inf\{J(y, v) \mid \quad (y, v) \in C(\overline{Q}) \times L^\infty(\Sigma), \quad (y, v) \text{ satisfies (1.1) and (1.2)}\},$$

where the cost functional is given by

$$J(y, v) = \int_Q F(x, t, y)\, dx\, dt + \int_\Sigma G(s, t, y, v)\, ds\, dt + \int_\Omega L(x, y(T))\, dx.$$

Let us give a simple example for which results of the paper may be applied. Consider the state equation

$$\frac{\partial y}{\partial t} - \Delta y = 0 \text{ in } Q, \quad \frac{\partial y}{\partial n} + |y|^3 y = v \text{ on } \Sigma, \quad y(0) = y_o \text{ in } \Omega,$$

with $y_o \geq 0$, $v \geq 0$, and the functional $J(y, v) = \int_\Omega a(x)|y(x, T) - y_d(x)|^2 dx + \int_\Sigma y^4(s, t)\, dsdt$. The nonnegative function a plays the role of a weight, y_d is a desired profile of temperature, and the term $\int_\Sigma y^4(s, t)\, dsdt$ may be interpreted as a term penalizing a too-high temperature on the boundary. In the above radiation boundary condition, the control v corresponds to y^4_{ext}, where y_{ext} is the exterior temperature. In industrial processes it may be important that the difference $y^4_{ext} - y^4$ be bounded from above. Thus a constraint of the form $0 \leq v \leq y^4 + c$ on Σ, with $c > 0$, is meaningful. Setting $g_1(y, v) = v - |y|^3 y - c$, $g_2(y, v) = -v$, the above constraint is equivalent to $g(y, v) = (g_1(y, v), g_2(y, v)) \in \mathcal{C}$, where $\mathcal{C} = \{z \in (L^\infty(\Sigma))^2 \mid z \leq 0\}$ ($z \leq 0$ must be understood componentwise). The existence of solutions for the corresponding problem $(\mathcal{P})$ can be obtained by standard arguments. In Corollary 5.3 we obtain optimality conditions for a class of problems including this example, with multipliers belonging to $L^\infty(\Sigma)$. In our knowledge, the existence of bounded measurable multipliers for this example cannot be deduced from known results in the literature. Examples of constraints for which we obtain the same results are given in section 5. Other examples of equations and functionals satisfying the assumptions of the paper are given in [15].

1.2. Assumptions and notation. We suppose that Ω is of class C^2 (the boundary Γ of Ω is an $(N-1)$-dimensional manifold of class C^2 such that Ω lies locally on one side of Γ). The operator A is of the form

$$Ay(x) = -\sum_{i,j=1}^N D_i(a_{ij}(x) D_j y(x))$$

(D_i denotes the partial derivative with respect to x_i), with coefficient a_{ij} belonging to $C^1(\overline{\Omega})$ and satisfying the conditions

$$a_{ij}(x) = a_{ji}(x) \quad \text{for all } \ i, j \in \{1, \ldots, N\}, \qquad m_0|\xi|^2 \leq \sum_{i,j=1}^N a_{ij}(x)\xi_i\xi_j$$

for every $\xi \in \mathbb{R}^N$ and every $x \in \overline{\Omega}$, with $m_0 > 0$. The conormal derivate of y with respect to A is denoted by $\frac{\partial y}{\partial n_A}$, that is

$$\frac{\partial y}{\partial n_A}(s,t) = \sum_{i,j} a_{ij}(s) D_j y(s,t) n_i(s),$$

where $n = (n_1, \ldots, n_N)$ is the unit normal to Γ outward Ω.

To prove the existence of multipliers in $(L^k(\Sigma))^\ell$ we need some regularity assumptions on the data. They are specified below. Some of them depend on k. Observe that $k = \infty$ is allowed (see (A3)–(A5) below). Throughout the text, the exponent k denotes a given fixed exponent belonging to $]1, \infty]$.

(A1) Φ is a Carathéodory function from $Q \times \mathbb{R}$ into $\mathbb{R}$. For almost every $(x,t) \in Q$, $\Phi(x,t,\cdot)$ is of class C^1. The following estimates hold:

$$|\Phi(x,t,0)| \le \eta(|y|), \qquad C_o \le \Phi'_y(x,t,y) \le \eta(|y|),$$

where $C_o \in \mathbb{R}$ and η is a nondecreasing function from $\mathbb{R}^+$ into $\mathbb{R}^+$.

(A2) Ψ is a Carathéodory function from $\Sigma \times \mathbb{R}^2$ into $\mathbb{R}$. For almost every $(s,t) \in \Sigma$, $\Psi(s,t,\cdot)$ is of class C^1, and it satisfies

$$\begin{aligned} &|\Psi(s,t,0,v)| + |\Psi'_v(s,t,y,v)| \le \eta(|y|)\ \eta(|v|), \\ &C_o \le \Psi'_y(s,t,y,v) \le \eta(|y|)\ \eta(|v|). \end{aligned}$$

(A3) F is a Carathéodory function from $Q \times \mathbb{R}$ into $\mathbb{R}$. For almost all $(x,t) \in Q$, $F(x,t,\cdot)$ is of class C^1. The following estimates hold:

$$|F(x,t,y)| \le F_1(x,t)\eta(|y|), \qquad |F'_y(x,t,y)| \le F_2(x,t)\eta(|y|),$$

where $F_1 \in L^1(Q)$, $F_2 \in L^{k_1}(Q)$, with $k_1 > \frac{(N+2)k}{N+1+2k}$ if $k < \infty$, and $k_1 > \frac{N}{2} + 1$ if $k = \infty$.

(A4) G is a Carathéodory function from $\Sigma \times \mathbb{R}^2$ into $\mathbb{R}$. For almost all $(s,t) \in \Sigma$, $G(s,t,\cdot)$ is of class C^1. The following estimates hold:

$$\begin{aligned} &|G(s,t,y,v)| \le G_1(s,t)\eta(|y|)\eta(|v|), \qquad |G'_v(s,t,y,v)| \le G_2(s,t)\eta(|y|)\eta(|v|), \\ &|G'_y(s,t,y,v)| \le G_3(s,t)\eta(|y|)\eta(|v|), \end{aligned}$$

where $G_1 \in L^1(\Sigma)$, $G_2 \in L^k(\Sigma)$, with $1 < k \le \infty$, $G_3 \in L^{k_2}(\Sigma)$ with $k_2 > \frac{(N+1)k}{N+1+k}$ if $k < \infty$, and $k_2 > N+1$ if $k = \infty$.

(A5) L is a Carathéodory function from $\Omega \times \mathbb{R}$ into $\mathbb{R}$. For almost all $x \in \Omega$, $L(x,\cdot)$ is of class C^1, and

$$|L(x,y)| \le L_1(x)\eta(|y|), \qquad |L'_y(x,y)| \le L_2(x)\eta(|y|),$$

where $L_1 \in L^1(\Omega)$, $L_2 \in L^{k_3}(\Omega)$, with $k_3 > \frac{Nk}{N+1}$ if $k < \infty$, and $k_3 = \infty$ if $k = \infty$.

(A6) g is a function from $\overline{\Sigma} \times \mathbb{R}^2$ into $\mathbb{R}^\ell$. For every $(s,t) \in \overline{\Sigma}$, $g(s,t,\cdot)$ is of class C^1, and for every $y \in \mathbb{R}$, $g(\cdot,y)$ is bounded measurable on Σ. The functions g, g'_y, and g'_v are bounded on compact subsets of $\overline{\Sigma} \times \mathbb{R}^2$.

For simplicity, we often write $g(y,v)$ in place of $g(s,t,y,v)$. We set $\overline{\Omega}_0 = \overline{\Omega} \times \{0\}$, $\overline{\Omega}_T = \overline{\Omega} \times \{T\}$, $\Gamma_T = \Gamma \times \{T\}$. For every $1 \le r \le \infty$, the usual norms in the spaces $L^r(\Omega)$, $L^r(Q)$, $L^r(\Sigma)$ will be denoted by $||\cdot||_{r,\Omega}$, $||\cdot||_{r,Q}$, $||\cdot||_{r,\Sigma}$. The Hilbert space

$W(0,T;H^1(\Omega),(H^1(\Omega))') = \{y \in L^2(0,T;H^1(\Omega)) \mid \frac{dy}{dt} \in L^2(0,T;(H^1(\Omega))')\}$ will be denoted by $W(0,T)$. Throughout what follows, we denote by C a generic constant. If $\mathcal{O}$ is a locally compact subset of $\mathbb{R}^{N+1}$, $C_b(\mathcal{O})$ denotes the space of bounded continuous functions on $\mathcal{O}$, and $C_0(\mathcal{O})$ denotes the space of continuous functions vanishing at infinity. The dual space of $C_0(\mathcal{O})$ will be denoted by $\mathcal{M}_b(\mathcal{O})$ (the space of bounded Radon measures on $\mathcal{O}$). If μ belongs to $\mathcal{M}_b(\mathcal{O})$ and y belongs to $C_b(\mathcal{O})$, we set $\langle \mu, y\rangle_{b,\mathcal{O}} = \int_{\mathcal{O}} y\, d\mu$. For $\sigma > 1$, $k \wedge \sigma$ stands for $\min(k,\sigma)$. For simplicity, $\langle \cdot,\cdot\rangle_{*,\Sigma}$ stands for the duality pairing between the spaces $((L^\infty(\Sigma))^\ell)'$ and $(L^\infty(\Sigma))^\ell$.

2. State and adjoint equations. We begin this section by recalling some existence, uniqueness, and regularity results concerning the state equation (1.1).

DEFINITION 2.1. *A function $y \in L^2(0,T;H^1(\Omega)) \cap C([0,T];L^2(\Omega))$ is a weak solution of* (1.1) *if and only if $\Phi(\cdot,y(\cdot)) \in L^1(Q)$, $\Psi(\cdot,y(\cdot),v(\cdot)) \in L^1(\Sigma)$, and*

$$\int_Q \left(-y\frac{\partial z}{\partial t} + \sum_{i,j=1}^N a_{ij} D_i y D_j z + \Phi(\cdot,y)z \right) dx\,dt - \int_\Omega y_o z(0)\,dx$$
$$= \;-\int_\Sigma \Psi(\cdot,y,v)z\,ds\,dt \qquad \text{for all } z \in C^1(\overline{Q}) \text{ such that } z(T)=0.$$

THEOREM 2.2 (see [16, Theorem 3.1]). *Let us suppose that* (A1)–(A2) *are satisfied. Equation* (1.1) *admits a unique weak solution $y_v \in W(0,T)\cap C(\overline{Q})$ satisfying*

$$\|y_v\|_{C(\overline{Q})} \le C(\|v\|_{\infty,\Sigma} + \|y_o\|_{C(\overline{\Omega})} + 1).$$

Now, let us recall some existence, uniqueness, and regularity results for the adjoint equation. Let (a,b) be in $L^\infty(Q)\times L^\infty(\Sigma)$, and consider the terminal boundary value problem

$$(2.1)\quad \frac{\partial p}{\partial t} + Ap + ap = \mu_Q \quad \text{in } Q, \qquad \frac{\partial p}{\partial n_A} + bp = \mu_\Sigma \quad \text{on } \Sigma, \qquad p(T) = \mu_{\overline{\Omega}_T} \quad \text{on } \overline{\Omega},$$

where $\mu = \mu_Q + \mu_\Sigma + \mu_{\overline{\Omega}_T}$ is a bounded Radon measure on $\overline{Q}\setminus\overline{\Omega}_0$, μ_Q is the restriction of μ to Q, μ_Σ is the restriction of μ to Σ, and $\mu_{\overline{\Omega}_T}$ is the restriction of μ to $\overline{\Omega}_T$.

DEFINITION 2.3. *A function $p \in L^1(0,T;W^{1,1}(\Omega))$ is a weak solution of* (2.1) *if and only if*

$$\int_Q \left(p\frac{\partial z}{\partial t} + \sum_{i,j=1}^N a_{ij} D_j p D_i z + azp \right) dxdt + \int_\Sigma bpz\,dsdt = \langle \mu, z\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0}$$

for all $z \in C^1(\overline{Q})$ satisfying $z(0)=0$ on $\overline{\Omega}$.

We recall an existence theorem for parabolic equations with measures as data.

THEOREM 2.4 (see [15, Theorem 4.1]). *Let (a,b) be in $L^\infty(Q)\times L^\infty(\Sigma)$, and let μ be in $\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)$. Equation* (2.1) *admits a unique solution p in $L^1(0,T;W^{1,1}(\Omega))$. For every (δ,d,β) satisfying $\delta \ge 1$, $d \ge 1$, $\beta \ge 0$, and $\frac{N}{2d} + \frac{1}{\delta} > \frac{N}{2} + \frac{\beta}{2}$, p belongs to $L^\delta(0,T;W^{\beta,d}(\Omega))$ and*

$$||p||_{L^\delta(0,T;W^{\beta,d}(\Omega))} \le C||\mu||_{\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)},$$

where $C \equiv C(\Omega, T, \delta, d, M)$ is a positive constant independent of a and b, and M is an upper bound for $||a||_{\infty,Q} + ||b||_{\infty,\Sigma}$. In particular, the trace of p on Σ satisfies

$$||p_{|\Sigma}||_{L^\sigma(\Sigma)} \leq C||\mu||_{\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)} \qquad \text{for every } 1 \leq \sigma < \frac{N+1}{N}.$$

Moreover, there exists a function in $L^1(\Omega)$, denoted by $p(0)$, such that

$$\int_Q \left(\frac{\partial z}{\partial t} + Az + az\right) p\,dx\,dt + \int_\Sigma \left(\frac{\partial z}{\partial n_A} + bz\right) p\,ds\,dt + \int_\Omega z(0)p(0)\,dx$$
$$= \langle \mu, z\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} \qquad \text{for all } z \in \mathcal{Y},$$

where $\mathcal{Y} = \{y \in W(0,T) \mid \frac{\partial y}{\partial t} + Ay \in L^q(Q),\ \frac{\partial y}{\partial n_A} \in L^\infty(\Sigma),\ y(0) \in C(\overline{\Omega})\}$.

Proposition 2.5. *Set $\Lambda(\mu_Q, \mu_\Sigma, \mu_{\overline{\Omega}_T}) = p_{|\Sigma}$, where p is the solution of* (2.1) *corresponding to $(\mu_Q, \mu_\Sigma, \mu_{\overline{\Omega}_T})$. The mapping $\mu_Q \mapsto \Lambda(\mu_Q, 0, 0)$ is continuous from $L^{\beta_1}(Q)$ into $L^\beta(\Sigma)$, with $\beta_1 > \frac{(N+2)\beta}{N+1+2\beta}$ if $1 \leq \beta < \infty$, and $\beta_1 > \frac{N}{2} + 1$ if $\beta = \infty$. The mapping $\mu_\Sigma \mapsto \Lambda(0, \mu_\Sigma, 0)$ is continuous from $L^{\beta_2}(\Sigma)$ into $L^\beta(\Sigma)$, with $\beta_2 > \frac{(N+1)\beta}{N+1+\beta}$ if $1 \leq \beta < \infty$, and $\beta_2 > N+1$ if $\beta = \infty$.*

The mapping $\mu_{\overline{\Omega}_T} \mapsto \Lambda(0, 0, \mu_{\overline{\Omega}_T})$ is continuous from $L^{\beta_3}(\Omega)$ into $L^\beta(\Sigma)$, with $\beta_3 > \frac{N\beta}{N+1}$ if $1 \leq \beta < \infty$, and $\beta_3 = \infty$ if $\beta = \infty$.

In particular, $p_{|\Sigma}$ belongs to $L^k(\Sigma)$ if $(\mu_Q, \mu_\Sigma, \mu_{\overline{\Omega}_T})$ belongs to $L^{k_1}(Q) \times L^{k_2}(\Sigma) \times L^{k_3}(\Omega)$, where k, k_1, k_2, and k_3 are the exponents in assumptions (A3)–(A5).

Proof. The proof may be performed by using estimates on the analytic semigroup as in [17], Propositions 3.1 and 3.2. □

3. Main results.

Define the Hamiltonian function $\mathcal{H} : \Sigma \times \mathbb{R}^4 \longrightarrow \mathbb{R}$ by

$$\mathcal{H}(s,t,y,v,p,\alpha) = \alpha\ G(s,t,y,v) + p\ \Psi(s,t,y,v).$$

We shall say that $(\bar{y}, \bar{v})$ is regular if there exists $\widehat{v} \in L^\infty(\Sigma)$ such that

$$g(\bar{y},\bar{v}) + g'_y(\bar{y},\bar{v})(z_{\widehat{v}} - z_{\bar{v}}) + g'_v(\bar{y},\bar{v})(\widehat{v} - \bar{v}) \in \text{int } \mathcal{C}, \tag{3.1}$$

where z_v (with $v = \hat{v}$ or $v = \bar{v}$) is the solution of

$$\begin{cases} \dfrac{\partial z}{\partial t} + Az + \Phi'_y(\cdot,\bar{y})z = 0 & \text{in Q}, \\ \dfrac{\partial z}{\partial n_A} + \Psi'_y(\cdot,\bar{y},\bar{v})z = -\Psi'_v(\cdot,\bar{y},\bar{v})\ v & \text{on } \Sigma, \qquad z(0) = 0 \text{ in } \Omega. \end{cases} \tag{3.2}$$

In (3.1) "int" denotes the interior for the topology of $(L^\infty(\Sigma))^\ell$. The qualification condition (3.1) is of Mangasarian–Fromowitz type. It appears in [19, (3.6), p. 20].

Theorem 3.1. *Suppose that* (A1)–(A6) *are fulfilled. If $(\bar{y}, \bar{v})$ is a solution of $(\mathcal{P})$, then there exist $\bar{\alpha} \in [0,1]$, $\bar{\zeta} \in ((L^\infty(\Sigma))^\ell)'$, and $\bar{p} \in L^1(0,T;W^{1,1}(\Omega))$, such that the following conditions hold.*

- *Nontriviality condition* :

$$(\bar{\alpha}, \bar{\zeta}) \neq 0. \tag{3.3}$$

- *Complementarity condition*:

$$\langle \bar{\zeta}, z - g(\bar{y},\bar{v})\rangle_{*,\Sigma} \leq 0 \qquad \text{for all } z \in \mathcal{C}. \tag{3.4}$$

- *Adjoint equation*:

$$(3.5)\qquad \begin{cases} -\dfrac{\partial \bar p}{\partial t} + A\bar p + \Phi'_y(\cdot,\bar y)\bar p = -\bar\alpha F'_y(\cdot,\bar y) & \textit{in } Q,\\[2mm] \dfrac{\partial \bar p}{\partial n_A} + \Psi'_y(\cdot,\bar y,\bar v)\bar p = -\bar\alpha\ G'_y(\cdot,\bar y,\bar v) - [g'_y(\bar y,\bar v)^*\bar\zeta]_{|\Sigma} & \textit{on } \Sigma,\\[2mm] \bar p(T) = -\bar\alpha L'_y(\cdot,\bar y(T)) - [g'_y(\bar y,\bar v)^*\bar\zeta]_{|\Gamma\times\{T\}} & \textit{on } \overline{\Omega}, \end{cases}$$

where $[g'_y(\bar y,\bar v)^*\bar\zeta]_{|\Sigma}$ *is the restriction of* $g'_y(\bar y,\bar v)^*\bar\zeta$ *to* Σ*,* $[g'_y(\bar y,\bar v)^*\bar\zeta]_{|\Gamma\times\{T\}}$ *is the restriction of* $g'_y(\bar y,\bar v)^*\bar\zeta$ *to* $\Gamma\times\{T\}$*, and* $g'_y(\bar y,\bar v)^*\bar\zeta$ *is the bounded Radon measure on* $\Sigma\cup(\Gamma\times\{T\})$ *defined by*

$$\langle g'_y(\bar y,\bar v)^*\bar\zeta, z\rangle_{b,\Sigma\cup\Gamma\times\{T\}} = \langle\bar\zeta, g'_y(\bar y,\bar v)z\rangle_{*,\Sigma} \qquad \textit{for all } z\in C_0(\Sigma\cup(\Gamma\times\{T\})).$$

- *Optimality condition for* $\bar v$:

$$(3.6)\quad \int_\Sigma \mathcal{H}'_v(s,t,\bar y,\bar v,\bar p,\bar\alpha)\chi\,ds\,dt + \langle\bar\zeta, g'_v(\bar y,\bar v)\ \chi\rangle_{*,\Sigma} = 0 \quad \textit{for all}\ \ \chi\in L^\infty(\Sigma).$$

Moreover, if there exists $\widehat v$ *satisfying* (3.1)*, then we can take* $\bar\alpha = 1$.

Remark 3.2. Here $g'_y(\bar y,\bar v)$ is a linear continuous operator from $C_0(\Sigma\cup(\Gamma\times\{T\}))$ into $(L^\infty(\Sigma))^\ell$, and $g'_y(\bar y,\bar v)^*$ denotes the corresponding adjoint operator.

Proof. Let us set

$$\begin{aligned} \mathcal{A} = \{(z,\lambda)\in(L^\infty(\Sigma))^\ell\times\mathbb{R} \mid\ & z = g(\bar y,\bar v) + g'_y(\bar y,\bar v)(z_v - z_{\bar v}) + g'_v(\bar y,\bar v)(v-\bar v),\\ & \lambda = J'_y(\bar y,\bar v)\ (z_v - z_{\bar v}) + J'_v(\bar y,\bar v)(v-\bar v) \ \text{ for some }\ v\in L^\infty(\Sigma)\},\\ \mathcal{B} = \text{int }\mathcal{C}\times&]-\infty,0[, \end{aligned}$$

where $z_{\bar v}$ (respectively, z_v) is the solution of (3.2) corresponding to $\bar v$ (respectively, v). The sets $\mathcal{A}$ and $\mathcal{B}$ are convex, and $\mathcal{B}$ is open. Let us prove that $\mathcal{A}\cap\mathcal{B}=\emptyset$. Argue by contradiction, and suppose that there exists $v_o\in L^\infty(\Sigma)$ such that

$$(3.7)\qquad g(\bar y,\bar v) + g'_y(\bar y,\bar v)(z_{v_o} - z_{\bar v}) + g'_v(\bar y,\bar v)(v_o-\bar v)\in\text{int }\mathcal{C},$$

$$(3.8)\qquad \lambda_o = J'_y(\bar y,\bar v)\ (z_{v_o} - z_{\bar v}) + J'_v(\bar y,\bar v)(v_o-\bar v) < 0.$$

Set $v_\rho = \bar v + \rho(v_o - \bar v)$. Let y_ρ be the solution of (1.1) corresponding to v_ρ, and set $g_\rho = g(\bar y,\bar v) + \frac{1}{\rho}(g(y_\rho,v_\rho) - g(\bar y,\bar v))$. Due to (3.7) and (3.8), there exists $0<\rho_o<1$ such that

$$g_\rho\in\text{int }\mathcal{C} \quad\text{and}\quad \frac{J(y_\rho,v_\rho)-J(\bar y,\bar v)}{\rho} < 0 \quad\text{for all } 0<\rho\le\rho_o<1.$$

Therefore, for every $0<\rho\le\rho_o<1$, we have

$$g(y_\rho,v_\rho) = \rho\ g_\rho + (1-\rho)\ g(\bar y,\bar v)\in\ \text{int }\mathcal{C}\ \text{ and }\ J(y_\rho,v_\rho) < J(\bar y,\bar v) = \inf(\mathcal{P}).$$

The pair (y_ρ,v_ρ) is admissible for $(\mathcal{P})$ and we have a contradiction. Thus, $\mathcal{A}\cap\mathcal{B}=\emptyset$. From a geometric version of the Hahn–Banach theorem (the Eidelheit theorem [18]), there exists $(\bar\alpha,\bar\zeta)\in\mathbb{R}\times((L^\infty(\Sigma))^\ell)'$, such that

$$(3.9)\qquad \bar\alpha\ \lambda_1 + \langle\bar\zeta,z_1\rangle_{*,\Sigma} > \bar\alpha\ \lambda_2 + \langle\bar\zeta,z_2\rangle_{*,\Sigma}\ \text{ for all } (z_1,\lambda_1,z_2,\lambda_2)\in\mathcal{A}\times\mathcal{B},$$

$$\bar\alpha\ \lambda_1 + \langle\bar\zeta, z_1\rangle_{*,\Sigma} \geq \bar\alpha\ \lambda_2 + \langle\bar\zeta, z_2\rangle_{*,\Sigma} \quad \text{for all } (z_1,\lambda_1,z_2,\lambda_2) \in \mathcal{A}\times\overline{\mathcal{B}}. \tag{3.10}$$

- Due to (3.9), $(\bar\alpha,\bar\zeta) \neq 0$. We easily see that $\bar\alpha$ is nonnegative. Indeed if $\bar\alpha < 0$, setting $z_1 = g(\bar y,\bar v)$, $\lambda_1 = 0$, and letting λ_2 tend to $-\infty$, we obtain a contradiction. Thus $\bar\alpha$ is nonnegative. For z fixed in $\mathcal{C}$, by setting $z_1 = g(\bar y,\bar v)$, $z_2 = z$, $\lambda_1 = \lambda_2 = 0$ in (3.10), we establish (3.4).
- Let $v \in L^\infty(\Sigma)$. By setting $z_1 = g(\bar y,\bar v) + g'_y(\bar y,\bar v)(z_v - z_{\bar v}) + g'_v(\bar y,\bar v)(v-\bar v)$, $\lambda_1 = J'_y(\bar y,\bar v)\ (z_v - z_{\bar v}) + J'_v(\bar y,\bar v)(v-\bar v)$, $z_2 = g(\bar y,\bar v)$, and $\lambda_2 = 0$ in (3.10), we obtain

$$\begin{aligned}&\bar\alpha J'_y(\bar y,\bar v)\ (z_v - z_{\bar v}) + \langle g'_y(\bar y,\bar v)^*\bar\zeta, z_v - z_{\bar v}\rangle_{b,\Sigma\cup\Gamma_T}\\ &+ \bar\alpha J'_v(\bar y,\bar v)(v-\bar v) + \langle\bar\zeta, g'_v(\bar y,\bar v)(v-\bar v)\rangle_{*,\Sigma} \geq 0 \quad \text{for all } v \in L^\infty(\Sigma).\end{aligned} \tag{3.11}$$

Since the above inequality is satisfied for all $v \in L^\infty(\Sigma)$, the inequality can be replaced by an equality. Let $\bar p$ be the weak solution of (3.5). With the Green formula of Theorem 2.4, we have

$$\begin{aligned}&\int_Q \bar\alpha\ F'_y(x,t,\bar y)\ (z_v - z_{\bar v})\,dx\,dt + \int_\Omega \bar\alpha\ L'_y(x,\bar y(T))\ (z_v - z_{\bar v})(T)\,dx\\ &+\int_\Sigma \bar\alpha\ G'_y(s,t,\bar y,\bar v)\ (z_v - z_{\bar v})\,ds\,dt + \langle g'_y(\bar y,\bar v)^*\bar\zeta, z_v - z_{\bar v}\rangle_{b,\Sigma\cup\Gamma_T}\\ &= \int_\Sigma \bar p\ \Psi'_v(s,t,\bar y,\bar v)\ (v-\bar v)\,ds\,dt \quad \text{for all } v\in L^\infty(\Sigma).\end{aligned}$$

This equality together with (3.11) gives (3.6).

- Finally, if there exists $\widehat v \in L^\infty(\Sigma)$ such that (3.1) is satisfied, then by setting $z_1 = z_2 = g(\bar y,\bar v) + g'_y(\bar y,\bar v)(z_{\widehat v} - z_{\bar v}) + g'_v(\bar y,\bar v)(\widehat v - \bar v)$ in (3.9), we prove that $\bar\alpha \neq 0$. □

4. Regularity of multipliers for purely mixed constraints. Throughout what follows, $(\bar y,\bar u)$ stands for an optimal solution to $(\mathcal{P})$. In the following two sections, we want to prove that the multiplier $\bar\zeta$ may be identified with $\bar\eta\,dsdt$, where $\bar\eta$ belongs to $(L^k(\Sigma))^\ell$. In this case, (3.3)–(3.6) are rewritten in the form

$$(\bar\alpha,\bar\eta)\neq 0, \qquad \int_\Sigma \bar\eta\ (z - g(\bar y,\bar v))\,ds\,dt \leq 0 \qquad \text{for all } z\in\mathcal{C}, \tag{4.1}$$

$$\begin{cases} -\dfrac{\partial\bar p}{\partial t} + A\bar p + \Phi'_y(\cdot,\bar y)\bar p = -\bar\alpha F'_y(\cdot,\bar y) & \text{in } Q,\\ \dfrac{\partial\bar p}{\partial n_A} + \Psi'_y(\cdot,\bar y,\bar v)\bar p = -\bar\alpha G'_y(\cdot,\bar y,\bar v) - g'_y(\bar y,\bar v)^*\bar\eta & \text{on } \Sigma,\\ \bar p(T) = -\bar\alpha L'_y(\cdot,\bar y(T)) & \text{on } \Omega,\end{cases} \tag{4.2}$$

$$\mathcal{H}'_v(s,t,\bar y(s,t),\bar v(s,t),\bar p(s,t),\bar\alpha) = -g'_v(\bar y,\bar v)^*\bar\eta(s,t) \qquad \text{for almost all } (s,t)\in\Sigma. \tag{4.3}$$

In this section, we consider the control problem $(\mathcal{P})$ when $\ell = 1$ and when the following regularity condition is satisfied.

(A7) $\ell = 1$ and the function $(g'_v(\bar y(\cdot),\bar v(\cdot)))^{-1}$ belongs to $L^\infty(\Sigma)$.

THEOREM 4.1. *Let $(\bar y,\bar v)$ be a solution to $(\mathcal{P})$. Suppose that* (A1)–(A7) *are fulfilled. There exist $\bar\alpha\in[0,1]$, $\bar\eta\in L^k(\Sigma)$ (k is the exponent in assumption* (A3)*),* *and* $\bar{p} \in L^1(0,T;W^{1,1}(\Omega))$*, such that* (4.1)–(4.3) *hold. Moreover, if* (3.1) *is satisfied, then we can take* $\bar{\alpha} = 1$.

Proof. Step 1. Due to Theorem 3.1, there exist $\bar{\alpha} \in [0,1]$, $\bar{\zeta} \in (L^\infty(\Sigma))'$, and $\bar{p} \in L^1(0,T;W^{1,1}(\Omega))$, such that (3.3)–(3.6) hold. From Proposition 2.5, we know that

$$\bar{p}_{|\Sigma} \in L^\sigma(\Sigma) \qquad \text{for all } \sigma < \frac{N+1}{N}. \tag{4.4}$$

Let σ be such that $\sigma < \frac{N+1}{N}$. From (4.4), with assumptions (A2) and (A4), we can easily see that $\mathcal{H}'_v(\cdot,\bar{y},\bar{v},\bar{p},\bar{\alpha})$ belongs to $L^{k\wedge\sigma}(\Sigma)$. Consider the continuous linear operators $\mathcal{S} : L^{(k\wedge\sigma)'}(\Sigma) \mapsto \mathbb{R}$ ($(k\wedge\sigma)'$ is the exponent conjugate to $k\wedge\sigma$) and $\mathcal{K} : L^\infty(\Sigma) \mapsto \mathbb{R}$, defined by

$$\mathcal{S}(\varphi) = -\int_\Sigma \mathcal{H}'_v(s,t,\bar{y},\bar{v},\bar{p},\bar{\alpha})\ \varphi\, ds\, dt \qquad \text{for all } \varphi \in L^{(k\wedge\sigma)'}(\Sigma),$$
$$\mathcal{K}(\chi) = \langle \bar{\zeta}, g'_v(\bar{y},\bar{v})\ \chi\rangle_{*,\Sigma} \qquad \text{for all } \chi \in L^\infty(\Sigma).$$

The optimality condition (3.6) can be rewritten as $\mathcal{S}(\chi) = \mathcal{K}(\chi)$ for all $\chi \in L^\infty(\Sigma)$. Due to the Hahn–Banach extension theorem, there exists $\bar{\mu} \in L^{k\wedge\sigma}(\Sigma)$ such that

$$\int_\Sigma \bar{\mu}\ \chi\, ds\, dt = \langle \bar{\zeta}, g'_v(\bar{y},\bar{v})\ \chi\rangle_{*,\Sigma} \qquad \text{for all } \chi \in L^\infty(\Sigma). \tag{4.5}$$

Moreover, we have

$$\int_\Sigma \bar{\mu}\ \varphi\, ds\, dt = \mathcal{S}(\varphi) \qquad \text{for all } \varphi \in L^{(k\wedge\sigma)'}(\Sigma).$$

If we set $\bar{\eta} = g'_v(\bar{y},\bar{v})^{-1}\ \bar{\mu}$, due to assumption (A7), $\bar{\eta}$ belongs to $L^{k\wedge\sigma}(\Sigma)$, and (4.5) is equivalent to the following equation:

$$\int_\Sigma \bar{\eta}\chi\, ds\, dt = \langle \bar{\zeta}, \chi\rangle_{*,\Sigma} \qquad \text{for all } \chi \in L^\infty(\Sigma). \tag{4.6}$$

The complementarity condition, the adjoint equation, and the optimality condition for $\bar{v}$ follow from (3.4), (3.5), (3.6), and (4.6). Let us prove the nontriviality condition. If $\bar{\alpha} \neq 0$, the proof is complete. If $\bar{\alpha} = 0$, then due to (3.3) we have $\bar{\zeta} \neq 0$, and from (4.6) it follows that $\bar{\eta} \neq 0$.

Step 2. Let us prove that $\bar{\eta}$ belongs to $L^k(\Sigma)$. Let σ_1 be such that $1 < \sigma_1 < \frac{N+1}{N}$. Due to step 1 and assumption (A6) the function $g'_y(\bar{y},\bar{v})\bar{\eta}$ belongs to $L^{k\wedge\sigma_1}(\Sigma)$. From Assumptions (A3)–(A5) and from Proposition 2.5, it follows that $\bar{p}_{|\Sigma}$ belongs to $L^{\sigma_2}(\Sigma)$ for all σ_2 satisfying

$$\sigma_1 < \sigma_2 \qquad \text{and} \qquad \frac{N+1}{2\sigma_1} < \frac{N+1}{2\sigma_2} + \frac{1}{2}. \tag{4.7}$$

Let σ_2 satisfy (4.7). From the regularity of $p_{|\Sigma}$, we deduce that $\mathcal{H}'_v(\cdot,\bar{y},\bar{v},\bar{p},\bar{\alpha})$ belongs to $L^{k\wedge\sigma_2}(\Sigma)$. From (4.3), it follows that $g'_v(\bar{y},\bar{v})\bar{\eta}$ belongs to $L^{k\wedge\sigma_2}(\Sigma)$. With (A7), we deduce that $\bar{\eta}$ belongs to $L^{k\wedge\sigma_2}(\Sigma)$, and with (A6) that $g'_y(\bar{y},\bar{v})\bar{\eta}$ belongs to $L^{k\wedge\sigma_2}(\Sigma)$ for all σ_2 satisfying (4.7). After a finite number of iterations, we can prove that $\bar{\eta}$ belongs to $L^k(\Sigma)$. The proof is complete. □

Remark 4.2. Let us observe that (3.1) can be easily verified when $\mathcal{C} = \{z \in L^\infty(\Sigma) \mid z \le 0\}$, and Ψ is of the form $\Psi(\cdot, y, v) = \psi(\cdot, y) - v$. For $\varepsilon > 0$, let us set $w_\varepsilon = \bar v - \varepsilon g'_v(\bar y, \bar v)^{-1} + g'_v(\bar y, \bar v)^{-1} g'_y(\bar y, \bar v) z_{\bar v}$. Let ξ_ε be the solution to the equation

$$\begin{aligned}
&\frac{\partial \xi}{\partial t} + A\xi + \Phi'_y(\cdot, \bar y)\xi = 0 \quad \text{ in } Q,\\
&\frac{\partial \xi}{\partial n_A} + \psi'_y(\cdot, \bar y)\xi + g'_v(\bar y, \bar v)^{-1} g'_y(\bar y, \bar v)\xi = w_\varepsilon \quad \text{ on } \Sigma, \qquad \xi(0) = 0 \quad \text{ in } \Omega.
\end{aligned}$$

It is clear that ξ_ε is the solution of (3.2) corresponding to $\widehat v = \bar v - \varepsilon g'_v(\bar y, \bar v)^{-1} - g'_v(\bar y, \bar v)^{-1} g'_y(\bar y, \bar v)(\xi_\varepsilon - z_{\bar v})$, which yields to

$$g(\bar y, \bar v) + g'_y(\bar y, \bar v)(z_{\hat v} - z_{\bar v}) + g'_v(\bar y, \bar v)^{-1}(\widehat v - \bar v) = -\varepsilon + g(\bar y, \bar v) \le -\varepsilon.$$

Therefore, the pair $(\widehat v, z_{\hat v})$ satisfies the condition (3.1). In this case, we can set $\bar\alpha = 1$ in the statement of Theorem 4.1.

5. Other regularity results.

In this section, we are concerned with problem $(\mathcal{P})$ when $\mathcal{C} = (\{z \in L^\infty(\Sigma) \mid z \le 0\})^\ell$. In this case, (4.1) is equivalent to

$$(5.1) \qquad (\bar\alpha, \bar\eta) \ne 0, \qquad \bar\eta \ge 0, \qquad \int_\Sigma \bar\eta \; g(\bar y, \bar v)\, ds\, dt = 0.$$

In section 5.1, we suppose that $\ell = 2$ and g satisfies the following separation condition.

(A8) There exists $\varepsilon > 0$ such that $g_1(\bar y, \bar v) + \varepsilon \le -g_2(\bar y, \bar v)$ almost everywhere (a.e.) on Σ. Moreover, we suppose that, for $i = 1, 2$, $(g'_{iv}(\bar y(\cdot), \bar v(\cdot)))^{-1}$ belongs to $L^\infty(\Sigma)$.

Remark 5.1. In examples studied in section 5 (Corollary 5.3 and examples in section 5.3), we are able to verify the qualification condition (3.1) by using both the separation condition and the properties of the state equation. But the separation condition is of a different nature since it gives the regularity of the multipliers (which is not the case of condition (3.1)), and in general it does not give optimality conditions in qualified form (which is the case of condition (3.1)).

With (A8) we are able to prove that the supports of the multipliers associated with the two constraints are disjoint. To prove such a result, we use the isomorphism between $(L^\infty(\Sigma))'$ and the space of bounded Radon measures on $\Sigma^\#$, where $\Sigma^\#$ is a compact Hausdorff space [9].

In section 5.2, we suppose that g satisfies a monotonicity condition of the following form.

(A9) For $i = 1, \ldots, \ell$, $g'_{iv}(\bar y, \bar v) \ge 0$ a.e. on Σ, and $(g'_{iv}(\bar y(\cdot), \bar v(\cdot)))^{-1}$ belongs to $L^\infty(\Sigma)$.

In this case, the regularity of multipliers follows from properties of nonnegative additive measures, and from the Radon–Nikodym theorem.

In section 5.3, we study a problem where the above separation and monotonicity conditions are coupled.

5.1. Regularity of multipliers with a separation assumption.

THEOREM 5.2. *Let* $(\bar y, \bar v)$ *be a solution to* $(\mathcal{P})$. *Suppose that* (A1)–(A6) *and* (A8) *are fulfilled. Then, there exist* $\bar\alpha \in [0, 1]$, $\bar\eta \in (L^k(\Sigma))^2$, *and* $\bar p \in L^1(0, T; W^{1,1}(\Omega))$ *satisfying* (4.2), (4.3), *and* (5.1). *If, in addition,* $\widehat v$ *satisfies* (3.1), *then we can take* $\bar\alpha = 1$.

Recall that k is the exponent introduced in the introduction. It is known that $(L^\infty(\Sigma))'$ can be identified with the space of bounded finitely additive measures vanishing on zero-measure sets [20, Theorem 2.3]. Below we identify $(L^\infty(\Sigma))'$ with $\mathcal{M}(\Sigma^\#)$, the space of Radon measures on $\Sigma^\#$, where $\Sigma^\#$ is a compact Hausdorff space. This identification is useful to characterize the supports of elements of $(L^\infty(\Sigma))'$.

Proof. *Step* 1. Due to Theorem 3.1, there exist $\bar\alpha \in [0,1]$, $\bar\zeta = (\bar\zeta_1, \bar\zeta_2) \in ((L^\infty(\Sigma))')^2$, and $\bar p \in L^1(0,T;W^{1,1}(\Omega))$ such that (3.3)–(3.6) hold. Condition (3.4) may be rewritten as

$$(5.2)\quad \bar\zeta_i \geq 0 \text{ and } \langle \bar\zeta_i, g_i(\bar y, \bar v)\rangle_{*,\Sigma} = \max\ \{\langle \bar\zeta_i, z\rangle_{*,\Sigma} \mid z \in \mathcal{C}\} = 0 \quad \text{for } i = 1,2.$$

Step 2. Let us denote by S_o the closed unit sphere of $(L^\infty(\Sigma))'$ (for the weak-star topology), and let us set

$$\Sigma^\# = \{q \in S_o \mid q \neq 0,\ \langle q, h\rangle_{*,\Sigma} = \langle q, f\rangle_{*,\Sigma}\ \langle q, g\rangle_{*,\Sigma} \ \text{ if } h = g\ f \text{ a.e. on } \Sigma\}.$$

It is well known [9, Theorem 11, p. 445] that $\Sigma^\#$ is a compact Hausdorff space. Moreover, there exists an isometric homomorphism τ from $L^\infty(\Sigma)$ onto $C(\Sigma^\#)$. The isomorphism τ maps nonnegative functions into nonnegative functions and is an algebraic isomorphism in the sense that if $\chi = \chi_1\ \chi_2$ a.e. on Σ, then $\tau(\chi) = \tau(\chi_1)\tau(\chi_2)$. If f is an arbitrary real continuous function, and χ is in $L^\infty(\Sigma)$, then $\tau(f(\chi)) = f(\tau(\chi))$. Hence, for $i = 1,2$, the measure $\bar\zeta_i \in (L^\infty(\Sigma))'$ can be identified with $\widehat{\bar\zeta}_i \in \mathcal{M}(\Sigma^\#)$ (the space of Radon measures on $\Sigma^\#$), via the formula

$$\langle \widehat{\bar\zeta}_i, \psi\rangle_{\Sigma^\#} = \langle \bar\zeta_i, \tau^{-1}(\psi)\rangle_{*,\Sigma} \quad \text{for all} \ \ \psi \in C(\Sigma^\#),$$

where $\langle\cdot,\cdot\rangle_{\Sigma^\#}$ denotes the duality pairing between $\mathcal{M}(\Sigma^\#)$ and $C(\Sigma^\#)$, and where τ^{-1} is the inverse mapping of τ. (For more details see [2].)

Let us prove that under the separation condition (A8), the supports of $\widehat{\bar\zeta}_1$ and $\widehat{\bar\zeta}_2$ (denoted by supp $\widehat{\bar\zeta}_i$) are disjoint. The condition (5.2) is rewritten as

$$\widehat{\bar\zeta}_i \geq 0 \quad \text{and} \quad \langle \widehat{\bar\zeta}_i, \tau(g_i(\bar y, \bar v))\rangle_{\Sigma^\#} = 0 \ \text{ for } i = 1,2.$$

Let us set $\Sigma_i = \{q \in \Sigma^\# \mid \tau(g_i(\bar y, \bar v))(q) = \langle\ q, g_i(\bar y, \bar v)\ \rangle_{*,\Sigma} = 0\}$. Since the mapping $q \mapsto \tau(g_i(\bar y, \bar v))(q)$ is continuous for the weak-star topology of $(L^\infty(\Sigma))'$, Σ_i is closed for this topology. Therefore we have

$$(5.3)\qquad \mathrm{supp}\widehat{\bar\zeta}_i \subset \Sigma_i.$$

On the other hand, due to the positivity property of τ and due to (A8), we have

$$(5.4)\ \tau(g_1(\bar y, \bar v))\ (q) - \varepsilon = \tau(g_1(\bar y, \bar v) - \varepsilon)\ (q) \leq \tau(g_2(\bar y, \bar v))\ (q) \qquad \text{for all } q \in \Sigma^\#.$$

From (5.3) and (5.4), we deduce that supp $\widehat{\bar\zeta}_1 \cap \mathrm{supp}\ \widehat{\bar\zeta}_2 = \emptyset$.

Step 3. Now, let us establish the regularity of $g_v'(\bar y, \bar v)^*\bar\zeta$. First, notice that (3.6) can be stated in the form

$$\int_\Sigma \mathcal{H}_v'(\cdot, \bar y, \bar v, \bar p, \bar\alpha)\chi\, ds\, dt + \langle \widehat{\bar\zeta}_1, \tau(g_{1v}'(\bar y, \bar v))\ \tau(\chi)\rangle_{\Sigma^\#} + \langle \widehat{\bar\zeta}_2, \tau(g_{2v}'(\bar y, \bar v))\ \tau(\chi)\rangle_{\Sigma^\#} = 0$$

for all $\chi \in L^\infty(\Sigma)$. Consider the linear operators $\mathcal{S} : L^{(k\wedge\sigma)'}(\Sigma) \mapsto \mathbb{R}$ and $\mathcal{K} : L^\infty(\Sigma) \mapsto \mathbb{R}$ defined by

$$\mathcal{S}(\varphi) = -\int_\Sigma \mathcal{H}'_v(\cdot, \bar{y}, \bar{v}, \bar{p}, \bar{\alpha})\ \varphi\, ds\, dt \qquad \text{for all } \varphi \in L^{(k\wedge\sigma)'}(\Sigma),$$

$$\mathcal{K}(\chi) = \langle \widehat{\bar{\zeta}}_1, \tau(g'_{1v}(\bar{y}, \bar{v}))\ \tau(\chi)\rangle_{\Sigma^\#} + \langle \widehat{\bar{\zeta}}_2, \tau(g'_{2v}(\bar{y}, \bar{v}))\ \tau(\chi)\rangle_{\Sigma^\#} \quad \text{for all } \chi \in L^\infty(\Sigma).$$

With arguments similar to those of Theorem 4.1, we can prove the existence of a function $\mu \in L^{k\wedge\sigma}(\Sigma)$ satisfying

$$\langle \widehat{\bar{\zeta}}_1, \tau(g'_{1v}(\bar{y}, \bar{v}))\ \tau(\chi)\rangle_{\Sigma^\#} + \langle \widehat{\bar{\zeta}}_2, \tau(g'_{2v}(\bar{y}, \bar{v}))\ \tau(\chi)\rangle_{\Sigma^\#} = \int_\Sigma \mu\ \chi\, ds\, dt \tag{5.5}$$

for all $\chi \in L^\infty(\Sigma)$. Since supp $\widehat{\bar{\zeta}}_1$ and supp $\widehat{\bar{\zeta}}_2$ are two disjoint compact subsets of $\Sigma^\#$, there exists $\psi_o \in C(\Sigma^\#)$, such that

$$0 \le \psi_o \le 1, \quad \psi_o \equiv 1 \quad \text{on supp } \widehat{\bar{\zeta}}_1, \quad \text{and} \quad \psi_o \equiv 0 \quad \text{on supp } \widehat{\bar{\zeta}}_2.$$

Letting $\tilde{\chi}$ be in $L^\infty(\Sigma)$ and setting $\chi_1 = \tau^{-1}(\psi_o)\ \tilde{\chi}$ and $\chi_2 = (1 - \psi_o)\ \tilde{\chi}$ in (5.5), we obtain

$$\langle \bar{\zeta}_i, g'_{iv}(\bar{y}, \bar{v})\tilde{\chi}\rangle_{*,\Sigma} = \langle \widehat{\bar{\zeta}}_i, \tau(g'_{iv}(\bar{y}, \bar{v}))\ \tau(\tilde{\chi})\rangle_{\Sigma^\#} = \int_\Sigma \bar{\mu}_i\ \tilde{\chi}\, ds\, dt \tag{5.6}$$

for $i = 1, 2$, with $\bar{\mu}_1 = \mu\ \tau^{-1}(\psi_o)$ and $\bar{\mu}_2 = \mu\ \tau^{-1}(1 - \psi_o)$. It is clear that $\bar{\mu}_1$ and $\bar{\mu}_2$ belong to $L^{k\wedge\sigma}(\Sigma)$ for all $\sigma < \frac{N+1}{N}$. Let us set $\bar{\eta}_i = (g'_{iv}(\bar{y}, \bar{v}))^{-1}\bar{\mu}_i$ for $i = 1, 2$. We have proved that (4.2), (4.3), and (5.1) are satisfied with $\bar{\eta} = (\bar{\eta}_1, \bar{\eta}_2) \in L^{k\wedge\sigma}(\Sigma)$. We conclude with a bootstrap process as in step 2 of the proof of Theorem 4.1. □

Consider the following example:

$$(\mathcal{P}_1) \qquad \inf\{J(y, v) \mid\ v \in L^\infty(\Sigma),\ (y, v) \text{ satisfies (1.1) and } 0 \le v \le \gamma(y) + c\},$$

where Ψ is of the form $\Psi(\cdot, y, v) = \psi(\cdot, y) - v$. It is a particular case of $(\mathcal{P})$ corresponding to $g_1(s, t, y, v) = v - \gamma(s, t, y) - c(s, t)$ and $g_2(y, v) = -v$. We suppose that c belongs to $L^\infty(\Sigma)$, and γ is defined either by $\gamma(s, t, y) = b(s, t)y$, or by $\gamma(s, t, y) = \phi(y)$, where b belongs to $L^\infty(\Sigma)$, $b \ge 0$, and ϕ is a nondecreasing function of class C^1.

COROLLARY 5.3. *Let $(\bar{y}, \bar{v})$ be a solution of $(\mathcal{P}_1)$, and suppose that* (A1)–(A5) *are fulfilled. Suppose in addition that there exists $\varepsilon > 0$ such that*

$$\gamma(\bar{y}) + c \ge \varepsilon \qquad a.e.\ on\ \Sigma. \tag{5.7}$$

Then, there exist $\bar{\eta} \in (L^k(\Sigma))^2$ and $\bar{p} \in L^1(0, T; W^{1,1}(\Omega))$, such that (4.2), (4.3), *and* (5.1) *are satisfied with $\bar{\alpha} = 1$.*

Remark 5.4. Observe that if $y_o \ge 0$, then $\bar{y} \ge 0$. If ϕ is nonnegative on $\mathbb{R}^+$ and if $c \ge \varepsilon$, then condition (5.7) is satisfied. The case when $\gamma(s, t, y) = b(s, t)y$ is studied in [6]. The proof in [6] is based on duality techniques. Here it is a direct consequence of Theorem 5.2.

Proof. The separation condition (A8) is nothing else than (5.7). Therefore, due to Theorem 5.2, we can state optimality conditions in nonqualified form. To prove the corollary, we have only to check that there exists $\hat{v} \in L^\infty(\Sigma)$ such that (3.1) is satisfied. If $\gamma(s,t,y) = b(s,t)y$, we set $\beta = b$, and if $\gamma(s,t,y) = \phi(y)$, we set $\beta = \phi'(\bar y)$. Therefore $\beta \geq 0$. For $\lambda > 0$, we set $w_\lambda = \bar v - \lambda\varepsilon - \beta z_{\bar v}$. Let ξ_λ be the solution to

$$\frac{\partial \xi}{\partial t} + A\xi + \bar\Phi'_y\,\xi = 0 \ \text{ in } Q, \quad \frac{\partial \xi}{\partial n_A} + \bar\psi'_y\,\xi - \beta\xi = w_\lambda \ \text{ on } \Sigma, \quad \xi(0) = 0 \ \text{ in } \Omega,$$

where $\bar\Phi'_y = \Phi'_y(\cdot,\bar y)$ and $\bar\psi'_y = \psi'_y(\cdot,\bar y)$. Observe that $\xi_\lambda = z_{\tilde v_\lambda}$ is the solution of (3.2) for $\widetilde v_\lambda = \bar v - \lambda\varepsilon + \beta(\xi_\lambda - z_{\bar v})$, and

$$-\beta(z_{\tilde v_\lambda} - z_{\bar v}) + \tilde v_\lambda - \bar v = -\lambda\varepsilon. \tag{5.8}$$

By making the difference between the equation satisfied by $z_{\bar v}$ and the equation satisfied by ξ_λ, we can prove that

$$\xi_\lambda = z_{\tilde v_\lambda} \leq z_{\bar v} \quad \text{and} \quad ||z_{\tilde v_\lambda} - z_{\bar v}||_{C(\overline{Q})} \leq C\lambda\varepsilon. \tag{5.9}$$

By combining (5.9) and (5.8), we prove that there exists a constant $\bar C > 0$ such that

$$||\widetilde v_\lambda - \bar v||_{\infty,\Sigma} \leq \bar C\ \lambda\varepsilon. \tag{5.10}$$

Set $\Sigma_0 = \{(s,t) \in \Sigma \mid \bar v = 0\}$, $\Sigma_\lambda = \{(s,t) \in \Sigma \mid 0 < \bar v \leq 2\bar C\lambda\varepsilon\}$, and let χ_λ (respectively, χ_0) be the characteristic function of Σ_λ (respectively, Σ_0). Set $\widehat v_\lambda = \lambda\varepsilon\chi_0 + (\bar v + \lambda\varepsilon)\chi_\lambda + \widetilde v_\lambda(1 - \chi_0 - \chi_\lambda)$. We claim that the pair $(z_{\hat v_\lambda}, \widehat v_\lambda)$ satisfies (3.1) for λ small enough.

- On $\Sigma_0 \cup \Sigma_\lambda$, we have $\widehat v_\lambda \geq \lambda\varepsilon > 0$. Due to (5.10), on $\Sigma \setminus (\Sigma_0 \cup \Sigma_\lambda)$ we have $\widehat v_\lambda = \widetilde v_\lambda \geq \bar v - |\widetilde v_\lambda - \bar v| \geq \bar C\lambda\varepsilon$. Thus we have $\widehat v_\lambda \geq \min(\lambda\varepsilon, \bar C\lambda\varepsilon)$.
- With (5.10) we obtain

$$||z_{\hat v_\lambda} - z_{\tilde v_\lambda}||_{C(\overline{Q})} \leq C||(\bar v - \widetilde v_\lambda + \lambda\varepsilon)(\chi_\lambda + \chi_0)||_{\infty,\Sigma} \leq C\lambda\varepsilon.$$

Now, from (5.9) we deduce $||z_{\hat v_\lambda} - z_{\bar v}||_{C(\overline{Q})} \leq ||z_{\hat v_\lambda} - z_{\tilde v_\lambda}||_{C(\overline{Q})} + ||z_{\tilde v_\lambda} - z_{\bar v}||_{C(\overline{Q})} \leq C\lambda\varepsilon$. Due to (5.7), we have

$$\bar v - \gamma(\bar y) - c - \beta(z_{\hat v_\lambda} - z_{\bar v}) + (\widehat v_\lambda - \bar v) \leq 2\bar C\lambda\varepsilon - \varepsilon - \beta(z_{\hat v_\lambda} - z_{\bar v}) + \lambda\varepsilon \ \text{ on } \Sigma_0 \cup \Sigma_\lambda.$$

Therefore, we can choose $\lambda > 0$ small enough to have

$$\bar v - \gamma(\bar y) - c - \beta(z_{\hat v_\lambda} - z_{\bar v}) + (\widehat v_\lambda - \bar v) \leq -\frac{\varepsilon}{2} \qquad \text{on } \Sigma_0 \cup \Sigma_\lambda.$$

Since $\tilde v_\lambda = -\lambda\varepsilon + \beta(z_{\tilde v_\lambda} - z_{\bar v}) < 0$ on Σ_0 (because of (5.9)), we have $\widehat v_\lambda - \tilde v_\lambda \geq 0$ on $\Sigma \setminus \Sigma_\lambda$. Hence if ζ_λ is the solution to (3.2) for $v = (\widehat v_\lambda - \tilde v_\lambda)\chi_\lambda = (\bar v + \lambda\varepsilon - \tilde v_\lambda)\chi_\lambda$, we have $z_{\hat v_\lambda} - z_{\tilde v_\lambda} \geq \zeta_\lambda$, and $||\zeta_\lambda||_{C(\overline{Q})} \leq C||(\bar v + \lambda\varepsilon - \tilde v_\lambda)\chi_\lambda||_{r,\Sigma} \leq C\lambda\varepsilon\ \mathcal{L}^N(\Sigma_\lambda)^{\frac{1}{r}}$, with $r > N+1$, and $\mathcal{L}^N(\Sigma_\lambda)$ is the Lebesgue measure of Σ_λ. With (5.8) and the bound for ζ_λ, we have the following estimate on $\Sigma \setminus (\Sigma_0 \cup \Sigma_\lambda)$:

$$\begin{aligned} &\bar v - \gamma(\bar y) - c - \beta(z_{\hat v_\lambda} - z_{\bar v}) + (\widehat v_\lambda - \bar v) \leq -\beta(z_{\tilde v_\lambda} - z_{\bar v}) + (\widetilde v_\lambda - \bar v) - \beta(z_{\hat v_\lambda} - z_{\tilde v_\lambda}) \\ &\leq -\lambda\varepsilon + ||\beta||_{\infty,\Sigma}\ ||\zeta_\lambda||_{\infty,\Sigma} \leq -\lambda\varepsilon + C\lambda\varepsilon\ \mathcal{L}^N(\Sigma_\lambda)^{\frac{1}{r}}. \end{aligned}$$

Since $-\lambda\varepsilon + C\lambda\varepsilon\ \mathcal{L}^N(\Sigma_\lambda)^{\frac{1}{r}} < -\frac{1}{2}\lambda\varepsilon$ for $\lambda > 0$ small enough, the proof is complete. □

5.2. Regularity of multipliers with a monotonicity assumption.

THEOREM 5.5. *Let* $(\bar y, \bar v)$ *be a solution of* $(\mathcal{P})$*. Suppose that* (A1)–(A6) *and* (A9) *are fulfilled. Then, there exist* $\bar\alpha \in [0,1]$, $\bar\eta \in (L^k(\Sigma))^\ell$, *and* $\bar p \in L^1(0,T;W^{1,1}(\Omega))$*), such that conditions* (4.2), (4.3), *and* (5.1) *are satisfied.*

Proof. Due to Theorem 3.1, there exist $\bar\alpha \in [0,1]$, $\bar\zeta = (\bar\zeta_1, \ldots, \bar\zeta_\ell) \in ((L^\infty(\Sigma))^\ell)'$, and $\bar p \in L^1(0,T;W^{1,1}(\Omega))$ such that (3.3)–(3.6) hold. With arguments similar to those of the proof of Theorem 5.2, we show that

$$\bar\zeta_i \geq 0, \quad \langle \bar\zeta_i, g_i(\bar y, \bar v)\rangle_{*,\Sigma} = 0 \qquad \text{for all } i = 1, \ldots, \ell. \tag{5.11}$$

Moreover, as in the proof of Theorem 4.1, with the Hahn–Banach extension theorem, we can establish the existence of a multiplier $\bar\mu$ belonging to $L^{k\wedge\sigma}(\Sigma)$ for all $\sigma < \frac{N+1}{N}$, such that

$$\langle g'_v(\bar y, \bar v)^* \bar\zeta, \chi\rangle_{*,\Sigma} = \int_\Sigma \bar\mu\, \chi\, ds\, dt \qquad \text{for all } \chi \in L^\infty(\Sigma).$$

Due to (A9), for $i = 1, \ldots, \ell$, the additive measures $g'_{iv}(\bar y, \bar v)^* \bar\zeta_i$ are nonnegative. From a decomposition theorem for nonnegative finitely additive measures [9, Theorem 8, p. 163], we deduce that $g'_{iv}(\bar y, \bar v)^* \bar\zeta_i$ admits a unique decomposition

$$g'_{iv}(\bar y, \bar v)^* \bar\zeta_i = \bar\mu_i + \bar\nu_i,$$

where $\bar\mu_i$ are nonnegative countably additive measures, and $\bar\nu_i$ are nonnegative purely finitely additive set functions in $(L^\infty(\Sigma))'$. It follows that

$$g'_v(\bar y, \bar v)^* \bar\zeta = \sum_{i=1}^{\ell} \bar\mu_i + \sum_{i=1}^{\ell} \bar\nu_i = \bar\mu \;\; ds\, dt.$$

Thus, the pair $(\Sigma_{i=1}^{\ell} \bar\mu_i, \Sigma_{i=1}^{\ell} \bar\nu_i)$ is a decomposition of $\bar\mu\, dsdt$ in a nonnegative countably additive measure and a nonnegative purely finitely additive set function. From the uniqueness of this decomposition, we deduce that $\bar\nu_i = 0$. Since $\bar\mu_i \geq 0$ for $i = 1, \ldots, \ell$, $\Sigma_{i=1}^{\ell} \bar\mu_i = \bar\mu\, dsdt$, and $\bar\mu$ belongs to $L^{k\wedge\sigma}(\Sigma)$ for all $\sigma < \frac{N+1}{N}$, there exist nonnegative functions $\xi_i \in L^{k\wedge\sigma}(\Sigma)$ such that $\bar\mu_i = \xi_i\, ds\, dt$ (it is a consequence of the Radon–Nikodym theorem). We set $\bar\eta_i = (g'_{iv}(\bar y, \bar u))^{-1} \xi_i$. We have

$$\langle g'_{iv}(\bar y, \bar v)^* \bar\zeta_i, \chi\rangle_{*,\Sigma} = \int_\Sigma \xi_i\, \chi\, ds\, dt = \int_\Sigma g'_{iv}(\bar y, \bar v) \bar\eta_i\, \chi\, ds\, dt$$

for all $\chi \in L^\infty(\Sigma)$. Thus, (4.2), (4.3), and (5.1) are satisfied with $\bar\eta = (\bar\eta_1, \ldots, \bar\eta_\ell) \in (L^{k\wedge\sigma}(\Sigma))^\ell$. We finish with a bootstrap argument as in the proof of Theorem 4.1. □

Consider the following example.

$$(\mathcal{P}_2) \;\; \inf\{J(y,v) \mid \; v \in L^\infty(\Sigma),\; (y,v) \text{ satisfies (1.1)},\; v \leq \gamma_i(y) + c_i \;\text{ for } i = 1, \ldots, \ell\},$$

where Ψ is of the form $\Psi(\cdot, y, v) = \psi(\cdot, y) - v$. It is a particular case of $(\mathcal{P})$ corresponding to $g_i(y,v) = v - \gamma_i(y) - c_i$ for $i = 1, \ldots, \ell$. We suppose that c_i belongs to $L^\infty(\Sigma)$, and γ_i is defined either by $\gamma_i(s,t,y) = b_i(s,t)y$, or by $\gamma_i(s,t,y) = \phi_i(y)$, where b_i belongs to $L^\infty(\Sigma)$, and ϕ_i is of class C^1.

COROLLARY 5.6. *Let* $(\bar y, \bar v)$ *be a solution of* $(\mathcal{P}_2)$*. Suppose that* (A1)–(A5) *are fulfilled. Then there exist* $\bar\eta \in (L^k(\Sigma))^\ell$ *and* $\bar p \in L^1(0,T;W^{1,1}(\Omega))$*), such that* (4.2), (4.3)*, and* (5.1) *are satisfied with* $\bar\alpha = 1$.

Proof. The assumptions of Theorem 5.5 are clearly satisfied. We have only to check that we can take $\bar{\alpha} = 1$. If $\gamma_i(s,t,y) = b_i(s,t)y$, we set $\beta_i = b_i$, and if $\gamma_i(s,t,y) = \phi_i(y)$, we set $\beta_i = \phi_i'(\bar{y})$. For $\lambda > 0$ and $i = 1, \dots, \ell$, we set $w_{i\lambda} = \bar{v} - \lambda\varepsilon - \beta_i z_{\bar{v}}$. Let ξ_λ be the solution to

$$\begin{aligned} &\frac{\partial \xi}{\partial t} + A\xi + \Phi_y'(\cdot,\bar{y})\xi = 0 && \text{in } Q, \\ &\frac{\partial \xi}{\partial n_A} + \psi_y'(\cdot,\bar{y})\xi = \min_{i=1,\dots,\ell}(w_{i\lambda} + \beta_i \xi) \quad \text{on } \Sigma, && \xi(0) = 0 \text{ in } \Omega. \end{aligned} \tag{5.12}$$

Since the function "min" is Lipschitz with respect to its arguments, the existence of a solution to (5.12) may be proved as in [16] (see also [6] where the same trick is used). Observe that ξ_λ is the solution of (3.2) corresponding to $\widetilde{v}_\lambda = \min_{i=1,\dots,\ell}(w_{i\lambda} + \beta_i\xi)$, that is, $\xi_\lambda = z_{\widetilde{v}_\lambda}$. Moreover,

$$\bar{v} - \gamma_i(\bar{y}) - c_i + (\widetilde{v}_\lambda - \bar{v}) - \beta_i(z_{\widetilde{v}_\lambda} - z_{\bar{v}}) \leq (\widetilde{v}_\lambda - \bar{v}) - \beta_i(z_{\widetilde{v}_\lambda} - z_{\bar{v}}) \leq -\lambda\varepsilon$$

for $i = 1, \dots, \ell$. Thus (3.1) is satisfied by $(\widetilde{v}_\lambda, z_{\widetilde{v}_\lambda})$. □

5.3. Coupling separation and monotonicity conditions. In this section we study a problem for which the monotonicity and separation conditions are coupled. We suppose that $\ell = 3$ and that $g = (g_1, g_2, g_3)$ satisfies the assumption below.

(A10) There exists $\varepsilon > 0$ such that $g_1(\bar{y},\bar{v}) + \varepsilon \leq -g_2(\bar{y},\bar{v})$ and $g_3(\bar{y},\bar{v}) + \varepsilon \leq -g_2(\bar{y},\bar{v})$ a.e. on Σ. The pair $(g_1(\bar{y},\bar{v}), g_3(\bar{y},\bar{v}))$ satisfies the monotonicity condition stated in (A9). The function $(g_{2v}'(\bar{y}(\cdot),\bar{v}(\cdot)))^{-1}$ belongs to $L^\infty(\Sigma)$.

The case of bottleneck type constraints studied in [5], [6] falls into this setting (see Remark 5.8). We prove the existence of regular multipliers when (A10) is fulfilled. At the end of the section we give two examples for which optimality conditions are obtained in qualified form (that is, $\bar{\alpha} = 1$). The existence of regular multipliers for these examples cannot be deduced from [6].

THEOREM 5.7. *Let $(\bar{y},\bar{u})$ be a solution to (P). Suppose that* (A1)–(A6) *and* (A10) *are fulfilled. Then, there exist $\bar{\alpha} \in [0,1]$, $\bar{\eta} \in (L^k(\Sigma))^3$, and $\bar{p} \in L^1(0,T;W^{1,1}(\Omega))$, such that* (4.2), (4.3)*, and* (5.1) *are satisfied. If, in addition, $\widehat{v}$ satisfies* (3.1)*, then we can take $\bar{\alpha} = 1$.*

Proof. Due to Theorem 3.1, there exist $\bar{\alpha} \in [0,1]$, $\bar{\zeta} \in ((L^\infty(\Sigma))^3)'$, and $\bar{p} \in L^1(0,T;W^{1,1}(\Omega))$ such that (3.3)–(3.6) are satisfied. In particular, the condition (3.4) is equivalent to

$$\bar{\zeta}_i \geq 0 \qquad \text{and} \qquad \langle \bar{\zeta}_i, g_i(\bar{y},\bar{v})\rangle_{*,\Sigma} = 0 \tag{5.13}$$

for $i = 1, 2, 3$. As in the proof of Theorem 5.2, we identify the additive measures $\bar{\zeta}_i$ with measures $\widehat{\bar{\zeta}}_i$ belonging to $\mathcal{M}(\Sigma^\#)$ (the space of Radon measures on $\Sigma^\#$), and the optimality condition (3.6) is

$$\int_\Sigma \mathcal{H}_v'(\cdot,\bar{y},\bar{v},\bar{p},\bar{\alpha})\chi \, ds\, dt + \sum_{i=1}^{3} \langle \widehat{\bar{\zeta}}_i, \tau(g_{iv}'(\bar{y},\bar{v}))\ \tau(\chi)\rangle_{\Sigma^\#} = 0$$

for all $\chi \in L^\infty(\Sigma)$. ($\tau$ is the isometric homomorphism used in the proof of Theorem 5.2, and $\langle\cdot,\cdot\rangle_{\Sigma^\#}$ denotes the duality pairing between $\mathcal{M}(\Sigma^\#)$ and $C(\Sigma^\#)$.) With the Hahn–Banach theorem, there exists a function μ belonging to $L^{k\wedge\sigma}(\Sigma)$ for all

$1 < \sigma < \frac{N+1}{N}$, such that

$$\sum_{i=1}^{3} \langle \widehat{\bar{\zeta}}_i, \tau(g'_{iv}(\bar{y}, \bar{v}))\ \tau(\chi) \rangle_{\Sigma^\#} = \int_\Sigma \mu\ \chi\, ds\, dt \qquad \text{for all } \chi \in L^\infty(\Sigma).$$

By using (5.13), and since $(g_1(\bar{y}, \bar{v}), g_2(\bar{y}, \bar{v}))$ and $(g_2(\bar{y}, \bar{v}), g_3(\bar{y}, \bar{v}))$ satisfy the separation condition stated in (A10), we can prove that

$$\text{supp}\ \widehat{\bar{\zeta}}_1 \cap \text{supp}\ \widehat{\bar{\zeta}}_2 = \text{supp}\ \widehat{\bar{\zeta}}_2 \cap \text{supp}\ \widehat{\bar{\zeta}}_3 = \emptyset. \tag{5.14}$$

Since (supp $\widehat{\bar{\zeta}}_1 \cup$ supp $\widehat{\bar{\zeta}}_3$) and supp $\widehat{\bar{\zeta}}_2$ are two disjoint compact subsets of $\Sigma^\#$, by using the same method as in the proof of Theorem 5.2 we deduce that there exist $\bar{\mu}_1$ and $\bar{\mu}_2$ belonging to $L^{k\wedge\sigma}(\Sigma)$ for all $1 < \sigma < \frac{N+1}{N}$, such that

$$\langle \bar{\zeta}_1, g'_{1v}(\bar{y}, \bar{v})\chi \rangle_{*,\Sigma} + \langle \bar{\zeta}_3, g'_{3v}(\bar{y}, \bar{v})\chi \rangle_{*,\Sigma} = \int_\Sigma \bar{\mu}_1\ \chi\, ds\, dt \quad \text{for all } \chi \in L^\infty(\Sigma), \tag{5.15}$$

$$\langle \bar{\zeta}_2, g'_{2v}(\bar{y}, \bar{v})\chi \rangle_{*,\Sigma} = \int_\Sigma \bar{\mu}_2\ \chi\, ds\, dt \qquad \text{for all } \chi \in L^\infty(\Sigma). \tag{5.16}$$

Let us set $\bar{\eta}_2 = (g'_{2v}(\bar{y}, \bar{v}))^{-1}\bar{\mu}_2$. The function $\bar{\eta}_2$ belongs to $L^{k\wedge\sigma}(\Sigma)$ for all $1 < \sigma < \frac{N+1}{N}$. Since $g'_{1v}(\bar{y}, \bar{v})$ and $g'_{3v}(\bar{y}, \bar{v})$ satisfy (A9) with (5.13) and (5.15), as in the proof of Theorem 5.5, we can establish the existence of two nonnegative functions $\bar{\xi}_1 \in L^{k\wedge\sigma}(\Sigma)$ and $\bar{\xi}_3 \in L^{k\wedge\sigma}(\Sigma)$ such that

$$\langle g'_{1v}(\bar{y}, \bar{v})\bar{\zeta}, \chi \rangle_{*,\Sigma} = \int_\Sigma \bar{\xi}_1 \chi\, ds\, dt, \qquad \langle g'_{3v}(\bar{y}, \bar{v})\bar{\zeta}_3, \chi \rangle_{*,\Sigma} = \int_\Sigma \bar{\xi}_3\ \chi\, ds\, dt \tag{5.17}$$

for all $\chi \in L^\infty(\Sigma)$. We set $\bar{\eta}_1 = (g'_{1v}(\bar{y}, \bar{v}))^{-1}\bar{\xi}_1$ and $\bar{\eta}_3 = (g'_{3v}(\bar{y}, \bar{v}))^{-1}\bar{\xi}_3$. The conditions (4.2), (4.3), and (5.1) are satisfied with $\bar{\eta} \in (L^{k\wedge\sigma}(\Sigma))^3$. We can conclude with a bootstrap argument as in the proof of Theorem 4.1. It is clear that if, in addition, $\widehat{v}$ satisfies (3.1), then we can take $\bar{\alpha} = 1$. The proof is complete. □

Finally, consider two examples for which the assumptions of Theorem 5.7 are satisfied.

Example 1. Set $g_1(y, v) = v - \beta_1 y - c_1$, $g_3(y, v) = v - \beta_3 y - c_3$, $g_2(y, v) = -v$, where β_1, β_3, c_1 and c_3 belong to $L^\infty(\Sigma)$, $\beta_1 \geq 0$, $\beta_3 \geq 0$. Suppose that Ψ is of the form $\Psi(\cdot, y, v) = \psi(\cdot, y) - v$. Suppose that there exists $\varepsilon > 0$ such that $\min(\beta_1\bar{y} + c_1, \beta_3\bar{y} + c_3) \geq \varepsilon$. The pairs $(g_1(\bar{y}, \bar{v}), g_2(\bar{y}, \bar{v}))$ and $(g_2(\bar{y}, \bar{v}), g_3(\bar{y}, \bar{v}))$ satisfy the separation condition stated in (A10). The pair $(g_1(\bar{y}, \bar{v}), g_3(\bar{y}, \bar{v}))$ satisfies the monotonicity condition stated in (A9). Thus (A10) is satisfied and Theorem 5.7 can be applied. Moreover, by combining the arguments of the proofs of Corollaries 5.3 and 5.6, we can prove the existence of $\widehat{v}$ obeying (3.1). Thus we can take $\bar{\alpha} = 1$.

Remark 5.8. If we set $\beta_3 = 0$ in the above example, we recover the case of constraints considered in [6].

Example 2. We replace g_1 and g_3 in the above example by $g_1(y, v) = v - \phi_1(y) - c_1$, $g_3(y, v) = v - \phi_3(y) - c_3$, where ϕ_1 and ϕ_3 are nondecreasing functions of class C^1. Suppose that there exists $\varepsilon > 0$ such that $\min(\phi_1(\bar{y}) + c_1, \phi_3(\bar{y}) + c_3) \geq \varepsilon$. Thus (A10) is satisfied and Theorem 5.7 can be applied. As above, we can take $\bar{\alpha} = 1$.

Acknowledgments. The authors wish to thank the anonymous referees for helpful remarks and suggestions that improved the presentation of the paper.

REFERENCES

[1] N. Arada and J.-P. Raymond, *Necessary optimality conditions for control problems and the Stone–Čech compactification*, SIAM J. Control Optim., 37 (1999), pp. 1011–1032.

[2] N. Arada and J.-P. Raymond, *Minimax control of parabolic systems with state constraints*, SIAM J. Control Optim., 38 (1999), pp. 254–271.

[3] A. V. Arutynov, *Perturbations of extremal problems with constraints and necessary optimality conditions*, J. Sov. Math., 54 (1991), pp. 1342–1400.

[4] A. V. Arutynov and A. I. Okoulevitch, *Necessary optimality conditions for optimal control problems with intermediate constraints*, J. Dynam. Control Systems, 4 (1998), pp. 49–58.

[5] M. Bergounioux and F. Tröltzsch, *Optimal control of linear bottleneck problems*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 235–250.

[6] M. Bergounioux and F. Tröltzsch, *Optimal control of semilinear parabolic equations with state constraints of bottleneck type*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 595–608.

[7] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[8] E. Casas, J.-P. Raymond, and H. Zidani, *Pontryagin's principle for local solutions of control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1182–1203.

[9] N. Dunford and J. T. Schwartz, *Linear Operators, Part* 1, Interscience Publishers, New York, London, 1958.

[10] H. O. Fattorini, *Infinite Dimensional Optimization and Control Theory, Encyclopedia of Mathematics and its Applications*, Cambridge University Press, Cambridge, UK, 1999.

[11] R. F. Hartl, S. P. Sethi, and R. G. Vickson, *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Rev., 37 (1995), pp. 181–218.

[12] X. J. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, Basel, Berlin, 1995.

[13] K. Makowski and L. W. Neustadt, *Optimal control problems with mixed control-phase variable equality and inequality constraints*, SIAM J. Control, 12 (1974), pp. 184–228.

[14] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962.

[15] J.-P. Raymond, *Nonlinear boundary control of semilinear parabolic equations with pointwise state constraints*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 341–370.

[16] J.-P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim., 39 (1999), pp. 143–177.

[17] J.-P. Raymond and H. Zidani, *Time optimal problems with boundary controls*, Differential Integral Equations, 13 (2000), pp. 1039–1072.

[18] T. Roubiček, *Relaxation in Optimization Theory and Variational Calculus*, de Gruyter Ser. Nonlinear Anal. Appl. 4, Walter de Gruyter, Berlin, 1997.

[19] F. Tröltzsch, *Optimality Conditions for Parabolic Control Problems and Applications*, Teubner-Texte Math. 62, B.G. Teubner Verlagsgesellschaft, Leipzig, 1984.

[20] K. Yosida and E. Hewitt, *Finitely additive measures*, Trans. Amer. Math. Soc., 72 (1952), pp. 46–66.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1408–1422

BOUNDEDNESS PROPERTIES FOR TIME-VARYING NONLINEAR SYSTEMS*

JOAN PEUTEMAN[†] [‡], DIRK AEYELS[†], AND RODOLPHE SEPULCHRE[§]

Abstract. A Liapunov theorem guaranteeing uniform boundedness and uniform ultimate boundedness for a time-varying nonlinear system $\dot{x}(t) = f(x(t), t)$ has been established. The study of uniform boundedness and uniform ultimate boundedness of particular classes of time-varying nonlinear systems $\dot{x}(t) = f(x(t), t)$ is reduced to the study of the corresponding time-invariant frozen systems $\dot{x}(t) = f(x(t), \sigma)$ for all $\sigma \in \mathbb{R}$. This approach is illustrated for time-varying homogeneous systems with a positive order, for particular classes of time-varying nonhomogeneous systems and for time-varying Lotka–Volterra equations.

Key words. nonlinear systems, homogeneous systems, uniform boundedness, uniform ultimate boundedness

AMS subject classifications. 34, 34D, 34D20, 34D40

PII. S0363012999361652

1. Introduction.

The stability analysis of time-varying systems $\dot{x}(t) = f(x(t), t)$ is, in general, more difficult than the stability analysis of time-invariant systems. For this reason, several approaches have been proposed in the literature to reduce the stability analysis of time-varying systems to the stability analysis of related time-invariant systems.

Averaging is the most popular of these techniques. Exponential stability of the (time-invariant) averaged system $\dot{x}(t) = \bar{f}(x(t))$ implies exponential stability of the original (time-varying) system provided that the time-variation of the original system is sufficiently fast [1, 2, 8]. In contrast, when the time-variation is sufficiently slow, other results have been proposed based on the stability analysis of the familiy of the frozen systems $\dot{x}(t) = f(x(t), \sigma)$ (where σ is treated as a constant parameter) [3, 8, 12, 13, 14, 15].

In the recent paper [11], it has been observed that the fast time-variation hypothesis necessary for averaging results can be replaced by a homogeneity assumption on the vector field $f(x, t)$. Because the homogeneity property affects state but not time, the time-variation of a homogeneous vector field $f(x, t)$ of positive order $\tau > 0$ is inherently fast for $\|x\|$ small and slow for $\|x\|$ large. This fast and slow variation of the vector field is, of course, to be understood relatively to the time-variation of the solutions.

Based on this observation, the main result in [11] concludes local uniform asymptotic stability of the equilibrium point $x = 0$ of the time-varying homogeneous system

*Received by the editors September 17, 1999; accepted for publication (in revised form) August 15, 2000; published electronically December 28, 2000. This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with its authors.

http://www.siam.org/journals/sicon/39-5/36165.html

[†] SYSTeMS, Universiteit Gent, Technologiepark-Zwijnaarde, 9, 9052 Gent (Zwijnaarde), Belgium (Dirk.Aeyels@rug.ac.be).

[‡] Present address: KHBO, Departement Industriële Wetenschappen en Technologie, Zeedijk 101, 8400 Oostende, Belgium (joan.peuteman@kh.khbo.be).

[§] Institut Montefiore, B28, Université de Liège, 4000 Liège Sart-Tilman, Belgium (r.sepulchre@ulg.ac.be).

from asymptotic stability of the averaged system. This result exploits the inherently *fast* character of homogeneous systems with a positive order *near the origin.*

In the present paper, we exploit the inherently *slow* character of such systems *far from the origin.* We develop a freezing result for homogeneous systems with a positive order: we show that asymptotic stability of each frozen system implies uniform boundedness and uniform ultimate boundedness of the original time-varying system. Boundedness properties rather than asymptotic stability of the equilibrium point follows from the fact that the time-variation is not slow near the origin.

Our result is further extended to systems that are not necessarily homogeneous but that possess a homogeneous approximation for $\|x\|$ sufficiently large. This robustness result is dual to the robustness of local asymptotic stability with respect to higher order perturbations [6, 10].

The paper is organized as follows. In section 2, we formulate a Liapunov result guaranteeing uniform boundedness and uniform ultimate boundedness of a time-varying system $\dot{x}(t) = f(x(t), t)$. We also explain how a Liapunov function can be constructed satisfying the conditions of this Liapunov result. This approach is used in section 3 to prove uniform boundedness and uniform ultimate boundedness of a time-varying homogeneous system with a positive order $\tau > 0$. In section 4, we show that the approach presented in sections 2 and 3 is not restricted to the class of homogeneous systems. In section 5, we illustrate the results by means of a time-varying Lotka–Volterra system defined in the first closed orthant of $\mathbb{R}^n$.

2. Boundedness properties and the freezing technique. We first specify the class of systems under study in the present paper.

Consider

$$\dot{x}(t) = f(x(t), t) \tag{2.1}$$

with $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$. We assume that conditions are imposed on (2.1) such that the existence and uniqueness of its solutions are secured for all initial conditions $x_0 \in \mathbb{R}^n$ and for all initial times t_0. The solution of (2.1) at t with initial condition x_0 at t_0 is denoted as $x(t, t_0, x_0)$. These existence and uniqueness conditions are imposed on all the differential equations mentioned in the present paper.

We now introduce the notions of uniform boundedness and uniform ultimate boundedness (see [16, pp. 36–37]).

DEFINITION 2.1. *The system* (2.1) *is* uniformly bounded *when*[1] *for all* $R_1 > 0$, *there exists an* $R_2(R_1) > 0$ *such that for all* $x_0 \in \mathbb{R}^n$, *for all* t_0, *and for all* $t \geq t_0$

$$\|x_0\| \leq R_1 \Rightarrow \|x(t, t_0, x_0)\| \leq R_2(R_1). \tag{2.2}$$

DEFINITION 2.2. *The system* (2.1) *is* uniformly ultimately bounded *when there exists an* $R > 0$ *such that for all* $R_1 > 0$, *there exists a* $T(R_1) > 0$ *such that for all* $x_0 \in \mathbb{R}^n$, *for all* t_0, *and for all* $t \geq t_0 + T(R_1)$

$$\|x_0\| \leq R_1 \Rightarrow \|x(t, t_0, x_0)\| \leq R. \tag{2.3}$$

The classical theorem of Liapunov proves uniform asymptotic stability of the equilibrium point $x = 0$ of a dynamical system $\dot{x}(t) = f(x(t), t)$ when there exists a positive definite and decrescent Liapunov function $V(x, t)$ whose derivative $\dot{V}(x, t)$

[1] In the present paper, we always use—without loss of generality—the Euclidean norm.

along the solutions of the system is negative definite. When there exists an $R_V > 0$ such that the derivative $\dot{V}(x,t)$ along the solutions of the system is negative for x with $\|x\| > R_V > 0$, the following proposition proves uniform boundedness and uniform ultimate boundedness.

Consider the system (2.1). Consider a continuously differentiable function $V : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$.

PROPOSITION 2.3. *Assume that for all $t \in \mathbb{R}$ and for all $x \in \mathbb{R}^n$*

$$\alpha(\|x\|) \leq V(x,t) \leq \beta(\|x\|). \tag{2.4}$$

The functions $\alpha(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ and $\beta(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ are class-$\mathcal{K}_\infty$ functions.[2]

If there exist a class-$\mathcal{K}$ function $\gamma(\cdot) : \mathbb{R}^+ \to \mathbb{R}$ and an $R_V > 0$ such that for all $t \in \mathbb{R}$ and for all $x \in \mathbb{R}^n$ with $\|x\| > R_V$

$$\frac{\partial V}{\partial t}(x,t) + \frac{\partial V}{\partial x}(x,t) f(x,t) \leq -\gamma(\|x\|), \tag{2.5}$$

then (2.1) *is uniformly bounded and uniformly ultimately bounded.*

Proof. The result of the present proposition has been formulated in [16, pp. 39–42]. For completeness, the proof has been included in the appendix. □

The study of uniform boundedness and uniform ultimate boundedness of a time-varying nonlinear system, by means of Proposition 2.3, is, in general, highly nontrivial. Reducing the problem to the study of time-invariant systems may be an important simplification.

Consider the time-varying system (2.1). For each $\sigma \in \mathbb{R}$, define the time-invariant system

$$\dot{x}(t) = f(x(t), \sigma). \tag{2.6}$$

We call the system (2.6) the frozen system of (2.1) at σ. Consider for each $\sigma \in \mathbb{R}$ a continuously differentiable function $V_\sigma : \mathbb{R}^n \to \mathbb{R}$. Define $V : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ as $V(x,\sigma) := V_\sigma(x)$ for each $\sigma \in \mathbb{R}$ and each $x \in \mathbb{R}^n$.

THEOREM 2.4. *Assume that for all $\sigma \in \mathbb{R}$ and for all $x \in \mathbb{R}^n$*

$$W_1(x) \leq V(x,\sigma) \leq W_2(x). \tag{2.7}$$

The functions $W_1 : \mathbb{R}^n \to \mathbb{R}$ and $W_2 : \mathbb{R}^n \to \mathbb{R}$ are positive definite and radially unbounded.[3]

If there exist an $R_V > 0$ and positive definite functions $W_3 : \mathbb{R}^n \to \mathbb{R}$, $W_4 : \mathbb{R}^n \to \mathbb{R}$, and $W_5 : \mathbb{R}^n \to \mathbb{R}$ such that for all $\sigma \in \mathbb{R}$ and for all $x \in \mathbb{R}^n$ with $\|x\| > R_V$

$$\left| \frac{\partial V}{\partial \sigma}(x,\sigma) \right| \leq W_3(x), \tag{2.8}$$

$$\frac{\partial V}{\partial x}(x,\sigma) f(x,\sigma) \leq -W_4(x), \tag{2.9}$$

[2] A continuous function $\eta : [0,a) \to [0,\infty)$ is said to be a class-$\mathcal{K}$ function if it is strictly increasing and $\eta(0) = 0$. It is said to be a class-$\mathcal{K}_\infty$ function if $a = \infty$ and $\eta(r) \to \infty$ as $r \to \infty$.

[3] A function $W : \mathbb{R}^n \to \mathbb{R}$ is positive definite when W is continuous, $W(0) = 0$, and $W(x) > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$. In the case when $W(x) \to +\infty$ as $\|x\| \to +\infty$, the positive definite function W is radially unbounded.

$$W_3(x) - W_4(x) \leq -W_5(x), \tag{2.10}$$

then (2.1) *is uniformly bounded and uniformly ultimately bounded.*

Proof. The proof of the present theorem is based on Proposition 2.3. For each $t \in \mathbb{R}$ and for each $x \in \mathbb{R}^n$, define

$$V(x,t) := V(x,\sigma)\Big|_{\sigma=t}. \tag{2.11}$$

By (2.7) and [8, pp. 138–139], (2.4) is satisfied. By (2.8) and (2.9), it is clear that for all $t \in \mathbb{R}$ and for all $x \in \mathbb{R}^n$ with $\|x\| > R_V$

$$\frac{\partial V}{\partial t}(x,t) + \frac{\partial V}{\partial x}(x,t)f(x,t) \leq W_3(x) - W_4(x) \leq -W_5(x) < 0. \tag{2.12}$$

By [8, pp. 138–139], (2.5) is satisfied, and by Proposition 2.3 this implies uniform boundedness and uniform ultimate boundedness for the original time-varying system (2.1). □

Remark 1. It is obvious that the statement of Theorem 2.4 can be relaxed by replacing (2.8) and (2.9) by $\frac{\partial V}{\partial \sigma}(x,\sigma) + \frac{\partial V}{\partial x}(x,\sigma)f(x,\sigma) \leq -W_5(x)$. However, for the purpose we have in mind (see section 3), the present formulation of Theorem 2.4 will be applied.

3. Homogeneous systems.

In the present section, we specialize the result of the previous sections to the class of homogeneous systems.

Given an n-tuple $r = (r_1, \dots, r_n)$ (for all $i \in \{1, \dots, n\} : r_i > 0$), we define the dilation δ to be the map

$$\delta : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n : (s,x) \to \delta(s,x) = (s^{r_1}x_1, \dots, s^{r_n}x_n), \tag{3.1}$$

where $x = (x_1, \dots, x_n)$.

A continuous function $h : \mathbb{R}^n \to \mathbb{R}$ is r-homogeneous of degree $m \geq 0$ if and only if

$$\forall x \in \mathbb{R}^n, \forall s \in \mathbb{R}^+ : h(\delta(s,x)) = s^m h(x). \tag{3.2}$$

A continuous function $f_H : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is r-homogeneous of order $\tau \geq 0$ if and only if

$$\forall x \in \mathbb{R}^n, \forall t \in \mathbb{R}, \forall s \in \mathbb{R}^+ : f_H(\delta(s,x),t) = s^\tau \delta(s, f_H(x,t)). \tag{3.3}$$

When f_H is r-homogeneous of order $\tau \geq 0$, then for all $p > 0$, f_H is r'-homogeneous of order $\tau' \geq 0$ with $r' = (\frac{r_1}{p}, \dots, \frac{r_n}{p})$ and $\tau' = \frac{\tau}{p}$. When h is r-homogeneous of degree $m \geq 0$, then for all $p > 0$, h is r'-homogeneous of degree $m' \geq 0$ with $r' = (\frac{r_1}{p}, \dots, \frac{r_n}{p})$ and $m' = \frac{m}{p}$. For this reason, taking $0 < r_i < 1$ for all $i \in \{1, \dots, n\}$ is not a restriction in the definition of homogeneity. In what follows, we always take $0 < r_i < 1$ for all $i \in \{1, \dots, n\}$.

An r-homogeneous norm ρ is a continuous function $\rho : \mathbb{R}^n \to \mathbb{R}$ which is positive definite and r-homogeneous of degree 1 ($0 < r_i < 1$: for all $i \in \{1, \dots, n\}$).

In the present paper, we will use the r-homogeneous norm

$$\rho(x) = \sum_{i=1}^{n} |x_i|^{\frac{1}{r_i}}. \tag{3.4}$$

This homogeneous norm is continuously differentiable in $\mathbb{R}^n$ and for all $i \in \{1, \dots, n\}$,

$$\frac{\partial \rho}{\partial x_i}(\delta(s,x)) = s^{1-r_i}\frac{\partial \rho}{\partial x_i}(x)\,. \tag{3.5}$$

Lemma 3.1. *The time-invariant r-homogeneous system $\dot{x}(t) = f_H(x(t))$ of order $\tau > 0$ is asymptotically stable if and only if there exists a $k > 1$ such that for all $x_0 \in \mathbb{R}^n$ and for all $t \geq 0$*

$$\rho(x(t,0,x_0)) \leq \frac{k\rho(x_0)}{(1+t\rho(x_0)^\tau)^{\frac{1}{\tau}}}. \tag{3.6}$$

Proof. The proof is omitted. The reader is referred to [4, pp. 278–284] for a proof when the dilation is the standard dilation. □

3.1. Main result. Consider the r-homogeneous system

$$\dot{x}(t) = f_H(x(t),t) \tag{3.7}$$

with order $\tau > 0$. Consider for each $\sigma \in \mathbb{R}$ the frozen system

$$\dot{x}(t) = f_H(x(t),\sigma). \tag{3.8}$$

The solution of (3.7) at t with initial condition $x_0 \in \mathbb{R}$ at t_0 is denoted as $x_H(t,t_0,x_0)$, and the solution of (3.8) is denoted as $x_{H\sigma}(t,t_0,x_0)$.

Theorem 3.2. *Assume the following.*

- *The equilibrium point $x = 0$ of each frozen system* (3.8) *is asymptotically stable, and the estimate* (3.6) *is uniform, i.e., there exists a $k > 1$ independent of σ such that for all $\sigma \in \mathbb{R}$, for all $x_0 \in \mathbb{R}^n$, and for all $t \geq 0$*

$$\rho(x_{H\sigma}(t,0,x_0)) \leq \frac{k\rho(x_0)}{(1+t\rho(x_0)^\tau)^{\frac{1}{\tau}}}. \tag{3.9}$$

- *$f_H(x,\sigma)$ is continuously differentiable with respect to x and σ.*
- *There exists a $c_f > 0$ such that for all $\sigma \in \mathbb{R}$, for all $y \in \mathbb{R}^n$ with $\rho(y) = 1$, and for all $i,k \in \{1,\dots,n\}$*

$$|f_{Hi}(y,\sigma)| \leq c_f \quad \textit{and} \quad \left|\frac{\partial f_{Hi}}{\partial x_k}(y,\sigma)\right| \leq c_f \quad \textit{and} \quad \left|\frac{\partial f_{Hi}}{\partial \sigma}(y,\sigma)\right| \leq c_f; \tag{3.10}$$

then the time-varying system (3.7) *is uniformly bounded and uniformly ultimately bounded.*

Proof. The proof is based on Theorem 2.4. Define for all $x \in \mathbb{R}^n$ and for all $\sigma \in \mathbb{R}$

$$V(x,\sigma) := \int_0^\infty \rho(x_{H\sigma}(t,0,x))^{m\tau}dt, \tag{3.11}$$

where m will be chosen later on in the proof.

By (3.3), (3.10), and [4, pp. 278–284], there exists a $k' > 0$ such that for all $\sigma \in \mathbb{R}$, for all $x \in \mathbb{R}^n$, and for all $t \geq 0$

$$\rho(x_{H\sigma}(t,0,x)) \geq \frac{k'\rho(x)}{(1+t\rho(x)^\tau)^{\frac{1}{\tau}}}. \tag{3.12}$$

We now prove (2.7), (2.8), and (2.9).

I. By (3.9) and (3.12), there exist a $c_1 > 0$ and a $c_2 > 0$ such that for all $x \in \mathbb{R}^n$ and for all $\sigma \in \mathbb{R}$

$$c_1 \rho(x)^{(m-1)\tau} \le V(x,\sigma) \le c_2 \rho(x)^{(m-1)\tau}. \tag{3.13}$$

This implies that (2.7) is satisfied when $m > 1$.

II. In order to verify (2.8), we calculate $\frac{\partial V}{\partial \sigma}(x,\sigma)$. Notice that

(3.14)
$$\begin{aligned}\frac{\partial V}{\partial \sigma}(x,\sigma) &= m\tau \int_0^\infty \rho(x_{H\sigma}(t,0,x))^{m\tau-1} \frac{\partial}{\partial \sigma}\left(\rho(x_{H\sigma}(t,0,x))\right) dt \\ &= m\tau \int_0^\infty \rho(x_{H\sigma}(t,0,x))^{m\tau-1} \left(\sum_{i=1}^n \frac{\partial \rho}{\partial x_i}(x_{H\sigma}(t,0,x)) \frac{\partial x_{H\sigma i}}{\partial \sigma}(t,0,x)\right) dt.\end{aligned}$$

Here, we assume that $m > \frac{1}{\tau}$. For all $i \in \{1, \dots, n\}$,

$$\begin{aligned}&\frac{\partial \rho}{\partial x_i}(x_{H\sigma}(t,0,x)) \\ &= \frac{\partial \rho}{\partial x_i}\left(\delta(\rho(x_{H\sigma}(t,0,x))^{-1}, x_{H\sigma}(t,0,x))\right) \rho(x_{H\sigma}(t,0,x))^{1-r_i}.\end{aligned} \tag{3.15}$$

The continuity of $\frac{\partial \rho}{\partial x_i}$ on the compact set $\{y : y \in \mathbb{R}^n, \rho(y) = 1\}$ implies the existence of a $c_\rho > 0$ such that $|\frac{\partial \rho}{\partial x_i}(y)| \le c_\rho$ for all $i \in \{1, \dots, n\}$ and for all $y \in \mathbb{R}^n$ with $\rho(y) = 1$. It is clear that[4]

(3.16)
$$\left|\frac{\partial V}{\partial \sigma}(x,\sigma)\right| \le m\tau \int_0^\infty \rho(x_{H\sigma}(t,0,x))^{m\tau-1} c_\rho \left(\sum_{i=1}^n \rho(x_{H\sigma}(t,0,x))^{1-r_i} \left| \frac{\partial x_{H\sigma i}}{\partial \sigma}(t,0,x)\right| \right) dt.$$

In order to obtain an upper bound for the right-hand side of (3.16), we first calculate an appropriate upper bound for

$$\sum_{i=1}^n \rho(x_{H\sigma}(t,0,x))^{1-r_i} \left| \frac{\partial x_{H\sigma i}}{\partial \sigma}(t,0,x)\right|. \tag{3.17}$$

By integrating (3.8), one obtains that for all $x \in \mathbb{R}^n$ and for all $t \ge 0$,

$$x_{H\sigma}(t,0,x) = x + \int_0^t f_H(x_{H\sigma}(s,0,x),\sigma) ds \tag{3.18}$$

and

$$\begin{aligned}\frac{\partial x_{H\sigma i}}{\partial \sigma}(t,0,x) = \int_0^t &\sum_{k=1}^n \frac{\partial f_{Hi}}{\partial x_k}(x_{H\sigma}(s,0,x),\sigma) \frac{\partial x_{H\sigma k}}{\partial \sigma}(s,0,x) \\ &+ \frac{\partial f_{Hi}}{\partial \sigma}(x_{H\sigma}(s,0,x),\sigma) ds.\end{aligned} \tag{3.19}$$

[4] Since $f_H(x,\sigma)$ is continuously differentiable with respect to x and σ, the solution $x_{H\sigma}(t,t_0,x_0)$ is continuously differentiable with respect to σ [5, Theorem 3.3, p. 21].

By multiplying (3.19) with $\rho(x_{H\sigma}(t,0,x))^{1-r_i}$ and invoking the triangle inequality, one obtains that the expression (3.17) is less than or equal to

$$\begin{aligned}(3.20)\qquad &\sum_{i=1}^{n}\rho(x_{H\sigma}(t,0,x))^{1-r_i}\int_0^t\sum_{k=1}^{n}\left|\frac{\partial f_{Hi}}{\partial x_k}(x_{H\sigma}(s,0,x),\sigma)\right|\ \left|\frac{\partial x_{H\sigma k}}{\partial\sigma}(s,0,x)\right|\\ &+\left|\frac{\partial f_{Hi}}{\partial\sigma}(x_{H\sigma}(s,0,x),\sigma)\right|ds.\end{aligned}$$

Notice that

$$\begin{aligned}(3.21)\qquad &\frac{\partial f_{Hi}}{\partial x_k}(x_{H\sigma}(s,0,x),\sigma)\\ &=\ \rho(x_{H\sigma}(s,0,x))^{\tau+r_i-r_k}\frac{\partial f_{Hi}}{\partial x_k}\left(\delta(\rho(x_{H\sigma}(s,0,x))^{-1},x_{H\sigma}(s,0,x)),\sigma\right)\end{aligned}$$

and

$$\begin{aligned}(3.22)\qquad &\frac{\partial f_{Hi}}{\partial\sigma}(x_{H\sigma}(s,0,x),\sigma)\\ &=\ \rho(x_{H\sigma}(s,0,x))^{\tau+r_i}\frac{\partial f_{Hi}}{\partial\sigma}\left(\delta(\rho(x_{H\sigma}(s,0,x))^{-1},x_{H\sigma}(s,0,x)),\sigma\right).\end{aligned}$$

By (3.10) and (3.20), the expression (3.17) is less than or equal to

$$\begin{aligned}(3.23)\qquad &c_f\sum_{i=1}^{n}\rho(x_{H\sigma}(t,0,x))^{1-r_i}\int_0^t\sum_{k=1}^{n}\rho(x_{H\sigma}(s,0,x))^{\tau+r_i-r_k}\left|\frac{\partial x_{H\sigma k}}{\partial\sigma}(s,0,x)\right|\\ &+\rho(x_{H\sigma}(s,0,x))^{\tau+r_i}ds.\end{aligned}$$

By (3.9), $\rho(x_{H\sigma}(\tau,0,x_0))\le k\rho(x_0)$ for all $x_0\in\mathbb{R}^n$ and for all $\tau\ge0$. By setting $x_0=x_{H\sigma}(s,0,x)$ and $\tau=t-s$, one obtains that for all $s\in[0,t]$: $\rho(x_{H\sigma}(t-s,0,x_{H\sigma}(s,0,x)))\le k\rho(x_{H\sigma}(s,0,x))$ such that $\rho(x_{H\sigma}(t,0,x))\le k\rho(x_{H\sigma}(s,0,x))$. This implies by (3.23) that the expression (3.17) is less than or equal to

$$\begin{aligned}&(3.24)\\ &c_f\sum_{i=1}^{n}k^{1-r_i}\int_0^t\rho(x_{H\sigma}(s,0,x))^{1-r_i}\left(\sum_{k=1}^{n}\rho(x_{H\sigma}(s,0,x))^{\tau+r_i-r_k}\left|\frac{\partial x_{H\sigma k}}{\partial\sigma}(s,0,x)\right|\right.\\ &\left.+\ \rho(x_{H\sigma}(s,0,x))^{\tau+r_i}\right)ds.\end{aligned}$$

By (3.9), there exists a $c_3>0$ such that for all $x\in\mathbb{R}^n$ and for all $t\ge0$,

$$(3.25)\qquad \int_0^t\rho(x_{H\sigma}(s,0,x))^{1+\tau}ds\le c_3\rho(x).$$

This implies the existence of a $c_4>0$ and a $c_5>0$ such that the expression (3.17) is less than or equal to

$$\begin{aligned}&(3.26)\\ &c_4\rho(x)+c_5\int_0^t\rho(x_{H\sigma}(s,0,x))^{\tau}\left(\sum_{k=1}^{n}\rho(x_{H\sigma}(s,0,x))^{1-r_k}\left|\frac{\partial x_{H\sigma k}}{\partial\sigma}(s,0,x)\right|\right)ds.\end{aligned}$$

By the Gronwall–Bellman lemma, it is clear that

$$\sum_{i=1}^{n} \rho(x_{H\sigma}(t,0,x))^{1-r_i} \left| \frac{\partial x_{H\sigma i}}{\partial \sigma}(t,0,x) \right| \le c_4 \rho(x) e^{c_5 \int_0^t \rho(x_{H\sigma}(s,0,x))^{\tau} ds}. \tag{3.27}$$

By (3.9), there exists a $c_6 > 0$ such that for all $x \in \mathbb{R}^n$ and for all $t \ge 0$,

$$c_5 \int_0^t \rho(x_{H\sigma}(s,0,x))^{\tau} ds \le c_6 \ln\left(1 + t\rho(x)^{\tau}\right), \tag{3.28}$$

and therefore

$$\sum_{i=1}^{n} \rho(x_{H\sigma}(t,0,x))^{1-r_i} \left| \frac{\partial x_{H\sigma i}}{\partial \sigma}(t,0,x) \right| \le c_4 \rho(x) \left(1 + t\rho(x)^{\tau}\right)^{c_6}. \tag{3.29}$$

This implies by (3.16) and (3.9) the existence of a $c_7 > 0$ such that

$$\left| \frac{\partial V}{\partial \sigma}(x,\sigma) \right| \le c_7 \rho(x)^{m\tau} \int_0^{\infty} \left(1 + t\rho(x)^{\tau}\right)^{c_6 + \frac{1}{\tau} - m} dt. \tag{3.30}$$

Take $m > c_6 + \frac{1}{\tau} + 1$. There exists a $c_8 > 0$ such that for all $\sigma \in \mathbb{R}$ and for all $x \in \mathbb{R}^n$,

$$\left| \frac{\partial V}{\partial \sigma}(x,\sigma) \right| \le c_8 \rho(x)^{(m-1)\tau}. \tag{3.31}$$

III. From the definition (3.11), one obtains that the derivative of $V(x,\sigma)$ along the solutions of (3.8) equals $\dot{V}(x,\sigma) = -\rho(x)^{m\tau}$. This implies that

$$\frac{\partial V}{\partial x}(x,\sigma) f_H(x,\sigma) = \dot{V}(x,\sigma) = -\rho(x)^{m\tau}. \tag{3.32}$$

IV. By (3.31) and (3.32), (2.8) and (2.9) are satisfied with $W_3(x) = c_8 \rho(x)^{(m-1)\tau}$ and $W_4(x) = \rho(x)^{m\tau}$. Since $W_3(x) - W_4(x)$ is a continuous function of x that tends to $-\infty$ as $\|x\|$ tends to $+\infty$, there exist an $R_V > 0$ and a positive definite $W_5 : \mathbb{R}^n \to \mathbb{R}$ such that for all x with $\|x\| > R_V$, $W_3(x) - W_4(x) \le -W_5(x)$. Theorem 2.4 implies uniform boundedness and uniform ultimate boundedness of the system (3.7). □

Remark 2. By taking $V(x,\sigma)$ as defined by (3.11) and by setting $V(x,t) = V(x,\sigma)|_{\sigma=t}$ as in the proof of Theorem 2.4, we not only prove uniform boundedness and uniform ultimate boundedness of (3.7). We also obtain the estimate

$$\rho(x_H(t,t_0,x_0)) \le \frac{k\rho(x_0)}{\left(1 + (t-t_0)\rho(x_0)^{\tau}\right)^{\frac{1}{\tau}}} \tag{3.33}$$

when $\|x_H(\tau,t_0,x_0)\|$ is sufficiently large for all $\tau \in [t_0,t]$. Indeed, by (3.31) and (3.32), the derivative of $V(x,t)$ along the trajectories of (3.7) satisfies the inequality

$$\dot{V}(x_H(t,t_0,x_0),t) \le c_8 \rho(x_H(t,t_0,x_0))^{(m-1)\tau} - \rho(x_H(t,t_0,x_0))^{m\tau}. \tag{3.34}$$

There exists an $R_{V2} > 0$ sufficiently large such that for all $x_H(t,t_0,x_0)$ with $\|x_H(t,t_0,x_0)\| > R_{V2}$, $\dot{V}(x_H(t,t_0,x_0),t) \le -0.5\rho(x_H(t,t_0,x_0))^{m\tau}$. By (3.13), there exist positive constants c_{10} and c_{11} such that

$$\dot{V}(x_H(t,t_0,x_0),t) \le -c_{10} V(x_H(t,t_0,x_0),t)^{\frac{m}{m-1}} \tag{3.35}$$

and by integration

$$V(x_H(t,t_0,x_0),t)^{\frac{1}{1-m}} \ge V(x_0,t_0)^{\frac{1}{1-m}} + c_{11}(t-t_0). \tag{3.36}$$

By (3.13), there exists a $k > 1$ such that for all t_0 and for all $t \ge t_0$ (3.33) is satisfied when $\|x_H(\tau,t_0,x_0)\| > R_{V2}$ for all $\tau \in [t_0,t]$.

3.2. Time-periodicity. For the time-periodic case, the conditions mentioned in Theorem 3.2 can be simplified. When the r-homogeneous system $\dot{x}(t) = f_H(x(t), t)$ with order $\tau > 0$ is time-periodic, it is possible to reformulate the first condition of Theorem 3.2 by simply requiring asymptotic stability for each frozen system $\dot{x}(t) = f_H(x(t), \sigma)$.

THEOREM 3.3. *Consider the system $\dot{x}(t) = f_H(x(t), t)$, where $f_H(x, t)$ is assumed to be time-periodic with period T_f. When $f_H(x, \sigma)$ is continuously differentiable with respect to x and σ and each frozen system $\dot{x}(t) = f_H(x(t), \sigma)$ is assumed to be asymptotically stable, i.e., for each $\sigma \in \mathbb{R}$ there exists a $k(\sigma) > 1$ such that for all $x_0 \in \mathbb{R}^n$ and for all $t \geq 0$*

$$\rho(x_{H\sigma}(t, 0, x_0)) \leq \frac{k(\sigma)\rho(x_0)}{(1 + t\rho(x_0)^\tau)^{\frac{1}{\tau}}}, \tag{3.37}$$

then the time-varying system $\dot{x}(t) = f_H(x(t), t)$ is uniformly bounded and uniformly ultimately bounded.

Proof. The proof is based on Theorem 3.2.

I. First we show that (3.9) is satisfied.

Since $f_H(x, \sigma)$ is continuously differentiable with respect to x and σ, $x_{H\sigma}(T, 0, x)$ is continuously differentiable with respect to σ for all $T > 0$, for all $x \in \mathbb{R}^n$, and for all $\sigma \in [0, T_f]$ [5, Theorem 3.3, p. 21]. This implies that for all $T > 0$, for all $x \in \mathbb{R}^n$, and for all $\sigma \in [0, T_f]$

$$\begin{aligned} &\lim\nolimits_{\sigma' \to \sigma} x_{H\sigma'}(T, 0, x) = x_{H\sigma}(T, 0, x) \quad \text{and} \\ &\lim\nolimits_{\sigma' \to \sigma} \rho(x_{H\sigma'}(T, 0, x)) = \rho(x_{H\sigma}(T, 0, x)). \end{aligned} \tag{3.38}$$

Take an arbitrary $\sigma \in [0, T_f]$. $k(\sigma)$ in (3.37) is not unique, but for each fixed σ, the set of the possible $k(\sigma)$ has an infimum $k_{\inf}(\sigma)$. We will now prove that $k_{\inf}(\sigma)$ is a continuous function of σ. The continuity of $k_{\inf}(\sigma)$ as a function of σ on a compact interval implies the existence of a bounded maximum $\bar{k}_{\inf}$ on $[0, T_f]$. By taking an arbitrary $\epsilon > 0$ and setting $k = \bar{k}_{\inf} + \epsilon$, (3.37) implies that (3.9) is satisfied.

Suppose $k_{\inf}(\sigma)$ as a function of σ has a discontinuity at σ', i.e., there exists an $\epsilon' > 0$ such that for each $\delta' > 0$ there is a $\sigma'' \in]\sigma' - \delta', \sigma' + \delta'[$ for which $|k_{\inf}(\sigma') - k_{\inf}(\sigma'')| > \epsilon'$. This means that $k_{\inf}(\sigma'') < k_{\inf}(\sigma') - \epsilon'$ or that $k_{\inf}(\sigma'') > k_{\inf}(\sigma') + \epsilon'$.

First, suppose that $k_{\inf}(\sigma'') < k_{\inf}(\sigma') - \epsilon'$. For this fixed σ', $k_{\inf}(\sigma')$ is the infimum of all possible $k(\sigma')$. By (3.37), there exist a $T' > 0$ and an $x' \in \mathbb{R}^n$ such that

$$\rho(x_{H\sigma'}(T', 0, x')) > \frac{(k_{\inf}(\sigma') - \frac{\epsilon'}{4})\rho(x')}{(1 + T'\rho(x')^\tau)^{\frac{1}{\tau}}}. \tag{3.39}$$

But by (3.37)

$$\rho(x_{H\sigma''}(T', 0, x')) \leq \frac{(k_{\inf}(\sigma'') + \frac{\epsilon'}{4})\rho(x')}{(1 + T'\rho(x')^\tau)^{\frac{1}{\tau}}} \leq \frac{(k_{\inf}(\sigma') - \frac{3\epsilon'}{4})\rho(x')}{(1 + T'\rho(x')^\tau)^{\frac{1}{\tau}}} \tag{3.40}$$

such that

$$\rho(x_{H\sigma'}(T', 0, x')) - \rho(x_{H\sigma''}(T', 0, x')) > \frac{\epsilon'}{2} \frac{\rho(x')}{(1 + T'\rho(x')^\tau)^{\frac{1}{\tau}}} \tag{3.41}$$

for all $\delta' > 0$ with $\sigma'' \in]\sigma' - \delta', \sigma' + \delta'[$. Since (3.41) contradicts with (3.38), the assumption that $k_{\inf}(\sigma'') < k_{\inf}(\sigma') - \epsilon'$ is false.

Suppose that $k_{\inf}(\sigma'') > k_{\inf}(\sigma') + \epsilon'$. For this fixed σ'', $k_{\inf}(\sigma'')$ is the infimum of all possible $k(\sigma'')$. There exist a $T'' > 0$ and an $x'' \in \mathbb{R}^n$ such that

$$\rho(x_{H\sigma''}(T'', 0, x'')) > \frac{(k_{\inf}(\sigma'') - \frac{\epsilon'}{4})\rho(x'')}{(1 + T''\rho(x'')^{\tau})^{\frac{1}{\tau}}}. \tag{3.42}$$

But by (3.37),

$$\rho(x_{H\sigma'}(T'', 0, x'')) \le \frac{(k_{\inf}(\sigma') + \frac{\epsilon'}{4})\rho(x'')}{(1 + T''\rho(x'')^{\tau})^{\frac{1}{\tau}}} \le \frac{(k_{\inf}(\sigma'') - \frac{3\epsilon'}{4})\rho(x'')}{(1 + T''\rho(x'')^{\tau})^{\frac{1}{\tau}}} \tag{3.43}$$

such that

$$\rho(x_{H\sigma''}(T'', 0, x'')) - \rho(x_{H\sigma'}(T'', 0, x'')) > \frac{\epsilon'}{2} \frac{\rho(x'')}{(1 + T''\rho(x'')^{\tau})^{\frac{1}{\tau}}} \tag{3.44}$$

for all $\delta' > 0$ with $\sigma'' \in]\sigma' - \delta', \sigma' + \delta'[$. Since (3.44) contradicts with (3.38), the assumption that $k_{\inf}(\sigma'') > k_{\inf}(\sigma') + \epsilon'$ is false.

Since the discontinuity assumptions lead to contradictions, $k_{\inf}(\sigma)$ is a continuous function of σ. Therefore, $k_{\inf}(\sigma)$ has a bounded maximum $\bar{k}_{\inf}$ on $[0, T_f]$. By taking an arbitrary $\epsilon > 0$ and setting $k = \bar{k}_{\inf} + \epsilon$, (3.37) implies that (3.9) is satisfied.

II. Since $f_H(x, \sigma)$ is periodic in the second variable and since $f_H(x, \sigma)$ is continuously differentiable with respect to x and σ, (3.10) is satisfied.

III. Theorem 3.2 implies uniform boundedness and uniform ultimate boundedness of the homogeneous system $\dot{x}(t) = f_H(x(t), t)$ with order $\tau > 0$. □

4. Homogeneous approximations far from the origin. In the present section, we generalize the results of section 3. We consider systems that have a dominant homogeneous approximation at infinity, i.e., systems represented as $\dot{x}(t) = f_H(x(t), t) + g(x(t), t)$. Here $f_H : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is homogeneous with a positive order τ and $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is a perturbation of f_H when $\|x\|$ is sufficiently large.

Consider $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$. There exist an $R_g > 0$ and a continuous nonincreasing function $F : \mathbb{R}^+ \to \mathbb{R}$ with $\lim_{s \to \infty} F(s) = 0$ such that for all $x \in \mathbb{R}^n$ with $\rho(x) > R_g$ and for all $t \in \mathbb{R}$

$$\|\delta(\rho(x)^{-1}, g(x, t))\| \le \rho(x)^{\tau} F(\rho(x)). \tag{4.1}$$

A typical example is the case where $g(x, t)$ is the sum of a finite number of homogeneous terms with the same dilation as $f_H(x, t)$ and with orders *smaller* than τ. For x with $\|x\|$ sufficiently large, $g(x, t)$ can be seen as a perturbation which does not affect the uniform boundedness and the uniform ultimately boundedness property.

Consider the system

$$\dot{x}(t) = f_H(x(t), t) + g(x(t), t) \tag{4.2}$$

and the frozen systems

$$\dot{x}(t) = f_H(x(t), \sigma) + g(x(t), \sigma). \tag{4.3}$$

The solution of $\dot{x}(t) = f_H(x(t), t)$ at t with initial condition $x_0 \in \mathbb{R}$ at t_0 is denoted as $x_H(t, t_0, x_0)$, the solution of $\dot{x}(t) = f_H(x(t), \sigma)$ is denoted as $x_{H\sigma}(t, t_0, x_0)$, the solution of (4.2) is denoted as $x(t, t_0, x_0)$, and the solution of (4.3) is denoted as $x_\sigma(t, t_0, x_0)$.

THEOREM 4.1. *Assume that all the conditions of Theorem* 3.2 *are satisfied; then the time-varying system* $\dot{x}(t) = f_H(x(t), t) + g(x(t), t)$ *is uniformly bounded and uniformly ultimately bounded.*

Proof. In order to prove the boundedness properties, consider the Liapunov function $V(x, \sigma)$, defined by (3.11), which satisfies (3.13), (3.31), and (3.32).

I. By making calculations similar to the calculations in part II in the proof of Theorem 3.2 leading to (3.31), one obtains $\frac{\partial V}{\partial x_j}(x, \sigma)$ for all $j \in \{1, \ldots, n\}$ and for all $x \in \mathbb{R}^n$. There exists a $c_9 > 0$ such that for all $j \in \{1, \ldots, n\}$ and for all $x \in \mathbb{R}^n$

$$\left|\frac{\partial V}{\partial x_j}(x, \sigma)\right| \leq c_9 \rho(x)^{(m-1)\tau - r_j} \tag{4.4}$$

when $m > c_6 + \frac{1}{\tau} + 1$.

II. By (4.1), for all $x \in \mathbb{R}^n$ with $\rho(x) > R_g$, for all $t \in \mathbb{R}$, and for all $j \in \{1, \ldots, n\}$

$$|g_j(x, t)| \leq \rho(x)^{r_j + \tau} F(\rho(x)). \tag{4.5}$$

From (3.32), (4.4), and (4.5),

$$\sum_{j=1}^{n} \frac{\partial V}{\partial x_j}(x, \sigma)(f_{Hj}(x, \sigma) + g_j(x, \sigma)) \leq -\rho(x)^{m\tau}(1 - nc_9 F(\rho(x))) \tag{4.6}$$

when $\rho(x) > R_g$. Since $\lim_{s \to \infty} F(s) = 0$, there exists a $\rho_F > R_g$ such that for all $x \in \mathbb{R}^n$ with $\rho(x) > \rho_F$, $F(\rho(x)) < \frac{1}{2nc_9}$. This implies that for all $x \in \mathbb{R}^n$ with $\rho(x) > \rho_F$

$$\frac{\partial V}{\partial x}(x, \sigma)(f_H(x, \sigma) + g(x, \sigma)) < -\frac{1}{2}\rho(x)^{m\tau}. \tag{4.7}$$

III. By (3.31) and (4.7), (2.8) and (2.9) are satisfied with $W_3(x) = c_8 \rho(x)^{(m-1)\tau}$ and $W_4(x) = \frac{1}{2}\rho(x)^{m\tau}$. Since $W_3(x) - W_4(x)$ is a continuous function of x that tends to $-\infty$ as $\|x\|$ tends to $+\infty$, there exist an $R_V > 0$ and a positive definite $W_5 : \mathbb{R}^n \to \mathbb{R}$ such that for all $x \in \mathbb{R}^n$ with $\|x\| > R_V$, $W_3(x) - W_4(x) \leq -W_5(x)$. Theorem 2.4 implies uniform boundedness and uniform ultimate boundedness of the system (4.2). □

The boundedness results of Theorem 4.1 do not require time-periodicity of the system $\dot{x}(t) = f_H(x(t), t)$. In Theorem 4.2, we consider the time-periodic case which allows a simplification of the conditions.

THEOREM 4.2. *Consider the system* $\dot{x}(t) = f_H(x(t), t)$ *with order* $\tau > 0$. *Here,* $f_H(x, t)$ *is assumed to be time-periodic with period* T_f. *When* $f_H(x, \sigma)$ *is continuously differentiable with respect to* x *and* σ *and each frozen system* $\dot{x}(t) = f_H(x(t), \sigma)$ *is assumed to be asymptotically stable, then the time-varying system* $\dot{x}(t) = f_H(x(t), t) + g(x(t), t)$ *is uniformly bounded and uniformly ultimately bounded.*

Proof. All the conditions imposed by Theorem 3.3 are satisfied. The proof of Theorem 3.3 implies that the conditions imposed by Theorem 4.1 (and, equivalently, by Theorem 3.2) are satisfied. This implies uniform boundedness and uniform ultimate boundedness of the time-varying system $\dot{x}(t) = f_H(x(t), t) + g(x(t), t)$. □

The condition that each frozen system $\dot{x}(t) = f_H(x(t), \sigma)$ is asymptotically stable in order to ensure boundedness properties cannot, in general, be relaxed to a stability condition for some frozen systems. This is illustrated by means of a scalar example. Consider the case where $f_H(x,t) = s(t)x^5$ with $s(t) = 0$ when $t \in [0,1)$ and $s(t) = -1$ when $t \in [1,2)$ with $s(t) = s(t+2)$ for all $t \in \mathbb{R}$. When $g(x,t) = x^3$, the system $\dot{x}(t) = s(t)x^5 + x^3$ is not bounded. (The system has a finite escape time since $s(t) = 0$ when $t \in [0,1)$.)

If we specialize the result of Theorem 4.2 to the case where $f_H(x,t) = f_H(x)$ is time-invariant, Theorem 4.2 shows that asymptotic stability of $\dot{x}(t) = f_H(x(t))$ implies uniform boundedness and uniform ultimate boundedness of

$$\dot{x}(t) = f_H(x(t)) + g(x(t),t). \tag{4.8}$$

The boundedness properties are determined by $f_H(x)$ and not by $g(x,t)$. By (4.1), for x with $\|x\|$ sufficiently large, $g(x,t)$ is a perturbation which does not affect the uniform boundedness and the uniform ultimate boundedness property.

On the other hand, for $\|x\|$ sufficiently small, the local asymptotic stability properties of (4.8) are determined by $g(x,t)$ and not by $f_H(x,t)$. By (4.1), $f_H(x,t)$ does not affect the local stability properties. For example, when $f_H(x) = -x^3$ and $g(x,t) = x$, the system $\dot{x}(t) = -x^3(t) + x(t)$ is uniformly bounded and uniformly ultimately bounded, but the origin is not stable.

There is a duality between these boundedness results and the results proved by Hermes [6, Theorem 3.3], the results proved by Morin and Samson ([10, Proposition 2], and the linearization technique [8, pp. 127–132 and pp. 147–148]. Hermes [6] proves that asymptotic stability of $\dot{x}(t) = f_H(x(t))$ implies local asymptotic stability of $\dot{x}(t) = f_H(x(t)) + g(x(t))$ in the case when $g(x)$ is the sum of a finite number of homogeneous terms with the same dilation as $f_H(x)$ and with orders *larger* than τ. This result is valid since for x with $\|x\|$ sufficiently small $g(x)$ can be seen as a perturbation which does not affect the local asymptotic stability property.

5. Example: Lotka–Volterra equations. Theorem 4.2 proves uniform boundedness and uniform ultimate boundedness for systems arising as $\dot{x}(t) = f_H(x(t),t) + g(x(t),t)$ when all the frozen systems $\dot{x}(t) = f_H(x(t),\sigma)$ of order $\tau > 0$ have an asymptotically stable equilibrium point $x = 0$. The verification of this asymptotic stability property is crucial in the application of Theorem 4.2. It is obvious that the verification of this asymptotic stability property becomes much easier when the frozen systems $\dot{x}(t) = f_H(x(t),\sigma)$ belong to a class of systems whose stability properties have been studied in the literature. We illustrate this by means of an example.

Consider the time-varying Lotka–Volterra system

$$\dot{x}_i(t) = x_i(t)\left((A(t)x)_i + r_i(t)\right), \tag{5.1}$$

where $x = (x_1, \dots, x_n)^T$. Here, $A : \mathbb{R} \to \mathbb{R}^{n\times n}$ is periodic with period T_A and for all $i \in \{1, \dots, n\}$, $r_i : \mathbb{R} \to \mathbb{R}$.

The time-varying Lotka–Volterra equation (5.1) is a positive system. A system is positive if its state-components are nonnegative, i.e., the first closed orthant of $\mathbb{R}^n$ is positively invariant. Examples of these systems are found in a variety of applied areas such as biology, chemistry, and sociology [9, 7].

Although the results in the previous sections are formulated for systems defined in $\mathbb{R}^n$, they also allow the study of positive systems defined in the first closed orthant of $\mathbb{R}^n$.

Indeed, if $\dot{x}(t) = f_H(x(t),t)$ is defined in the first orthant of $\mathbb{R}^n$ with the additional condition that this first closed orthant is positively invariant for the original time-varying system and for all the time-invariant frozen systems $\dot{x}(t) = f_H(x(t), \sigma)$, the results of Theorems 3.2 and 3.3 remain valid.

Suppose that $\dot{x}(t) = f_H(x(t), t) + g(x(t), t)$ is defined in the first orthant of $\mathbb{R}^n$ with the additional condition that this first closed orthant is positively invariant for the time-varying systems $\dot{x}(t) = f_H(x(t), t) + g(x(t), t)$ and $\dot{x}(t) = f_H(x(t), t)$. Suppose also that the first closed orthant of $\mathbb{R}^n$ is positively invariant for all the time-invariant frozen systems $\dot{x}(t) = f_H(x(t), \sigma) + g(x(t), \sigma)$ and $\dot{x}(t) = f_H(x(t), \sigma)$; then the results of Theorems 4.1 and 4.2 remain valid.

Example 1. Assume the following.

- Whenever

  $$x_i \left(A(\sigma)x\right)_i = \lambda(\sigma)x_i, \qquad i = 1, \ldots, n, \tag{5.2}$$

  holds for some σ and for some $x \neq 0$ with $x_i \geq 0$ for all $i \in \{1, \ldots, n\}$, then $\lambda(\sigma) < 0$.
- $A(\sigma)$ is continuously differentiable. There exists a $c_A > 0$ such that for all $\sigma \in \mathbb{R}$

  $$\|A(\sigma)\| \leq c_A \quad \text{and} \quad \|\dot{A}(\sigma)\| \leq c_A. \tag{5.3}$$

- There exists a $c_r > 0$ such that for all $\sigma \in \mathbb{R}$ and for all $i \in \{1, \ldots, n\}$

  $$|r_i(\sigma)| \leq c_r; \tag{5.4}$$

then the time-varying system (5.1) is uniformly bounded and uniformly ultimately bounded.

Proof. By [7, pp. 185–187], all the systems

$$\dot{x}_i(t) = x_i(t)(A(\sigma)x(t))_i \tag{5.5}$$

are asymptotically stable. Take an arbitrary $r \in]0, 1[$. All the systems (5.5) are homogeneous with respect to the dilation $(r, \ldots, r)$ with order $\tau = r > 0$. Take $f_H(x, t) = (f_{H1}(x, t), \ldots, f_{Hn}(x, t))^T$ with $f_{Hi}(x, t) = x_i(A(t)x)_i$, and take $g_i(x, t) = r_i(t)x_i$ for all x and for all t. By setting $F(s) = \frac{\sqrt{n}c_r}{s^r}$, (4.1) is satisfied. By the asymptotic stability property of (5.5), by setting $T_f = T_A$, and by (5.3), the conditions required by Theorems 3.3 and 4.2 are satisfied. By Theorem 4.2, we obtain uniform boundedness and uniform ultimate boundedness for the time-varying positive system (5.1). □

6. Conclusions. In the present paper, we have reduced the study of uniform boundedness and uniform ultimate boundedness of a time-varying system to the study of the time-invariant frozen systems.

Appendix A. The appendix contains the proof of Proposition 2.3.

Proof. Take an arbitray $R_1 > 0$. Define $R_2(R_1) := \max\{\alpha^{-1}(\beta(R_V)), \alpha^{-1}(\beta(R_1))\}$. Take an arbitrary $x_0 \in \mathbb{R}^n$ with $\|x_0\| \leq R_1$. In order to prove (2.2), suppose that for some $t_1 > t_0$, $\|x(t_1, t_0, x_0)\| > R_2(R_1)$. Because of continuity of solutions and since $R_2(R_1) \geq \max\{R_1, R_V\}$, there exists a $t_1' \in [t_0, t_1[$ such that $\|x(t_1', t_0, x_0)\| = \max\{R_1, R_V\}$ and $\|x(t, t_0, x_0)\| > \max\{R_1, R_V\}$ for all $t \in]t_1', t_1]$. Since

$$V(x(t_1, t_0, x_0)) = V(x(t_1', t_0, x_0)) + \int_{t_1'}^{t_1} \dot{V}(x(t, t_0, x_0), t)dt \tag{A.1}$$

and by (2.5) $\dot{V}(x(t,t_0,x_0)) \leq -\gamma(\|x(t,t_0,x_0)\|)$ for all $t \in]t_1', t_1]$, it is clear that

$$V(x(t_1,t_0,x_0)) \leq V(x(t_1',t_0,x_0)) \leq \beta(\max\{R_1, R_V\}). \tag{A.2}$$

By (A.2) and (2.4), $\|x(t_1,t_0,x_0)\| \leq \max\{\alpha^{-1}(\beta(R_V)), \alpha^{-1}(\beta(R_1))\} = R_2(R_1)$. This contradicts the assumption that $\|x(t_1,t_0,x_0)\| > R_2(R_1)$. Therefore, $\|x(t,t_0,x_0)\| \leq R_2(R_1)$ for all $t \geq t_0$ and (2.2) follows.

In order to prove (2.3), take $R = \alpha^{-1}(\beta(R_V))$. Take an arbitrary $R_1 > 0$. Take an arbitrary t_0 and $x_0 \in \mathbb{R}^n$ with $\|x_0\| \leq R_1$. The solution $x(t,t_0,x_0)$ exists for all $t \geq t_0$ since by the first part of the proof $\|x(t,t_0,x_0)\| \leq \max\{\alpha^{-1}(\beta(R_V)), \alpha^{-1}(\beta(R_1))\}$. Define

$$T(R_1) = \max\left\{0, \frac{\beta(R_1) - \alpha(\beta^{-1}(\alpha(R)))}{\gamma(R_V)}\right\}. \tag{A.3}$$

Assume that for all $t_1 \in [t_0, t_0 + T(R_1)]$, $\|x(t_1,t_0,x_0)\| > \beta^{-1}(\alpha(R)) = R_V$ such that $\|x(t,t_0,x_0)\| > \beta^{-1}(\alpha(R)) = R_V$ for all $t \in [t_0, t_1]$ and by (2.5), $\dot{V}(x(t,t_0,x_0),t) \leq -\gamma(R_V)$ for all $t \in [t_0,t_1]$. Since for all $t_1 \in [t_0, t_0 + T(R_1)]$,

$$\begin{aligned} V(x(t_1,t_0,x_0),t_1) &= V(x_0,t_0) + \int_{t_0}^{t_1} \dot{V}(x(t,t_0,x_0),t)dt \leq V(x_0,t_0) \\ &- (t_1 - t_0)\gamma(R_V), \end{aligned} \tag{A.4}$$

we also have

$$\begin{aligned} &V(x(t_0 + T(R_1), t_0, x_0), t_0 + T(R_1)) \\ &\leq \beta(R_1) - T(R_1)\gamma(R_V) \leq \alpha(\beta^{-1}(\alpha(R))). \end{aligned} \tag{A.5}$$

This implies by (2.4) that $\|x(t_0 + T(R_1), t_0, x_0)\| \leq \beta^{-1}(\alpha(R))$, which contradicts the assumption that $\|x(t_1,t_0,x_0)\| > \beta^{-1}(\alpha(R))$ for all $t_1 \in [t_0, t_0 + T(R_1)]$. Consequently, there exists a $t_1 \in [t_0, t_0 + T(R_1)]$ such that $\|x(t_1)\| \leq \beta^{-1}(\alpha(R))$. By the first part of the proof, $\|x(t)\| \leq R$ when $t \geq t_1$ and (2.3) follows. □

REFERENCES

[1] D. Aeyels and J. Peuteman, *A new asymptotic stability criterion for nonlinear time-variant differential equations*, IEEE Trans. Automat. Control, 43 (1998), pp. 968–971.

[2] D. Aeyels and J. Peuteman, *On exponential stability of nonlinear time-varying differential equations*, Automatica J. IFAC, 35 (1999), pp. 1091–1100.

[3] C. A. Desoer, *Slowly varying $\dot{x} = A(t)x$*, IEEE Trans. Automat. Control, 14 (1969), pp. 780–781.

[4] W. Hahn, *Stability of Motion*, Springer-Verlag, New York, 1967.

[5] J. K. Hale, *Ordinary Differential Equations*, Robert E. Krieger Publishing Company, Malabar, FL, 1980.

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

[7] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[8] H. K. Khalil, *Nonlinear Systems*, Prentice Hall, Englewood Cliffs, NJ, 1996.

[9] P. De Leenheer and D. Aeyels, *A note on uniform boundedness of a class of positive systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, Arizona, 1999, pp. 2575–2579.

[10] P. Morin and C. Samson, *Time-varying exponential stabilization of a rigid spacecraft with two control torques*, IEEE Trans. Automat. Control, 42 (1997), pp. 528–535.

[11] J. Peuteman and D. Aeyels, *Averaging results and the study of uniform asymptotic stability of homogeneous differential equations that are not fast time-varying*, SIAM J. Control Optim., 37 (1999), pp. 997–1010.

[12] J. Peuteman and D. Aeyels, *A note on exponential stability of partially slowly time-varying nonlinear systems*, in Preprints of the 4th IFAC Nonlinear Control Systems Design Symposium, Enschede, The Netherlands, 1998, pp. 471–475.

[13] H. H. Rosenbrock, *The stability of linear time-dependent control systems*, J. Electronics Control (1), 15 (1963), pp. 73–80.

[14] W. J. Rugh, *Linear System Theory*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[15] V. Solo, *On the stability of slowly time-varying linear systems*, Math. Control Signals Systems, 7 (1994), pp. 331–350.

[16] T. Yoshizawa, *Stability Theory by Liapunov's Second Method*, Math. Soc. Jap., Tokyo, 1966.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1423–1440

A FREQUENCY DOMAIN ROBUST STABILITY THEOREM FOR INFINITE DIMENSIONAL SYSTEMS WITH PARAMETRIC UNCERTAINTY*

COLEMAN BROSILOW† AND MARSHALL J. LEITMAN‡

Abstract. A robust stability theorem of non-Kharitonov type for parametric uncertainty is formulated and proved: *A closed loop system depending on a connected set of parameters whose maximum frequency response is bounded is either stable for all the parameters or for none.* Realistic illustrative examples are provided.

Key words. robust stability, process control, H-infinity, closed loop frequency response, Mp-tuning, parametric uncertainty, root locus

AMS subject classifications. 32A20, 93D09

PII. S0363012999357550

1. Introduction.

In this investigation we formulate and prove a frequency domain robust stability theorem for infinite dimensional systems with parametric uncertainty which is not of Kharitonov [7] type. That is, we are not restricted to polynomial functions or to interval parameter uncertainty. We formulate our result in terms of sufficiently regular meromorphic functions and rather arbitrary parameter sets. Although we state and prove our main result for scalar-valued transfer functions, it can be extended to the matrix-valued case without too much modification. Ours is a parametric approach and differs from the gap metric approach introduced by Zames and El-Sakkary [13]. In the context of frequency response measurements, Vinnicombe [10] has proved some sharp gap metric robust stability results.

The importance of our result is that it justifies a relatively simple computational procedure for the design and tuning of control systems for stable or unstable, finite or infinite dimensional processes with uncertain parameters. The tuning procedure, which we call Mp-tuning, adjusts controller parameters to achieve a specified maximum peak of the frequency response of a closed loop transfer function over all processes in the uncertainty set. If the maximum magnitude of the closed loop frequency response is finite over all uncertain processes, our robust stability theorem then ensures that the resulting control system is stable for all processes in the uncertainty set provided it is stable for any one process in the uncertainty set; conversely, if the control system is unstable for any one process in the uncertainty set, then the control system is unstable for all processes in the uncertainty set.

Examples are provided to illustrate the use of our theorem for stable and unstable processes, infinite dimensional systems, and to emphasize the key role played by the requirement that the transfer function of the closed loop control system be a suitable meromorphic function, a condition which is usually easy to assess in practice.

*Received by the editors June 4, 1999; accepted for publication (in revised form) August 8, 2000; published electronically January 19, 2001.
http://www.siam.org/journals/sicon/39-5/35755.html

†Department of Chemical Engineering, Case Western Reserve University, Cleveland, OH 44106-7217 (cbb@po.cwru.edu).

‡Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106-7058 (mxl5@po.cwru.edu).

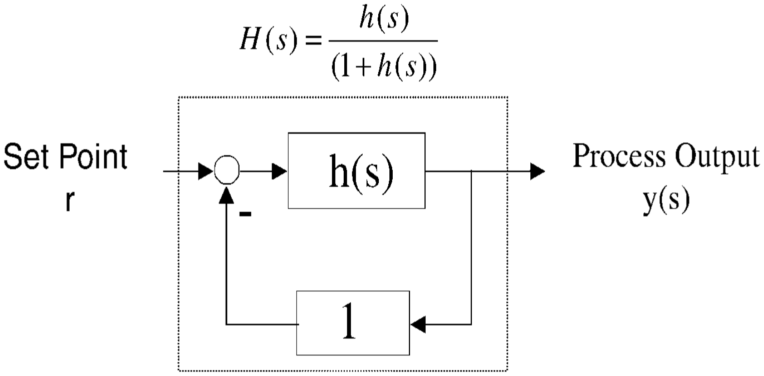


FIG. 1. *A simple feedback system.*

2. Definitions, notation, and hypotheses.

2.1. Transfer functions. An *(open loop) transfer function* is a complex-valued function $(s,p) \mapsto h(s,p)$ of a complex variable s and a parameter p.[1] To the open loop transfer function h we associate the *(closed loop) transfer function* H given by

$$h \mapsto H = \frac{h}{1+h}. \tag{1}$$

The feedback system is shown in Figure 1.

A *parameter* p is an n-dimensional real-valued vector lying in a subset Π of $\mathbb{R}^n$, Euclidean n-space, called the *parameter set* or the *uncertainty set.* If the complex variable s lies on the imaginary axis, we write $s = i\omega$ and call the real number ω the *frequency.* By the *closed loop frequency response (for the parameter* p) we mean the complex-valued function of a real variable given by

$$\omega \mapsto H(i\omega,p) = \frac{h(i\omega,p)}{1+h(i\omega,p)}. \tag{2}$$

To a distinct pair of open loop transfer functions, h and $\widetilde{h}$, we associate a transfer function $\widetilde{H}$ of the form

$$\left(h,\widetilde{h}\right) \mapsto \widetilde{H} = \frac{\widetilde{h}}{1+h}. \tag{3}$$

The corresponding closed loop frequency response is $\omega \longmapsto \widetilde{H}(i\omega,p)$.

2.2. Complex notation and meromorphic functions. The symbol s always denotes a complex number. Complex numbers are represented as points $s = \alpha + i\omega$ in the complex plane, $\mathbb{C}$, or as points on the complex projective sphere (Riemann sphere), $\mathbb{P}$. The complex points at infinity, $s = \infty$, and at zero, $s = 0$, are identified with the north and south poles of the projective sphere, $\mathbb{P}$. If $s \neq \infty$ is in $\mathbb{P}$, we still write $s = \alpha + i\omega$ for some real pair (α,ω). The open complex half-plane $\mathbb{C}^+ = \{s : \Re(s) > 0\} \subset \mathbb{C}$ corresponds to an open hemisphere of $\mathbb{P}$, denoted by $\mathbb{P}^+$. $\overline{\mathbb{C}^+}$ denotes the closed

[1] In this context, for each p, the function $s \mapsto h(s,p)$ may be construed as the Laplace transform of some real-valued function depending on the parameter p.

complex half-plane $\overline{\mathbb{C}^+} = \{s : \Re(s) \geq 0\} \subset \mathbb{C}$. $\overline{\mathbb{P}^+}$ denotes the closure, in $\mathbb{P}$, of the open hemisphere $\mathbb{P}^+$. The real number line, $(-\infty, +\infty)$, is denoted by $\mathbb{R}$, and the extended real number line, $[-\infty, +\infty]$, is denoted by $\overline{\mathbb{R}}$. The topologies of $\mathbb{P}$ and $\mathbb{C}$ are the same provided we identify the neighborhood of infinity $\{s : |s| > M\} \subset \mathbb{C}$ with the north polar cap of $\mathbb{P}$.

A map $(s,p) \to f(s,p)$ on a region in $\mathbb{C}\times\mathbb{C}^n$ is a *holomorphic function* whenever $(s,(p_1,p_2,\ldots,p_n)) \to f(s,(p_1,p_2,\ldots,p_n))$ is holomorphic separately in each of the variables $s, p_1, p_2, \ldots, p_n$. A map $(s,p) \to f(s,p)$ on a region in $\mathbb{C}\times\mathbb{C}^n$ is a *meromorphic function* if, on every neighborhood, $\mathcal{U}_i$, in the region, it is the ratio of two relatively prime holomorphic functions, say, n_i and m_i; that is, $m_i(s,p) f(s,p) = n_i(s,p)$ on $\mathcal{U}_i$. The m's and the n's must agree on overlapping neighborhoods in the sense that $n_i m_j = n_j m_i$ on $\mathcal{U}_i \cap \mathcal{U}_j$. Regarding the *value,* $f(s_0,p_0)$, of a meromorphic function f at a point (s_0,p_0) we have the following important definition.

The values of a meromorphic function.

$$f(s_0,p_0) = \begin{cases} 0 & \text{if} \quad n_i(s_0,p_0) = 0, m_i(s_0,p_0) \neq 0 \quad \text{(zero)}, \\ \frac{n_i(s_0,p_0)}{m_i(s_0,p_0)} & \text{if} \quad n_i(s_0,p_0) \neq 0, m_i(s_0,p_0) \neq 0, \\ \infty & \text{if} \quad n_i(s_0,p_0) \neq 0, m_i(s_0,p_0) = 0 \quad \text{(pole)}, \\ \text{undefined} & \text{if} \quad n_i(s_0,p_0) = 0, m_i(s_0,p_0) = 0. \end{cases} \tag{4}$$

In the first three cases f has a definite value in $\mathbb{P}$ at (s_0,p_0), denoted by $f(s_0,p_0)$; in the last case, the value of f must remain undefined since there is no consistent way to assign a value at such a point.[2] Indeed, for any complex number c, there is a point (s,p), arbitrarily close to (s_0,p_0), at which $f(s,p) = c$.

2.3. Hypotheses on the parameter set and the closed loop transfer function. Regarding the parameter set Π, we always assume the following:

(P) *The parameter set* Π *is an open connected subset of* $\mathbb{R}^n$.

Since Π is *connected,* for any two points $\widehat{p}$ and $\widehat{q}$ in Π, there is a *path* σ *in* Π *from* $\widehat{p}$ *to* $\widehat{q}$; that is, there is a continuous function $\sigma : [0,1] \to \Pi$ such that $\sigma(0) = \widehat{p}$ and $\sigma(1) = \widehat{q}$.

In the following hypotheses, h is a function defined on $\overline{\mathbb{P}^+} \times \Pi$ with values in $\mathbb{P}$; that is, h has a value at every point in $\overline{\mathbb{C}^+} \times \Pi$ and at the points $\{\infty\} \times \Pi$.

(M) *There is a connected open set* $\widehat{\Pi}$ *in* $\mathbb{C}^n$ *such that* $\widehat{\Pi} \cap \mathbb{R}^n = \Pi$ *and the function* $(s,p) \mapsto h(s,p)$ *has a meromorphic extension to* $\mathbb{C}^+ \times \widehat{\Pi}$*, also denoted by* h. (In particular, h has a value at every point in $\mathbb{C}^+ \times \Pi$ in the sense of the definition of the values of a meromorphic function (4).[3])

(S) *For each parameter* $p \in \Pi$*, the function* $s \mapsto h(s,p)$ *is not constant on* $\mathbb{C}^+$.

(R) *For each parameter* $p \in \Pi$*, the function* $s \mapsto h(s,p)$ *is real-valued or* ∞ *on the positive real axis* $\{s : \Re(s) > 0, \Im(s) = 0\}$.

(C) *The function* $(s,p) \mapsto h(s,p)$ *is (jointly) continuous as a function from* $\overline{\mathbb{P}^+} \times \Pi$ *into* $\mathbb{P}$.[4]

[2] See, for example, Hörmander [6, Theorem 6.2.3].

[3] If $(s,p) \mapsto h(s,p)$ is "meromorphic for real parameters p" in the sense of convergent power series expansions, this extension is always possible.

[4] If h, regarded as a map from $\overline{\mathbb{P}^+} \times \widehat{\Pi}$ into $\mathbb{P}$, satisfies (M), then $(s,p) \mapsto h(s,p)$ is always (jointly) continuous at points in $\mathbb{P}^+ \times \Pi$ (since h is meromorphic and has a value at every point in $\mathbb{P}^+ \times \widehat{\Pi}$). Since h has a value at every point in $\overline{\mathbb{P}^+} \times \Pi$, the condition (C) amounts to the additional assumption that h be (jointly) continuous at boundary points of the form $(i\omega, p) \in \overline{\mathbb{P}^+} \times \Pi$ for $\omega \in \mathbb{R}$ or $(\infty, p) \in \overline{\mathbb{P}^+} \times \Pi$.

The hypotheses (M), (S), and (R) are essentially regularity or consistency conditions on h. However, the connectivity of the parameter set, hypothesis (P), and the continuity of h, hypothesis (C), especially at points on the imaginary axis, are structurally necessary for our results to hold. (See Example 4.1.)

If h satisfies any of the hypotheses (M), (S), (R), or (C), so does the function $1 + h$. Furthermore, if h satisfies the hypothesis (M), so does the function H defined by (1) since we can assign H values in a manner consistent with the definition of the values of a meromorphic function (4). If, in addition to (M), h satisfies hypotheses (R) and/or (C), so does H.

We will be concerned with either a single open loop transfer function h and its associated closed loop transfer function H or a distinct pair of open loop transfer functions h and $\widetilde{h}$ together with the associated closed loop transfer function $\widetilde{H}$ defined by (3). In general, if h and $\widetilde{h}$ each satisfy hypotheses (M) and/or (C), it does not follow that $\widetilde{H}$ satisfies (M) and/or (C), for it may not be possible to assign values to $\widetilde{H}$ consistent with definition of the values of a meromorphic function (4) at every point where h and $\widetilde{h}$ have values. In such a case we must add the following compatibility hypothesis.

(D) *The open loop transfer functions h and $\widetilde{h}$ are* compatible *in the sense that each satisfies hypotheses (M) and (C) and the function $\widetilde{H} = \frac{\widetilde{h}}{1+h}$ also satisfies hypotheses (M) and (C).* Clearly, h is always compatible with itself.

For a given open loop transfer function h, the *maximum magnitude of the closed loop frequency response over all parameters* or, succinctly, the *maximum magnitude*, we mean the extended-real-valued function $\mathcal{H} : [-\infty, +\infty] \to [0, +\infty]$ defined by

$$\omega \mapsto \mathcal{H}(\omega) = \sup_{p\in\Pi} |H(i\omega, p)| = \sup_{p\in\Pi} \left| \frac{h(i\omega, p)}{1 + h(i\omega, p)} \right|. \tag{5}$$

Whenever hypothesis (C) holds, $\mathcal{H}$ is well defined on the extended real line $\overline{\mathbb{R}}$. By an *extended*-real-valued function, we mean that the supremum in the definition of $\mathcal{H}$ can be infinite at a given frequency, say, $\widetilde{\omega}$, in which case we write $\mathcal{H}(\widetilde{\omega}) = +\infty$; in particular $\mathcal{H}(\widetilde{\omega}) = +\infty$ whenever $h(i\widetilde{\omega}, \widetilde{p}) + 1 = 0$ for some frequency $\widetilde{\omega}$ and some parameter $\widetilde{p}$. If hypotheses (M), (C), and (R) hold, $h(\overline{s}, p) = \overline{h(s, p)}$ on $\overline{\mathbb{P}^+} \times \Pi$, since $h(s, p)$ is real-valued or ∞ on the positive real axis $\{s : \Re(s) > 0, \Im(s) = 0\}$. In particular, $h(-i\omega, p) = \overline{h(i\omega, p)}$ for all $\omega \in \mathbb{R}$ and each $p \in \Pi$.[5] In this case, the maximum magnitude is an (extended-real-valued) *even* function of the frequency; hence, it is completely determined by its values for $\omega \in [0, \infty]$. Definition (5) and the assertions above remain meaningful and valid mutatis mutandis when we have a distinct pair of compatible transfer functions h and $\widetilde{h}$, in which case we write $\widetilde{\mathcal{H}}$ instead of $\mathcal{H}$.

For a single open loop transfer function h, and a parameter p_0, the corresponding closed loop system is said to be *stable at the parameter* p_0 whenever the function $s \mapsto H(s, p_0) = \frac{h(s,p_0)}{1+h(s,p_0)}$ has no pole ($\neq \infty$) in the *closed* hemisphere $\overline{\mathbb{P}^+}$. This condition is equivalent to the assertion that the function $s \mapsto 1 + h(s, p_0)$ has no zero in the *closed* hemisphere $\overline{\mathbb{P}^+}$. The closed loop system is said to be *unstable at the parameter* p_0 if it is not stable at p_0; that is, $s \mapsto H(s, p_0) = \frac{h(s,p_0)}{1+h(s,p_0)}$ has at least one pole in the *closed* hemisphere $\overline{\mathbb{P}^+}$.

[5] This is a consequence of the reflection principle. Note that $h(\infty, p)$ is real-valued or ∞ for every parameter p.

For a pair of distinct compatible open loop transfer functions h and $\widetilde{h}$, we also say that the corresponding closed loop system is *stable at the parameter* p_0 whenever the transfer function $s \mapsto \widetilde{H}(s,p_0) = \frac{\widetilde{h}(s,p_0)}{1+h(s,p_0)}$ has no pole ($\neq \infty$) in the *closed* hemisphere $\overline{\mathbb{P}^+}$. The closed loop system is again said to be *unstable at the parameter* p_0 if it is not stable at p_0; that is, $s \mapsto \widetilde{H}(s,p_0) = \frac{\widetilde{h}(s,p_0)}{1+h(s,p_0)}$ has at least one pole in the *closed* hemisphere $\overline{\mathbb{P}^+}$. In the case of distinct compatible open loop transfer functions, stability at parameter p_0 is not equivalent to the assertion that the function $s \mapsto 1 + h(s,p_0)$ has no zero in the *closed* hemisphere $\overline{\mathbb{P}^+}$. Indeed, $1 + h(s,p_0)$ could be nonzero at a point where $\widetilde{h}(s,p_0)$ has a pole.

3. Main results and proofs. We now state our main result—*the robust stability theorem (RST)*—for a single open loop transfer function.

Theorem 3.1 (RST: single transfer function). *Assume that the open loop transfer function satisfies hypotheses* (P), (M), (S), (R), *and* (C). *Let the maximum magnitude of the closed loop frequency response over all parameters be bounded. Then, if the closed loop system is stable for* at least one *parameter in the parameter set* Π, *the closed loop system is stable for* all *parameters in the parameter set* Π.

Proof (outline of the proof). Suppose Theorem 3.1 was false under the stated hypotheses. Then there is a parameter q_0 for which the system is stable and another parameter p_0 for which the system is unstable. The parameter set Π is assumed open and connected, so we can choose a path from p_0 to q_0 lying entirely inside the parameter set Π. Since the system is unstable for p_0, there is at least one point, say, s_0, in $\overline{\mathbb{P}^+}$ such that $1+h(s_0,p_0)=0$; and, since the system is stable for q_0, there is no such point corresponding to q_0. We then use a "root locus" or "continuity" argument to conclude that there is a parameter $\widehat{p} \in \Pi$ and a frequency $\widehat{\omega} \in [0,\infty]$ such that $1+h(i\widehat{\omega},\widehat{p})=0$, and, hence, $\mathcal{H}(\widehat{\omega}) = +\infty$. The maximum magnitude is not bounded, which contradicts our hypothesis. □

In order to develop the root locus argument and complete the proof of Theorem 3.1, we establish some preliminary definitions and results. It will be convenient to write g for the function $h+1$. The hypotheses (M), (R), (S), and (C) hold for g if and only if they hold for h. If, for some parameter p_0, $s \longmapsto g(s,p_0)$ has no zero in $\overline{\mathbb{P}^+}$, we say that g is *stable* at p_0. We say g is *unstable* at p_0 if it is not stable at p_0.

We will first be concerned with the zeros of g in the set $\mathbb{P}^+ \times \Pi$; that is, those points $(s_0,p_0) \in \mathbb{P}^+ \times \Pi$ such that $g(s_0,p_0)=0$. Fix such a zero, (s_0,p_0). Henceforth, suppose (M) and (S) hold.[6] Then there is a positive integer k_0 and neighborhood in $\mathbb{P}^+ \times \widehat{\Pi}$ of (s_0,p_0) on which g is holomorphic and satisfies

$$g(s,p_0) = (s-s_0)^{k_0} f(s,p_0) \tag{6}$$

for some function $s \mapsto f(s,p_0)$, holomorphic near s_0, such that $f(s_0,p_0) \neq 0$. We call k_0 the *multiplicity* of the zero (s_0,p_0). We need the following immediate consequence of the Weierstrass preparation theorem (WPT).[7]

Theorem 3.2 (WPT). *Let g satisfy* (M) *and* (S). *Suppose* $g(s_0,p_0)=0$ *at* $(s_0,p_0) \in \mathbb{P}^+ \times \Pi$. *Then there is a neighborhood U_0 in $\mathbb{P}^+ \times \widehat{\Pi}$ of the zero (s_0,p_0) and k_0 functions $p \mapsto \lambda_k(p)$, $\lambda_k(p_0)=0$, $k=0,1,2,3,\ldots,(k_0-1)$, holomorphic on some common neighborhood W_0 in $\widehat{\Pi}$ of p_0, such that the function* $(s,p) \mapsto g(s,p)$

[6] Hypothesis (S) guarantees that the function $s \mapsto g(s,p_0)$ is not identically zero near s_0.

[7] For a complete statement and proof of the Weierstrass preparation theorem, see Hörmander [6].

and the function

$$(s,p)\mapsto (s-s_0)^{k_0}+\sum_{j=0}^{k_0-1}\lambda_j(p)(s-s_0)^j \tag{7}$$

have the same set of zeros on U_0. *This function is called the* Weierstrass polynomial *associated with* g *at the point* (s_0,p_0).

Let p_0 and q_0 be distinct parameters in Π, and let σ be a path in Π from p_0 to q_0. Let (s_0,p_0) be a zero of g such that $s_0\in\mathbb{P}^+$. For each $t\in[0,1]$ define a set $L_0^\sigma(t)\subset\mathbb{C}^+$ by

$$L_0^\sigma(t)=\left\{\widehat{s}\in\mathbb{P}^+ : \exists\varsigma:[0,t]\overset{cont}{\longrightarrow}\mathbb{P}^+,\left\{\begin{array}{l}\varsigma(0)=s_0\\ \varsigma(t)=\widehat{s}\\ \forall\tau\in[0,t],g(\varsigma(\tau),\sigma(\tau))=0\end{array}\right\}\right\}. \tag{8}$$

In other words, $L_0^\sigma(t)$ is the set of those zeros, $\widehat{s}$, of $s\to g(s,\sigma(t))$ in $\mathbb{P}^+$ for which there is a continuous path $\tau\mapsto(\varsigma(\tau),\sigma(\tau))$ on $[0,t]$ into $\mathbb{P}^+\times\Pi$, consisting entirely of zeros of g, connecting the zero (s_0,p_0) to the zero $(\widehat{s},\sigma(t))$. We call $L_0^\sigma(t)$ the *root locus, at* t, *emanating from* s_0 *induced by* σ or, simply, *the root locus emanating from* s_0.

The following three properties of the root locus are implicit in its definition:

(L1) $L_0^\sigma(0)=\{s_0\}$,

(L2) $s\in L_0^\sigma(t)\Rightarrow g(s,\sigma(t))=0$,[8]

(L3) $L_0^\sigma(t)\neq\emptyset\Rightarrow\forall\tau\in[0,t),L_0^\sigma(\tau)\neq\emptyset$.

Thus far, there is no guarantee that $L_0^\sigma(\tau)\neq\emptyset$ for *any* $\tau>0$. The next property of the root locus asserts that there are such τ's.

(L4) *There is an* $\epsilon\in(0,1]$ *such that* $\forall\tau\in[0,\epsilon),L_0^\sigma(\tau)\neq\emptyset$.

This is an immediate consequence of the following lemma, which asserts that the local behavior of the root locus is entirely determined by the Weierstrass polynomial.

LEMMA 3.1 (local root locus lemma). *Let* g *satisfy* (P), (M), *and* (S). *Fix* $\widehat{\tau}\in[0,1]$. *Let* $(\widehat{s},\sigma(\widehat{\tau}))\in\mathbb{P}^+\times\Pi$ *be a zero of* g; *that is,* $g(\widehat{s},\sigma(\widehat{\tau}))=0$. *Suppose this zero has multiplicity* $\widehat{k}$. *Then, for some* $\epsilon>0$, *there are* $\widehat{k}$ *continuous functions* $\tau\mapsto\zeta_l(\tau),l=1,2,3,\ldots,\widehat{k}$, *defined on a common subinterval of* $[0,1]$ *of the form* $[\widehat{\tau},\widehat{\tau}+\epsilon)$ *or* $(\widehat{\tau}-\epsilon,\widehat{\tau}]$, *depending on whether* $\widehat{\tau}\in[0,1)$ *or* $\widehat{\tau}\in(0,1]$, *with values in* $\mathbb{P}^+$, *each of which satisfies* $g(\zeta_l(\tau),\sigma(\tau))=0$ *for all* τ *in the common interval of definition. We say that* $\tau\mapsto(\zeta_l(\tau),\sigma(\tau))$ *is a forward or backward path of zeros of* g *induced by* σ *emanating from* $(\widehat{s},\sigma(\widehat{\tau}))$. *Of course, if* $\hat{\tau}\in(0,1)$, *there are* $\hat{k}$ *forward and* $\hat{k}$ *backward paths of zeros induced by* σ *emanating from* $(\widehat{s},\sigma(\widehat{\tau}))$ *and, hence, as many as* $\hat{k}^2$ *continuous paths of zeros of* g *on* $(\hat{\tau}-\epsilon,\hat{\tau}+\epsilon)$ *containing* $(\widehat{s},\sigma(\widehat{\tau}))$.

Proof. The proof is given for the case $\widehat{\tau}=0$, which yields property (L4). The other cases are similarly proved. Let U_0 and W_0 be the neighborhoods described in the WPT associated with $(\widehat{s},\sigma(\widehat{\tau}))=(s_0,p_0)$. Choose $\epsilon\in(0,1)$ so that $\sigma(\tau)\in W_0$ whenever $\tau\in[0,\epsilon)$. This is clearly possible. For each $\tau\in[0,\epsilon)$, consider the roots of the Weierstrass polynomial associated with g at the given zero:

$$(s-s_0)^{\hat{k}}+\sum_{k=0}^{\hat{k}-1}\lambda_k(\sigma(\tau))(s-s_0)^k. \tag{9}$$

[8] In general, the converse assertion $g(s,\sigma(t))=0\Rightarrow s\in L(t)$ is false.

By the WPT the roots of this polynomial are precisely those of $s \mapsto g(s, \sigma(\tau))$, provided that $(s, \sigma(\tau)) \in U_0$. (Note that the $\hat{k}$ coefficient functions $\tau \mapsto \lambda_k(\sigma(\tau))$, $k = 0, 1, 2, 3, \ldots, (\hat{k} - 1)$, are each continuous and satisfy $\lambda_k(\sigma(0)) = \lambda_k(p_0) = 0$.) It follows that there are $\hat{k}$ continuous complex-valued functions $\tau \mapsto \zeta_l(\tau)$, $l = 1, 2, 3, \ldots, \hat{k}$, on $[0, \epsilon)$ such that each $\zeta_l(0) = s_0$ and

$$(10) \qquad (\zeta_l(\tau) - s_0)^{\hat{k}} + \sum_{k=0}^{\hat{k}-1} \lambda_k(\sigma(\tau))(\zeta_l(\tau) - s_0)^k = 0.$$

The functions $\tau \mapsto (\zeta_l(\tau), \sigma(\tau))$ are paths of zeros of g induced by σ emanating from $(\widehat{s}, \sigma(\widehat{\tau})) = (s_0, p_0)$. Any one of the $\hat{k}$ functions $\tau \mapsto \zeta_l(\tau)$ can serve as a path, or a segment of a path, in the definition of L_0^σ. In particular, $L_0^\sigma(\tau) \neq \emptyset$ for $\tau \in [0, \epsilon)$. □

There is a *maximal* interval of the form $[0, \epsilon_0) \subseteq [0, 1)$ such that $\tau \in [0, \epsilon_0) \Rightarrow L_0^\sigma(\tau) \neq \emptyset$. Indeed, the maximal interval $[0, \epsilon_0)$ can be realized by

$$(11) \qquad [0, \epsilon_0) = \bigcup \{[0, \epsilon) \subseteq [0, 1) : \forall \tau \in [0, \epsilon), L_0^\sigma(\tau) \neq \emptyset\}.$$

Consider the *Ω-limit set,* $\Omega(\sigma)$, of the root locus $t \mapsto L_0^\sigma(t)$ on the maximal interval $[0, \epsilon_0)$. The set $\Omega(\sigma)$ is defined by[9]

$$(12) \qquad \Omega(\sigma) = \bigcap_{\tau \in [0, \epsilon_0)} cl\left(\bigcup_{t \in [\tau, \epsilon_0)} L_0^\sigma(t)\right).$$

By its construction, $\Omega(\sigma)$ is a nonempty, compact subset of $\overline{\mathbb{P}^+}$. We call $\Omega(\sigma)$ *the Ω-limit set (for g) induced by σ emanating from s_0.*

We can now state and prove the following lemma.

Lemma 3.2 (Ω-limit set). *Assume that g satisfies* (P), (M)*, and* (S). *Let p_0 and q_0 be parameters in Π such that g is stable at q_0 and unstable at p_0. For the unstable parameter p_0, let there be a point $s_0 \in \mathbb{P}^+$ such that $g(s_0, p_0) = 0$. Fix a path σ in Π from p_0 to q_0. Let $\Omega(\sigma)$ be the Ω-limit set for g induced by σ emanating from s_0. Then $\Omega(\epsilon_0, \sigma) \bigcap \mathbb{P}^+ = \emptyset$. Equivalently, the Ω-limit set, $\Omega(\sigma)$, regarded as a subset of $\overline{\mathbb{P}^+}$, is a nonempty, compact subset of the extended imaginary axis $\{s : \Re(s) = 0\} \cup \{\infty\}$.*[10]

Proof. Consider the root locus $t \mapsto L_0^\sigma(t)$ of g induced by σ emanating from s_0 and its Ω-limit set $\Omega(\epsilon_0, \sigma)$. Observe that $L_0^\sigma(1) = \emptyset$, since g is stable at q_0. Suppose $\Omega(\sigma) \bigcap \mathbb{P}^+ \neq \emptyset$. Let $\widehat{s} \in \Omega(\sigma) \bigcap \mathbb{P}^+$. Then there is an increasing sequence $t_n \uparrow \epsilon_0$ and a sequence $s_n \in \mathbb{P}^+$ such that $s_n \to \widehat{s}$ and $g(s_n, \sigma(t_n)) = 0$. Now g must be holomorphic in some neighborhood of $(\widehat{s}, \sigma(\epsilon_0))$ in $\mathbb{P}^+ \times \widehat{\Pi}$, since it is defined at $(\widehat{s}, \sigma(\epsilon_0))$; moreover, it vanishes at points arbitrarily close to $(\widehat{s}, \sigma(\epsilon_0))$. Since g is jointly continuous at $(\widehat{s}, \sigma(\epsilon_0))$, it follows that $g(\widehat{s}, \sigma(\epsilon_0)) = 0$. It turns out that $\widehat{s} \in L_0^\sigma(\epsilon_0)$, so $L_0^\sigma(\epsilon_0) \neq \emptyset$. To see this, apply the local root locus Lemma 3.1 at the point $(\widehat{s}, \sigma(\epsilon_0))$. For N sufficiently large there must be a subsequence s_{n_k}, $n_k \geq N$, which lies entirely on one of the paths of zeros of g induced by σ emanating *backward* from $(\widehat{s}, \sigma(\epsilon_0))$ as described in the local root locus Lemma 3.1. This follows since *all* the zeros of $s \mapsto g(s, \sigma(\epsilon_0))$ close to $\hat{s}$ are determined by the corresponding Weierstrass

[9] The function cl( ) denotes topological closure in the complex plane $\mathbb{C}$ or on the projective sphere $\mathbb{P}$.

[10] The extended imaginary axis, $\{s : \Re(s) = 0\} \bigcup \{\infty\}$, should be construed as a great circle on $\mathbb{P}$.

polynomial. For any corresponding t_{n_k}, we have $s_{n_k} \in L_0^\sigma(t_{n_k})$. This implies that $\widehat{s} \in L_0^\sigma(\epsilon_0)$. Since $L_0^\sigma(1) = \emptyset$, we must have $\epsilon_0 < 1$. But then the interval $[0, \epsilon_0)$ is not maximal. Indeed, since $\widehat{s} \in L_0^\sigma(\epsilon_0)$, we can apply the local root locus Lemma (3.1) *forward* at $(\widehat{s}, \sigma(\epsilon_0))$, as we did at $(s_0, \sigma(0))$. Hence, there must be an ϵ_1, $1 \geq \epsilon_1 > \epsilon_0$ such that $L_0^\sigma(t) \neq \emptyset$ for $t \in [\epsilon_0, \epsilon_1)$. The proof is complete. □

Finally, we can prove Theorem (3.1).

Proof (details of the proof). Suppose Theorem 3.1 to be false. If, for the unstable parameter p_0, there is a frequency $\widehat{\omega} \in [0, \infty]$ such that $1 + h(i\widehat{\omega}, p_0) = 0$, then $\mathcal{H}(\widehat{\omega}) = \infty$. This is an immediate contradiction, and the RST follows. Suppose that for the unstable parameter p_0 there is no such point; that is, $1 + h(i\omega, p_0) \neq 0$ for all $\omega \in [0, \infty]$. Then there is at least one point $s_0 \in \mathbb{P}^+$ such that $g(s_0, p_0) = 1 + h(s_0, p_0) = 0$. Fix a path σ in Π from p_0 to q_0. By the Ω-limit set Lemma 3.2, the Ω-limit set $\Omega(\sigma)$ for g induced by σ emanating from s_0 is a subset of the extended imaginary axis $\{s : \Re(s) = 0\} \bigcup \{\infty\}$. Let $\widehat{s} \in \Omega(\sigma)$. Either $\widehat{s} = i\widehat{\omega}$, for some $\widehat{\omega} \in \mathbb{R}$, or $\widehat{s} = \infty$. In either case, there is a sequence $t_n \uparrow \epsilon_0$ and a sequence of points $s_n \in \mathbb{P}^+$ such that $s_n \to \widehat{s}$ and $g(s_n, \sigma(t_n)) = 0$. By the joint continuity assumption (C), we must have $g(\widehat{s}, \widehat{p}) = 0$, where $\widehat{p} = \sigma(\epsilon_0)$. This means that $\mathcal{H}(\widehat{\omega}) = +\infty$ or $\mathcal{H}(\infty) = +\infty$; that is, the maximum magnitude of the closed loop frequency response over all parameters is unbounded, which contradicts our hypothesis. The theorem is proved. □

The RST also holds for a pair of distinct compatible open loop transfer functions. However, the proof requires a bit more care.

Theorem 3.3 (RST: distinct compatible transfer functions). *Assume that the distinct pair of open loop transfer functions satisfy hypotheses* (P), (M), (S), (R), *and* (C), *and are compatible. Let the maximum magnitude of the closed loop frequency response over all parameters be bounded. Then, if the closed loop system is stable for* at least one *parameter in the parameter set* Π, *the closed loop system is stable for* all *parameters in the parameter set* Π.

Proof. The details of the proof are essentially the same as those for Theorem 3.1 once the following observations are made.

If $s \longmapsto \widetilde{H}(s, r_0)$ has a pole at s_0, then $\widetilde{H}$ has a pole at (s_0, r_0). In this case, $\widetilde{H}$ can be represented in a neighborhood of (s_0, r_0) as a ratio of a pair of relatively prime holomorphic functions, say, $\widetilde{m}$ and $\widetilde{n}$ ($\widetilde{H} = \frac{\widetilde{n}}{\widetilde{m}}$), such that $\widetilde{n}(s_0, r_0) \neq 0$ and $\widetilde{m}(s_0, r_0) = 0$. The two lemmas (Lemmas 3.1 and 3.2) and the proof of the theorem remain essentially unchanged, except that we use the $\widetilde{m}$'s locally in place of g. In this case we refer to $\Omega(\sigma)$ as *the* Ω-limit set (for $\widetilde{H}$) induced by σ emanating from s_0.

Suppose Theorem 3.3 to be false. If, for the unstable parameter p_0, there is a frequency $\widehat{\omega} \in [0, \infty]$ such that $\widetilde{H}(i\widehat{\omega}, p_0) = \infty$, then $\widetilde{\mathcal{H}}(\widehat{\omega}) = \infty$. This is an immediate contradiction, and the robust stability Theorem 3.3 follows. Suppose that for the unstable parameter p_0 there is no such point; that is, $\widetilde{H}(i\omega, p_0) \neq \infty$ for all $\omega \in [0, \infty]$. Then there is at least one point $s_0 \in \mathbb{P}^+$ such that $\widetilde{H}(s_0, p_0) = \infty$. Fix a path σ in Π from p_0 to q_0. By the root locus Lemma 3.1, the Ω-limit set $\Omega(\sigma)$ for $\widetilde{H}$ induced by σ emanating from s_0 is a nonempty, compact subset of the extended imaginary axis $\{s : \Re(s) = 0\} \bigcup \{\infty\}$. Let $\widehat{s} \in \Omega(\sigma)$. Either $\widehat{s} = i\widehat{\omega}$, for some $\widehat{\omega} \in \mathbb{R}$, or $\widehat{s} = \infty$. In either case, there is a sequence $t_n \uparrow \epsilon_0$ and a sequence of points $s_n \in \mathbb{P}^+$ such that $s_n \to \widehat{s}$ and $\widetilde{H}(s_n, \sigma(t_n)) = \infty$. By the joint continuity assumption (C), we must have $\widetilde{H}(\widehat{s}, \widehat{p}) = \infty$, where $\widehat{p} = \sigma(\epsilon_0)$. This means that $\widetilde{\mathcal{H}}(\widehat{\omega}) = +\infty$ or $\widetilde{\mathcal{H}}(\infty) = +\infty$; that is, the maximum magnitude of the closed loop frequency response over all parameters is unbounded, which contradicts our hypothesis. The theorem is proved. □

4. Examples. We provide three useful examples. The first, Example 4.1, illustrates the key role played by hypotheses (S) and (C). The second two, Examples 4.2 and 4.3, deal with unstable uncertain processes and illustrate the use of our robust stability Theorem 3.3, for a pair of distinct transfer functions. To deal with such processes, we will use two degree of freedom control structures shown in Figures 2 and 3. Since the configuration of Figure 3 is internally unstable for unstable processes (Morari and Zafiriou [8]), we will use Figure 3 only for design purposes. Actual control system implementation must be by Figure 2, or by algorithms proposed by Berber and Brosilow [1, 2] and Cheng and Brosilow [4], which yield the same input-output transfer functions as in Figure 3 but which are internally stable.

Example 4.1 (the role of hypotheses (P) and (C)). This example[11] illustrates the role of hypotheses (P) and (C) and why they are necessary for our result to hold.

Suppose h is given by

$$h(s,a) = \frac{a(a-s)}{a(1-a)+s(1+a)}, \tag{13}$$

where the parameter set is the open interval $\Pi = \{a : -1 < a < 1\}$. In this case the closed loop transfer function is

$$\frac{h(s,a)}{1+h(s,a)} = a\frac{(a-s)}{a+s}, \tag{14}$$

and the maximum magnitude function, $\mathcal{H}$, is

$$\mathcal{H}(\omega) = \sup_{-1<a<1} \left| \frac{h(s,i\omega)}{1+h(s,i\omega)} \right| = 1. \tag{15}$$

Clearly $\mathcal{H}$ is bounded. On the other hand, for $a < 0$ the system is unstable, while for $a > 0$, the system is stable. This seems to violate the conclusion of our RST! To see that this compelling example is not a counterexample to Theorem 3.1, examine the hypotheses. The parameter set, Π, is just an interval, so it is connected. From the formula above, h is well defined as a meromorphic function for all parameters $a \in \Pi$ and $s \in \mathbb{C}^+$; moreover, h is real-valued for $a \in \Pi$ and $s \in \{s : \Im(s) = 0, \Re(s) > 0\}$. Hence, hypotheses (P), (M), and (R) are satisfied. However, hypothesis (S) requires that, for any $a \in \Pi$, the function $s \to h(s,a)$ not be constant in $\mathbb{C}^+$. However, at $a = 0$, $s \to h(s,0) = 0$ for $s \in \mathbb{C}^+$. More importantly, hypothesis (C) requires that h be well defined on the extended imaginary axis for all parameters $a \in \Pi$ in such a way that it is jointly continuous there. But h cannot be defined at $(s_0, a_0) = (0,0)$ in such a way that it is jointly continuous at that point. In fact, the defining expression (13) for h represents a meromorphic function for all complex s and a which has no value at $s = a = 0$ in the sense of the definition of the values of a meromorphic function (4). An obvious way to avoid violating hypotheses (S) and/or (C) is to exclude $a = 0$ from the parameter set. However, since Π is an interval, this would disconnect the parameter set, violating hypothesis (P).

[11] The authors are indebted to Professor Sebastian Engell, Chemietechnik, Universität Dortmund, for suggesting this example.

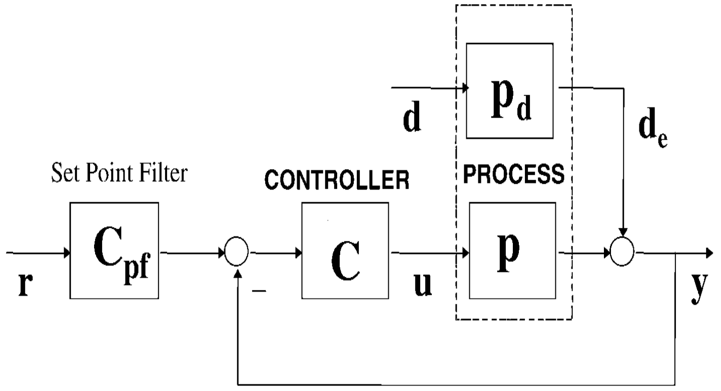


FIG. 2. *Standard feedback two degree of freedom control system.*

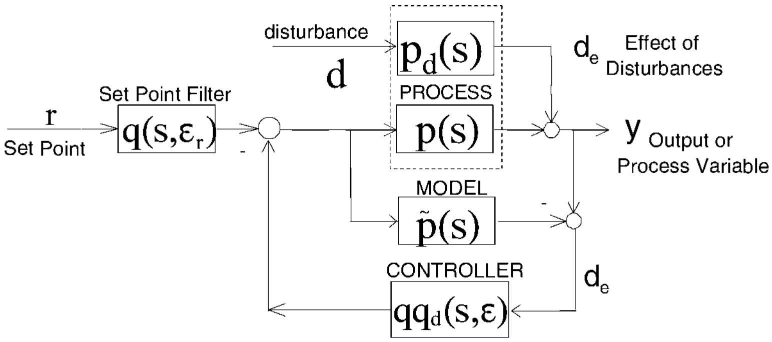


FIG. 3. *IMC form of two degree of freedom control system.*

Example 4.2 (an undamped two-mass spring system). This example presents the solution for an unstable "benchmark" control problem proposed by Wie and Bernstein [11, 12]. It consists of an undamped two-mass spring system modeled by

$$y(s) = p(s)u(s) + p_d(s)d(s), \tag{16}$$

where

$$p(s) = \frac{\frac{1}{2}}{s^2\left(\frac{1}{2k}s^2 + 1\right)}, \tag{17}$$

$$p_d(s) = \frac{\frac{1}{2}\left(\frac{1}{k}s^2 + 1\right)}{s^2\left(\frac{1}{2k}s^2 + 1\right)}, \tag{18}$$

and where

k =process spring constant ($0.5 \le k \le 2$),

u =control effort,

y =position of second mass.

The control objective is to design a robust controller to suppress impulse disturbances for all the plants in the uncertainty set. Further, the "nominal" process is to have a settling time of 15 seconds.

Referring to Figures 2 and 3, the closed transfer functions between the set points and the disturbances and the process are

$$y(s) = \frac{c_{pf}c(s)p(s)r(s) + p_d(s)d(s)}{(1 + c(s)p(s))} \tag{19}$$

for Figure 2 and

$$y(s) = \frac{q(s, \varepsilon_r)p(s)r(s) + (1 - \tilde{p}(s)qq_d(s, \varepsilon))\, p_d(s)d(s)}{(1 + (p(s) - \tilde{p}(s))\, qq_d(s, \varepsilon))} \tag{20}$$

for Figure 3.

The transfer functions given by (19) and (20) are equivalent (in terms of input-output transfer functions) if $c_{pf}(s)$ and $c(s)$ are chosen as

$$c_{pf}(s) = q(s, \varepsilon_r)\, (qq_d(s, \varepsilon))^{-1} \tag{21}$$

and

$$c(s) = \frac{qq_d(s, \varepsilon)}{(1 - \tilde{p}(s)qq_d(s, \varepsilon))}. \tag{22}$$

To design the controller $c(s)$ in (22) we select the controller, $qq_d(s, \varepsilon)$, so that the zeros of $(1 - \tilde{p}(s)qq_d(s, \varepsilon))$ cancel the poles of $\tilde{p}_d(s)$. An easy way of accomplishing this task is to take the controller $qq_d(s, \varepsilon)$ to be the product of two terms, $q(s, \varepsilon)$ and $q_d(s, \varepsilon)$:

$$qq_d(s, \varepsilon) = q(s, \varepsilon)q_d(s, \varepsilon). \tag{23}$$

The term $q(s, \varepsilon)$ is taken as the inverse of the model process as given below:

$$q(s, \varepsilon) = \frac{s^2 \left(\frac{1}{2\tilde{k}}s^2 + 1\right)}{\frac{1}{2}\,(\varepsilon s + 1)^4}, \tag{24}$$

where

$\tilde{k}$=model spring constant,

ε=adjustable controller filter time constant.

The above choice for $q(s, \varepsilon)$ reduces the problem to that of choosing $q_d(s, \varepsilon)$ so that the zeros of $(1 - q_d(s, \varepsilon))\,/\,(\varepsilon s + 1)^4$ cancel the poles of $\tilde{p}_d(s)$. This requires that $q_d(s, \varepsilon)$ be at least of third order as given by

$$q_d(s) = \frac{\tau_3 s^3 + \tau_2 s^2 + \tau_1 s + 1}{(\varepsilon s + 1)^3}. \tag{25}$$

The constants τ_1, τ_2, and τ_3 in (25) are chosen so that the numerator of $(1 - q_d(s, \varepsilon))\,/\,(\varepsilon s + 1)^4$ has zeros at $s = 0$ and at $s = \pm i\sqrt{2\tilde{k}}$. There is automatically an additional zero at $s = 0$ by virtue of the fact that $q_d(0, \varepsilon)$ is one. Equating the coefficients of the requisite polynomials yields

$$\tau_1 = 7\varepsilon, \tag{26}$$

$$\tau_2 = 21\varepsilon^2 - 70\tilde{k}\varepsilon^4 + 28\tilde{k}^2\varepsilon^6, \tag{27}$$

$$\tau_3 = 35\varepsilon^3 - 42\tilde{k}\varepsilon^5 + 4\tilde{k}^2\varepsilon^7. \tag{28}$$

These constants depend on the value of the filter time constant, ε, and the model parameter $\tilde{k}$. The filter time constant, ε, is selected to be $\varepsilon = 1$ so that the nominal plant with $\tilde{k} = 1$ has the desired settling time of about 15 seconds as required in the benchmark problem specifications (Wie and Bernstein [11, 12]). Having thus chosen ε, our only remaining task is to find an optimal value for $\tilde{k}$.

We will choose that value of $\tilde{k}$ that minimizes the maximum peak of $y(i\omega)/d(i\omega)$ from (20). By exploring values of the model spring constant, $\tilde{k}$, between 0.5 and 2, we determined that this minimum peak is obtained by using the model spring constant $\tilde{k} = 0.7$. The controller for this spring constant is given by

$$C_1(s) = \frac{10.8\left(s^3 - 1.88s^2 + 0.93s + 0.13\right)}{s^3 + 7s^2 + 19.6s + 25.2}. \tag{29}$$

Values for $\tilde{k}$ near 0.5 and 2 yield unstable control systems which are indicated by very large peaks of the maximum of $y(i\omega)/d(i\omega)$.

Braatz and Morari [3] solved the above control problem using the D-K iteration method (Doyle [5]) obtaining the following "μ-optimal" controller:

$$C_\mu(s) = \frac{0.443\,(9.402s + 1)\,(-2.697s + 1)\,(0.4789s + 1)}{(0.216s^2 + 0.861s + 1)\,(0.118s^2 + 0.369s + 1)}. \tag{30}$$

For the hypotheses of the robust stability Theorem 3.3, the pair of distinct transfer functions h, $\tilde{h}$, and the associated $\tilde{H}$ are constructed through

$$h(s) = (p(s) - \tilde{p}(s))qq_d(s), \tag{31}$$

$$\tilde{h}(s) = (1 - \tilde{p}(s)qq_d(s)), \tag{32}$$

and, recalling (3),

$$\tilde{H}(s) = \frac{(1 - \tilde{p}(s)qq_d(s))}{1 + (p(s) - \tilde{p}(s))qq_d(s)}. \tag{33}$$

Compare the transfer functions $\tilde{H}(s)$ in (33) and $y(s)$ in (20).

Figure 4 shows that the maximum value of the sensitivity function is bounded for both control systems. Since a separate Nyquist analysis of the control systems for a process with a spring constant of $k = 0.8$ shows that each control system is stable for that value of the spring constant, we conclude from our robust stability Theorem 3.3 that the control system is stable for all values of the spring constant in the uncertainty set. Further, since the maximum of the sensitivity function over all frequencies is smaller for controller $C_1(s)$ than for controller $C_\mu(s)$, we expect that the time responses for the control system using $C_1(s)$ are likely to be less oscillatory than those for $C_\mu(s)$, at least for some values of the spring constant. The time responses in Figures 5 and 6 confirm this expectation.

Example 4.3 (unstable first-order lag plus deadtime process). This example demonstrates the application of our robust stability Theorem 3.3 to an infinite dimensional system. It also points out the need to check the stability of one process which is not equal to the model in order to draw correct conclusions regarding the stability of all processes in the set of uncertain processes. The process is

$$p(s) = \frac{Ke^{-s}}{(-s+1)},\ 0.9 \leq K \leq 1.1. \tag{34}$$

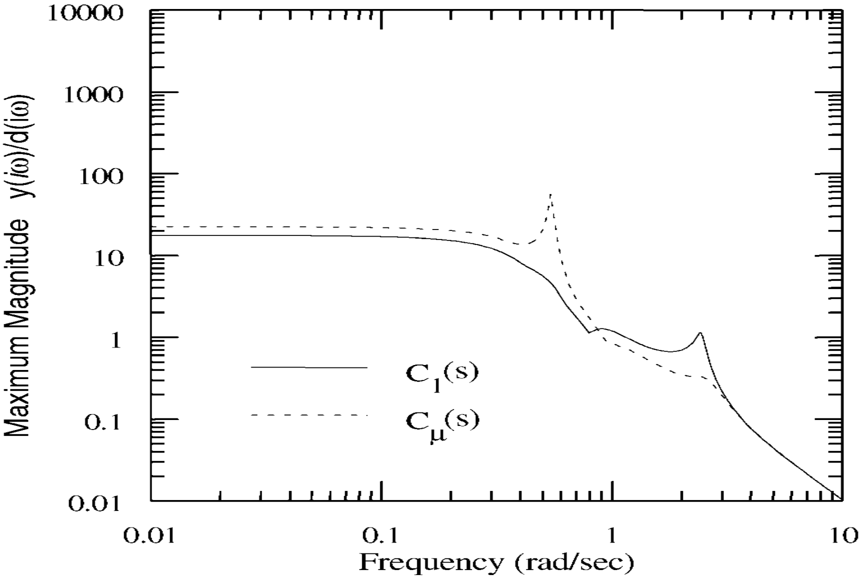


FIG. 4. *Maximum magnitude of the sensitivity function using controllers* $C_1(s)$ *and* $C_\mu(s)$*.*

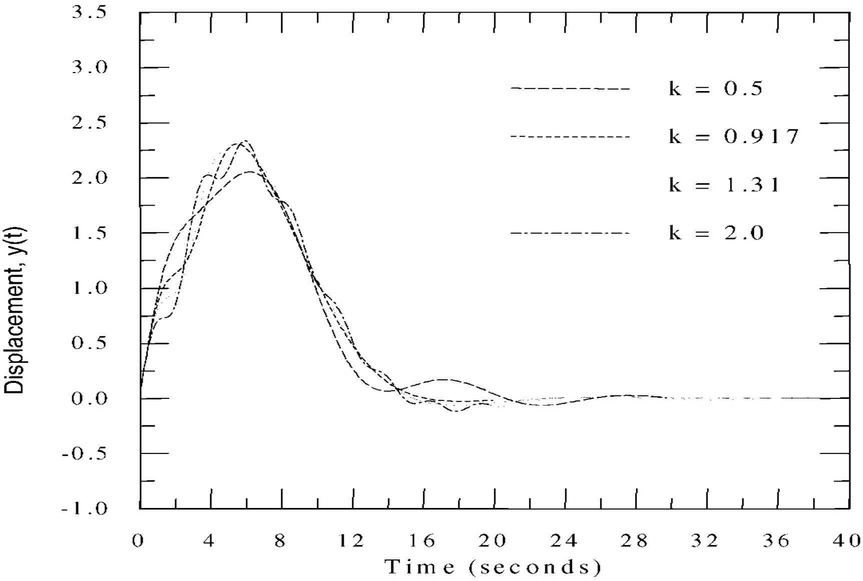


FIG. 5. *Response of body* 2 *position,* $y(t)$*, to an impulse disturbance for different spring constant values using controller* $C_1(s)$*.*

This process is one that is not easy to control, in spite of the relatively small range of uncertain gains, because of the large magnitude of the deadtime to time constant ratio.

The process model is taken as

$$\tilde{p}(s) = \frac{e^{-s}}{(-s+1)}. \tag{35}$$

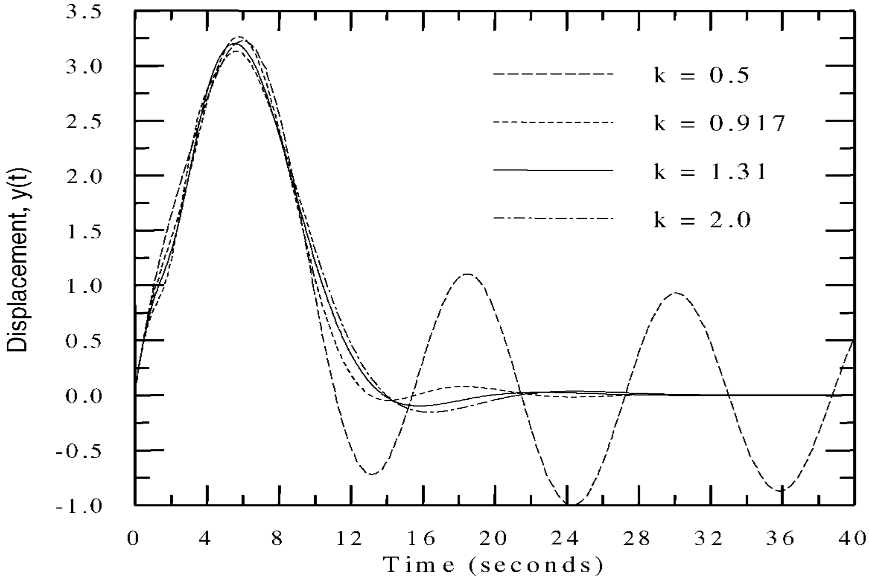


FIG. 6. *Response of body* 2 *position, $y(t)$, to an impulse disturbance for different spring constant values using controller $C_\mu(s)$.*

Selecting a filter time constant, ε, of 0.5 gives

$$qq_d(s, 0.5) = \frac{(-s+1)(5.116s+1)}{(0.5s+1)^2}. \tag{36}$$

Substituting (34), (35), and (36) into (20) and computing the sensitivity function yields Figure 7. For the hypotheses of the robust stability Theorem 3.3, the pair of distinct transfer functions h, $\tilde{h}$, and the associated $\tilde{H}$ are constructed, as in the previous example, through (31), (32), and (33).

Figure 7 indicates that the sensitivity function is finite over all positive frequencies. However, to conclude from this that the control system is stable over all uncertain processes, we must also see if the control system is stable for any single process gain, K, in the range 0.9 to 1.1, but not equal to 1.0. Since the control system given by Figure 3 is internally unstable, a common alternative is to use the control system of Figure 2 with the controller, $C(s)$, given by (22). A Nyquist analysis of this controller with $qq_d(s)$ given by (36) shows that the controller has six right half-plane poles. Further, a Nyquist analysis of the sensitivity function given by (19), using the same controller, shows that the disturbance response is unstable for a process gain of 1.1. Thus, from our robust stability Theorem 3.3, we can conclude that the control system is unstable for all process gains in the uncertainty range.

If the controller filter time constant, ε, is increased to 2.8, then the two degree of freedom IMC controller becomes

$$qq_d(s, 2.8) = \frac{(-s+1)(38.25s+1)}{(2.8s+1)^2}. \tag{37}$$

A Nyquist analysis of the controller found by substituting (37) into (22) shows that the denominator has one right half-plane zero, but this zero is exactly at 1, and so is cancelled by the numerator zero at 1. This cancellation must be enforced so that the

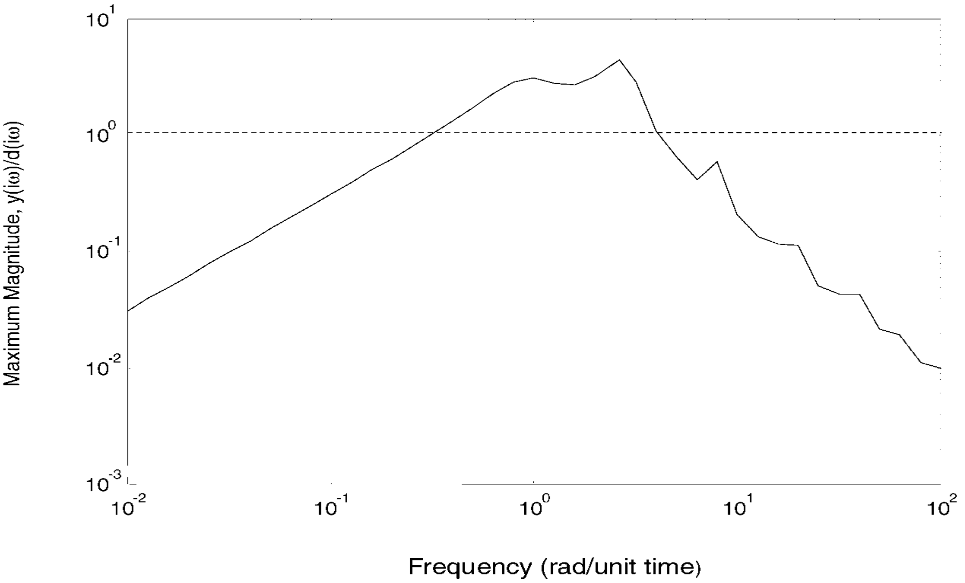


FIG. 7. *Upper bound of sensitivity function when* $\varepsilon = 0.5$.

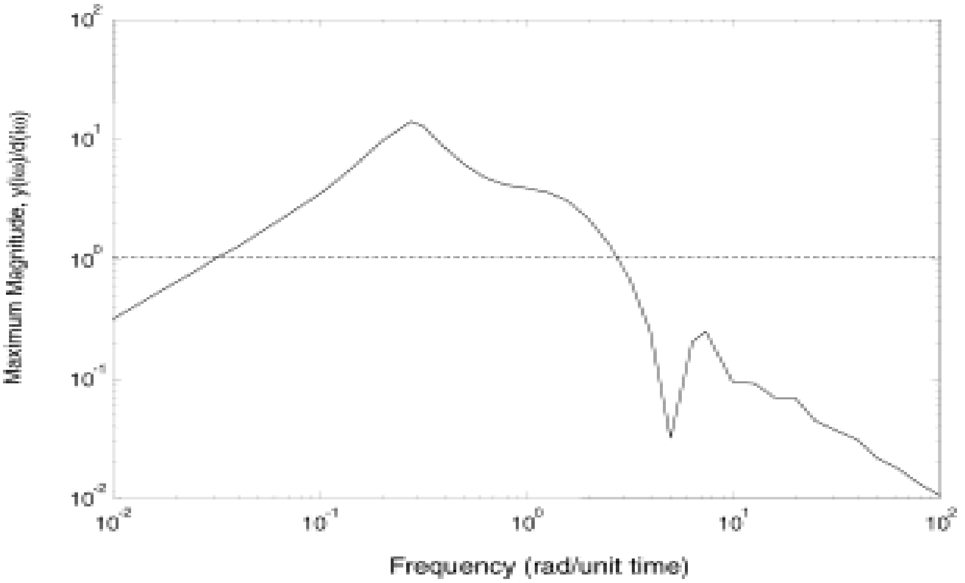


FIG. 8. *Upper bound of sensitivity function when* $\varepsilon = 2.8$.

controller that is actually implemented is stable, except of course for the pole at the origin. The upper bound of the sensitivity function for the control system associated with (37) is shown in Figure 8.

From Figure 8, and our robust stability Theorem 3.3, we conclude that the control system is stable for all process gains between 0.9 and 1.1 because a Nyquist analysis for a gain of 1.1 shows that this control system is stable. The time response for a unit step disturbance and a unit step set point change for the control system with $\varepsilon = 2.8$ and $\varepsilon_r = 2.8$ are shown in Figures 9 and 10.

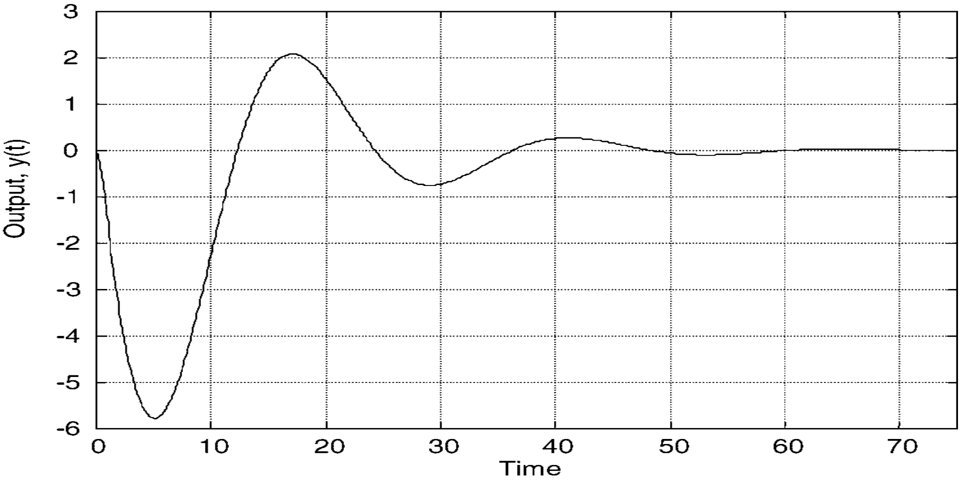


FIG. 9. *Disturbance step response when $\varepsilon_r = 3$ and $\varepsilon = 2.8$.*

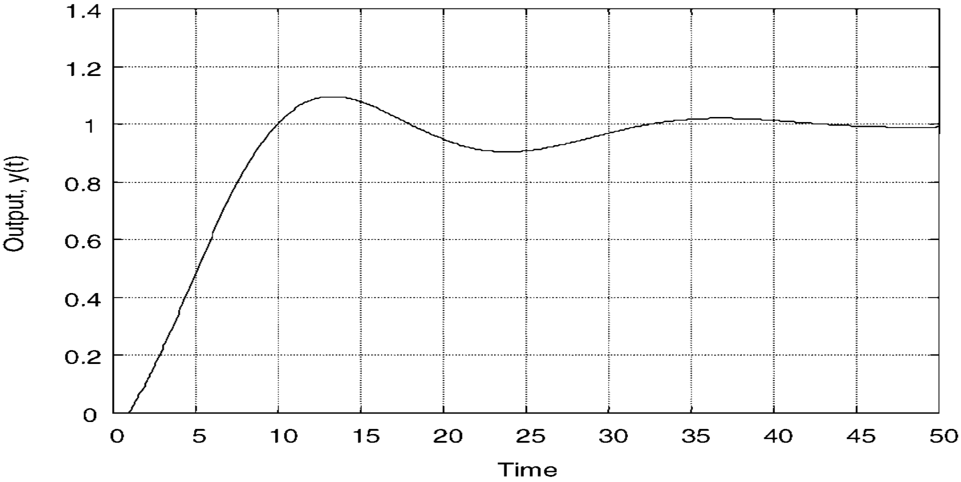


FIG. 10. *Set point step response when $\varepsilon_r = 3$ and $\varepsilon = 2.8$.*

It is instructive to compare the responses of Figures 9 and 10 with those of an internally stable control system such as that proposed by Berber and Brosilow [1, 2]. These responses are given in Figures 11 and 12 for the same process (i.e., $K = 0.9$). Based on the height and position of the peaks in Figures 7 and 8, one would expect the time responses associated with Figure 7 to be faster and less oscillatory than those associated with Figure 8. As can be seen from Figures 11 and 12, this is indeed the case.

5. Computational considerations. From a practical point of view, implementation of our results depends upon our ability to compute efficiently the maximum magnitude function, $\mathcal{H}$, in (5). In Examples 4.2 and 4.3, we used our robust stability results to establish the stability or instability of closed loop systems, and, in Example 4.2, to select an optimal model parameter. This was done with the aid of

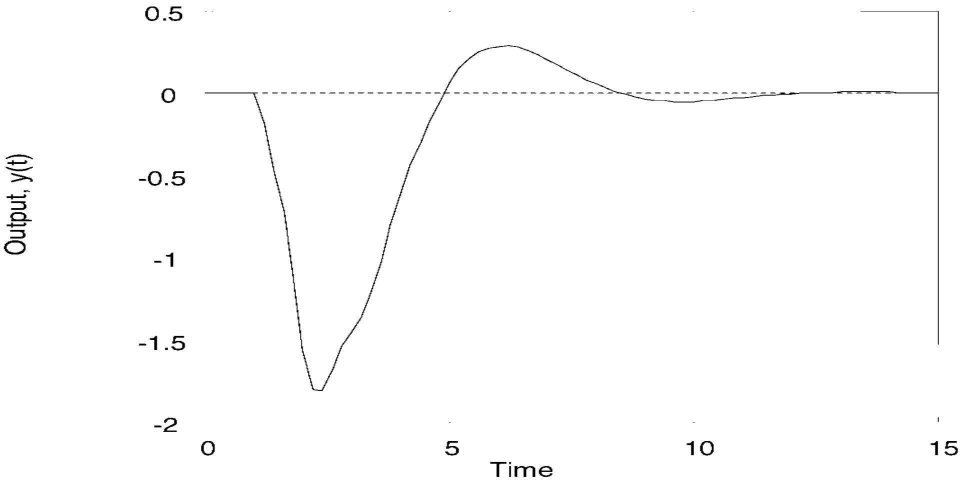


FIG. 11. *Disturbance step response when $\varepsilon_r = 1.3$ and $\varepsilon = 0.5$.*

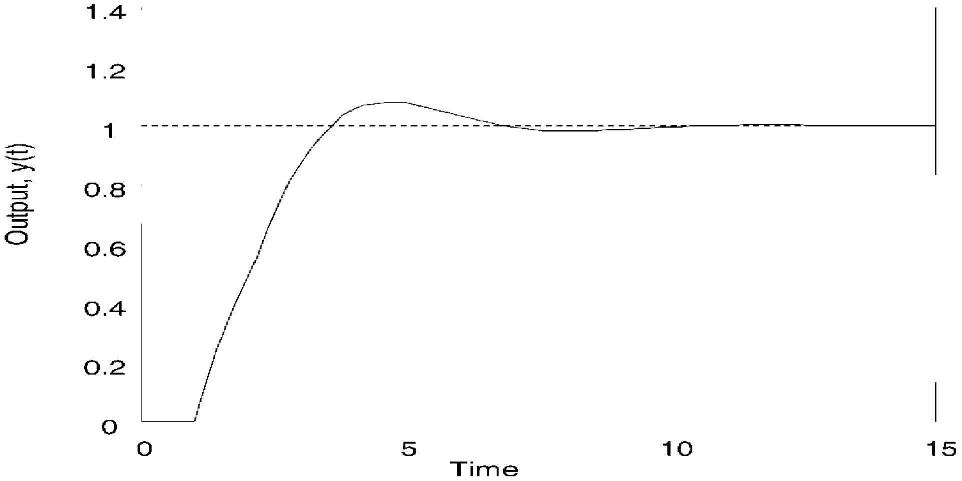


FIG. 12. *Set point step response when $\varepsilon_r = 1.3$ and $\varepsilon = 0.5$.*

MATLAB-based software developed by Strycezk [9].[12] This software uses the following approach to keep the computations tractable. Starting at a very high frequency,[13] the algorithm seeks the maximum magnitude of the specified closed loop transfer function over frequency and all uncertain parameters. The result of this optimization is the highest frequency local (possibly global) maximum. The algorithm then searches for all local maxima between the frequency just found and a frequency below which

[12] This software is available free of charge at the web site http://cheme.cwru.edu/People/Faculty/brosilow/brosilow.htm#brosilow. It requires MATLAB 5.3, or higher, as well as the Control System and Optimization Toolboxes. The user interface makes data input quite comfortable and online help is available. Results of the computations are presented as both graphs and tables. Since URLs tend to change over time, a relatively sure method of arriving at the site containing this software is first to access the home page of Case Western Reserve University (http://www.cwru.edu), and from there proceed to the home page of the Chemical Engineering Department, and finally to the home page of C. Brosilow.

[13] In this case, 1000/ the smallest model, uncertain process, or controller time constant.

the magnitude of the closed loop frequency response is constant (e.g., 1 for integral control systems). The results of these calculations are presented in a graph of magnitude versus frequency. Unstable control systems usually show magnitudes higher than 10^8. However, for practical purposes, closed loop magnitudes higher than 100 times the closed loop gain can be taken as evidence of an effectively unstable control system.

The optimization algorithm used in the above calculation is the constrained optimization algorithm provided by the MATLAB Optimization Toolbox. The optimization algorithm has worked well with less than six uncertain parameters, and we have experience with such processes. The algorithm should also work well with many more than six uncertain parameters, but we do not yet have experience with such problems.

Acknowledgments. One of the authors (M.J.L.) is indebted to his colleagues Professors Alejandro de Acosta, Michael Hurley, Thomas Ivey, and Arthur Obrock, all of Case Western Reserve University, for their valuable comments.

REFERENCES

[1] R. Berber and C. Brosilow, *Internally stable linear and nonlinear algorithmic internal model control of unstable systems*, in Proceedings of the NATO ASI Series, Series E: Applied Sciences, Vol. 353, 1997, pp. 209–234.

[2] R. Berber and C. Brosilow, *Algorithmic internal model control of unstable systems*, in Proceedings of the 7th IEEE Mediterranean Conference on Control & Automation, Haifa, Israel, 1999, pp. 28–30.

[3] R. Braatz and M. Morari, *Robust control for a noncollocated spring-mass system*, J. Guidance Control Dynamics, 15 (1992), pp. 1103–1109.

[4] C. Cheng and C. Brosilow, *Model Predictive Control of Unstable Systems*, presented at the Annual AIChE Meeting, New York, 1987.

[5] J. C. Doyle, *Structured uncertainty in control system design*, in Proceedings of the 24th IEEE Conference on Decision and Control, 1985, pp. 260–265.

[6] L. Hörmander, *An Introduction to Complex Analysis in Several Variables*, Elsevier, New York, 1973.

[7] V. L. Kharitonov, *Asymptotic stability of a family of systems of linear differential equations*, Differentsial'nye Uravneniya, 14 (1978), pp. 2986–2088 (in Russian). English translation in Differential Equations, 14 (1978), pp. 1483–1485.

[8] M. Morari and E. Zafiriou, *Robust Process Control*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[9] K. Strycezk, *Tuning Multivariable Control Systems for Parametric Uncertainty*, Ph.D. thesis, Case Western Reserve University, Cleveland, OH, 1996.

[10] G. Vinnicombe, *Frequency domain uncertainty and the graph topology*, IEEE Trans. Automat. Control, 38 (1993), pp. 1371–1383.

[11] B. Wei and D. S. Bernstein, *A benchmark problem for robust control design*, in Proceedings of the 1990 American Control Conference, 1990, pp. 961–962.

[12] B. Wei and D. S. Bernstein, *Benchmark problems for robust control design*, J. Guidance Control Dynamics, 15 (1992), pp. 1057–1059.

[13] G. Zames and A. J. El-Sakkary, *Unstable systems and feedback: The gap metric*, in Proceedings of the Allerton Conference, 1980, pp. 380–385.

SIAM J. Control Optim.
Vol. 39, No. 5, pp. 1441–1460

MULTIPLIER RULES UNDER MIXED ASSUMPTIONS OF DIFFERENTIABILITY AND LIPSCHITZ CONTINUITY*

JANE J. YE†

Abstract. In this paper we study nonlinear programming problems with equality, inequality, and abstract constraints where some of the functions are Fréchet differentiable at the optimal solution, some of the functions are Lipschitz near the optimal solution, and the abstract constraint set may be nonconvex. We derive Fritz John type and Karush–Kuhn–Tucker (KKT) type first order necessary optimality conditions for the above problem where Fréchet derivatives are used for the differentiable functions and subdifferentials are used for the Lipschitz continuous functions. Constraint qualifications for the KKT type first order necessary optimality conditions to hold include the generalized Mangasarian–Fromovitz constraint qualification, the no nonzero abnormal multiplier constraint qualification, the metric regularity of the constraint region, and the calmness constraint qualification.

Key words. necessary optimality conditions, Fréchet differentiability, subdifferentials, constraint qualifications, metric regularity, calmness

AMS subject classifications. 49K10, 90C30

PII. S0363012999358476

1. Introduction.

The classical multiplier rule usually requires that the objective function and the inequality constraints be differentiable, the equality constraints be continuously differentiable at the optimal solution, and the abstract constraint set be convex with nonempty interior (e.g., see Bazaraa, Sherali, and Shetty [1] and Mangasarian [14]).

Over the last three decades, the classical multiplier rule was extended under two different assumptions: differentiability and Lipschitz continuity.

On the one hand, the classical multiplier rule was extended in the direction of eliminating the smoothness assumption while keeping the differentiability assumption. In the case where there is no abstract constraint, based on a correction theorem, Halkin [9] proved that the classical multiplier rule holds under the weaker assumption which requires only that the equality constraints be Fréchet differentiable at the optimal solution and continuous in a neighborhood of the optimal solution. Based on a multidimensional intermediate value theorem, Di [7] derived some first order and second order multiplier rules for nonlinear programming problems with equality, inequality, and abstract constraints where all functions are Fréchet differentiable at the optimal solution and continuous in a neighborhood of the optimal solution and the abstract constraint set is convex.

On the other hand, in nonsmooth analysis the classical multiplier rule was generalized in the direction of replacing the differentiability assumption by the Lipschitz continuity assumption. Under the assumption that all functions are Lipschitz near the optimal solution and the abstract constraint set is closed, Clarke [3] derived a generalized multiplier rule involving the Clarke generalized gradient and the Clarke normal

*Received by the editors July 6, 1999; accepted for publication (in revised form) August 23, 2000; published electronically January 19, 2001. The research of this paper was partially supported by NSERC and the University of Victoria.

http://www.siam.org/journals/sicon/39-5/35847.html

†Department of Mathematics and Statistics, University of Victoria, Victoria BC, Canada V8W 3P4 (janeye@Math.UVic.CA).

cone. The Clarke generalized gradient of a function would reduce to the usual derivative only when the function is strictly differentiable (for example, when the function is continuously differentiable). Hence, when all functions involved are continuously differentiable and the abstract constraint set is convex, the generalized multiplier rule of Clarke would recover the classical multiplier rule. However, the Clarke generalized gradient of a Lipschitz continuous function may be strictly larger than the set which consists of the usual derivative when the function is Fréchet differentiable but not strictly differentiable. In the case when the abstract set is convex, Ioffe [11] showed that the Clarke generalized multiplier rule can be sharpened by replacing the Clarke generalized gradient by the Michel–Penot subdifferential which coincides with the usual derivative when the function is Gâteaux differentiable. Other results in this direction also include Mordukhovich's combined multiplier rule [16] and the Treiman's multiplier rule [19].

In this paper we study first order necessary optimality conditions for nonlinear programming problems with equality, inequality, and abstract constraints where some of the functions are Fréchet differentiable at the optimal solution, some of the functions are Lipschitz near the optimal solution, and the abstract constraint set may be nonconvex. For the above nonlinear programming problem with mixed assumptions on differentiability and Lipschitz continuity, since a differentiable function may not be Lipschitz continuous, the only applicable necessary optimality conditions in the literature are fuzzy multiplier rules (see, e.g., Borwein and Zhu [2]). Although in a finite dimensional space the fuzzy multiplier rule reduces to an exact multiplier rule, it involves the singular subdifferential of the non-Lipschitz functions. Our purpose is to derive exact (i.e., nonfuzzy) first order multipler rules which do not involve any singular subdifferentials for the above problem where Fréchet derivatives are used for the differentiable functions and subdifferentials are used for the Lipschitz continuous functions.

To be more precise, we consider the following optimization problem:

$$
\begin{array}{lll}
\text{(P)} & \text{minimize} & f(x) \\
& \text{subject to} & g_i(x) \le 0, \quad i = 1, 2, \ldots, I, \\
& & h_j(x) = 0, \quad j = 1, 2, \ldots, J, \\
& & \phi_k(x) \le 0, \quad k = 1, 2, \ldots, K, \\
& & \psi_l(x) = 0, \quad l = 1, 2, \ldots, L, \\
& & x \in \Omega,
\end{array}
$$

where $f, g_i(i = 1, 2, \ldots, I), h_i(j = 1, 2, \ldots, J), \phi_k(k = 1, 2, \ldots, K), \psi_l(l = 1, 2, \ldots, L)$ are the objective function and the constraint functions from a Banach space X to R. Ω is a closed subset of X and I, J, K, L are given integers. Generally one has $I \ge 1, J \ge 1, K \ge 1, L \ge 1$, but we allow I, J, K, or $L = 0$ to signify the case in which there are no explicit constraints of the type.

Let $\bar{x}$ be a local optimal solution to (P). Denote by $I(\bar{x}) := \{i : g_i(\bar{x}) = 0\}$ and $K(\bar{x}) := \{k : \phi_k(\bar{x}) = 0\}$ the index sets of the binding constraints. We always make the following basic assumptions on the constraint functions.

(A) $g_i(i \in I(\bar{x})), h_j(j = 1, 2, \ldots, J)$ are Fréchet differentiable at $\bar{x}$ and $g_i(i \notin I(\bar{x}))$ are continuous at $\bar{x}$. $\phi_k(k \in K(\bar{x})), \psi_l(l = 1, 2, \ldots, L)$ are Lipschitz near $\bar{x}$ and $\phi_k(k \notin K(\bar{x}))$ are continuous at $\bar{x}$.

Our main results include the following multiplier rules.

Theorem 1.1 (Fritz John necessary optimality conditions for the case $L = 0$).

Let $\bar{x}$ be a local optimal solution of (P) *with $L = 0$. Suppose that f is either Fréchet differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, in addition to assumption* (A), *$h_j (j = 1, 2, \dots, J)$ are continuous in a neighborhood of $\bar{x}$, and there exists a vector that is hypertangent* (*see Definition* 2.4) *to the abstract constraint set Ω at $\bar{x}$. Then there exist scalars $\lambda \geq 0, \alpha_i \geq 0 (i \in I(\bar{x})), \beta_j (j = 1, 2, \dots, J), \gamma_k \geq 0 (k \in K(\bar{x}))$ not all zero such that*

$$0 \in \lambda \partial^{\diamond} f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \partial^{\diamond} \phi_k(\bar{x}) + N(\bar{x}, \Omega),$$

where $\partial^{\diamond}$ denotes the Michel–Penot subdifferential, ∇ denotes the Fréchet derivative, and $N(\bar{x}, \Omega)$ denotes the Clarke normal cone to Ω at $\bar{x}$.

Remark 1. Note that in the case where f is Fréchet differentiable at $\bar{x}$, $\partial^{\diamond} f(\bar{x}) = \{\nabla f(\bar{x})\}$ in the above multiplier rule. As it was shown by Fernandez [8], the continuity assumption of the equality constraints h_j in Theorem 1.1 cannot be removed.

THEOREM 1.2 (Fritz John necessary optimality conditions for the case $I = J = 0$). *Let $\bar{x}$ be a local optimal solution of* (P) *with $I = J = 0$. Suppose that the objective function f is Fréchet differentiable at $\bar{x}$, $\phi_k (k \in K(\bar{x})), \psi_l (l = 1, 2, \dots, L)$ are Lipschitz near $\bar{x}$ and $\phi_k (k \notin K(\bar{x}))$ are continuous at $\bar{x}$. Then there exist scalars $\lambda \geq 0, \gamma_k \geq 0 (k \in K(\bar{x})), \eta_l (l = 1, 2, \dots, L)$ not all zero such that*

$$0 \in \lambda \nabla f(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \partial \phi_k(\bar{x}) + \sum_{l=1}^{L} \eta_l \partial \psi_l(\bar{x}) + N(\bar{x}, \Omega),$$

where ∂ denotes the Clarke generalized gradient and $N(x, \Omega)$ denotes the Clarke normal cone to Ω at $\bar{x}$. Moreover, if X is an Asplund space which is a Banach space whose separable subspaces have separable duals (as is the case for reflexive spaces), under the above assumptions, there exist scalars $\lambda \geq 0, \gamma_k \geq 0 (k \in K(\bar{x})), \eta_l (l = 1, 2, \dots, L)$ not all zero such that

$$0 \in \lambda \nabla f(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \hat{\partial} \phi_k(\bar{x}) + \hat{\partial} \left(\sum_{l=1}^{L} \eta_l \psi_l \right) (\bar{x}) + \hat{N}(\bar{x}, \Omega),$$

where $\hat{\partial}$ denotes the limiting subdifferential and $\hat{N}(\bar{x}, \Omega)$ denotes the limiting normal cone to Ω at $\bar{x}$.

As in smooth and Lipschitz optimization we also give constraint qualifications under which the scalar λ in the above theorems is nonzero such as the generalized Mangasarian–Fromovitz constraint qualification (GMFCQ), the no nonzero abnormal multiplier constraint qualification (NNAMCQ), the metric regularity of the constraint region (metric regularity CQ), and the calmness constraint qualification (calmness CQ).

We organize the paper as follows. In the next section we provide preliminaries that will be used in the paper. In section 3, we prove Theorem 1.1, the Fritz John type necessary optimality condition for the case where there are no Lipschitz continuous equality constraints. In section 4, we introduce constraint qualifications, discuss the relationship between the (GMFCQ) and (NNAMCQ), and prove that under constraint qualifications such as the (NNAMCQ), the metric regularity CQ and the calmness CQ, λ in Theorems 1.1 can be taken as 1. An example is given to show that when the objective function is not Lipschitz but only Fréchet differentiable, the metric regularity CQ may not imply the calmness CQ. Hence the well-known relationships between these constraint qualifications may not hold when some of the functions are not Lipschitz but Fréchet differentiable. However, it turns out that the Karush–Kuhn–Tucker (KKT) conditions can usually be derived directly. We prove that unlike the Fritz John type condition (Theorem 1.1), under the metric regularity CQ and the calmness CQ the KKT condition holds even in the case where $L \neq 0$, and the continuity assumption of the Fréchet differentiable equality constraints is not needed. In section 5, we derive KKT type necessary optimality conditions for the case where all constraint functions are Lipschitz continuous and the objective function is Fréchet differentiable under the constraint qualification (NNAMCQ), the metric regularity CQ, and the calmness CQ. Theorem 1.2, the Fritz John type necessary optimality condition, then follows as an easy consequence.

2. Preliminaries.

This section contains some background material on nonsmooth analysis which will be used throughout the paper. We give only concise definitions that will be needed in the paper. For more detailed information on the subject, our references are Clarke [4], Clarke, Ledyaev, Stern, and Wolenski [5], Loewen [13], and Mordukhovich and Shao [17].

We first give the following notations that will be used throughout the paper. For a vector $v \in R^n$, v_i is the ith components of v. For any Banach space X we denote its norm by $\|\cdot\|$ and consider the dual space X^* equipped with the weak-star topology w^*, where $\langle\cdot,\cdot\rangle$ means the canonic pairing. As usual, B and B^* stand for the open unit balls in the space and the dual space in question. Note that intΩ, clΩ, and *co*Ω mean, respectively, the interior, the closure, and the convex hull of an arbitrary nonempty set $\Omega \subset X$, while the notation cl^* is used for the weak-star topological closure in X^*.

For a set-valued map $\Phi : X \rightrightarrows X^*$, we denote by

$$\limsup_{x\to\bar{x}} \Phi(x)$$

the sequential Kuratowski–Painlevé upper limit with respect to the norm topology in X and the weak-star topology in X^*, i.e.,

$$\limsup_{x\to\bar{x}} \Phi(x) := \{x^* \in X^* | \exists \text{ sequences } x_k \to \bar{x}, x_k^* \xrightarrow{w^*} x^*, \\ \text{with } x_k^* \in \Phi(x_k) \forall k = 1, 2, \ldots\}.$$

We now give some concepts for various normal cones.

Definition 2.1. *Let Ω be a nonempty subset of a Banach space X and let $\epsilon \geq 0$.*

(i) *Given $x \in \text{cl}\Omega$, the set*

$$N_\epsilon^F(x,\Omega) := \left\{ x^* \in X^* | \limsup_{x'\to x, x'\in\Omega} \frac{\langle x^*, x'-x\rangle}{\|x'-x\|} \leq \epsilon \right\} \tag{2.1}$$

is called the set of Fréchet ϵ-normals to Ω at x. When $\epsilon = 0$, the set (2.1) is a cone which is called the Fréchet normal cone to Ω at x and is denoted by $N^F(x,\Omega)$.

(ii) *Let $\bar{x} \in \text{cl}\Omega$. The nonempty cone*

$$\hat{N}(\bar{x},\Omega) := \limsup_{x\to\bar{x},\epsilon\downarrow 0} N_\epsilon^F(x,\Omega) \tag{2.2}$$

is called the limiting normal cone to Ω at $\bar{x}$.

Using the definitions for normal cones, we now give definitions for corresponding subdifferentials of a single-valued map.

DEFINITION 2.2. *Let X be a Banach space and $\varphi : X \to R \cup \{+\infty\}$ be l.s.c. (lower semicontinuous) and finite at $x \in X$. The sets*

$$\begin{aligned} \partial^F_\epsilon \varphi(x) &:= \{x^* \in X^* | (x^*, -1) \in N^F_\epsilon((x, \varphi(x)), epi\varphi)\}, \\ \hat{\partial}\varphi(x) &:= \{x^* \in X^* | (x^*, -1) \in \hat{N}((x, \varphi(x)), epi\varphi)\}, \end{aligned} \tag{2.3}$$

where $epi\varphi := \{(x, v) : v \geq \varphi(x)\}$ denotes the epigraph of φ, are called, respectively, the Fréchet ϵ-subdifferential and the limiting subdifferential of φ at x. When $\epsilon = 0$, the set (2.3) *is called the Fréchet subdifferential of φ at x and is denoted by $\partial^F\varphi(x)$.* It is known that the Fréchet subdifferential has the following analytic expression:

$$\partial^F\varphi(x) = \left\{x^* \in X^* | \liminf_{x' \to x} \frac{\varphi(x') - \varphi(x) - \langle x^*, x' - x\rangle}{\|x' - x\|} \geq 0\right\}. \tag{2.4}$$

Let X be any Banach space, $\bar{x} \in X$, and $\varphi : X \to R$ be any continuous function. Then the Michel–Penot directional derivative of φ at $\bar{x}$ in the direction $v \in X$ introduced in [15] is given by

$$\varphi^\square(\bar{x}; v) := \sup_{w \in X} \limsup_{t \downarrow 0} \frac{\varphi(\bar{x} + t(v + w)) - \varphi(\bar{x} + tw)}{t}$$

and the Michel–Penot subdifferential of φ at $\bar{x}$ is given by the set

$$\partial^\diamond\varphi(\bar{x}) := \{x^* \in X^* | \langle x^*, v\rangle \leq \varphi^\square(\bar{x}; v)\ \forall v \in X\}.$$

It is known (see [15, Proposition 1.3]) that when a function φ is Gâteaux differentiable at $\bar{x}$, $\partial^\diamond\varphi(\bar{x}) = \{\nabla\varphi(\bar{x})\}$.

The following properties of the Michel–Penot directional derivatives and the Michel–Penot subdifferentials will be useful.

PROPOSITION 2.3 (see [15, 11]). *Let X be a Banach space, $x \in X$, and f be Lipschitz near x with constant L_f. Then*

(i) *The function $v \to f^\square(x; v)$ is finite, positively homogeneous, and subadditive on X.*

(ii) *As a function of v, $f^\square(x; v)$ is Lipschitz continuous with constant L_f on X.*

(iii) *$\partial^\diamond f(x)$ is a nonempty, convex, weak*-compact subset of X^* and $\|x^*\| \leq L_f$ for every $x^* \in \partial^\diamond f(x)$.*

Let X be any Banach space, $\bar{x} \in X$, and $\varphi : X \to R$ be Lipschitz near $\bar{x}$. Then the Clarke generalized derivative of φ at $\bar{x}$ in the direction $v \in X$ is given by

$$\varphi^0(\bar{x}; v) := \limsup_{x \to \bar{x}, t \downarrow 0} \frac{\varphi(x + tv) - \varphi(x)}{t}$$

and the Clarke generalized gradient of φ at $\bar{x}$ is given by the set

$$\partial\varphi(\bar{x}) := \{x^* \in X^* | \langle x^*, v\rangle \leq \varphi^0(\bar{x}; v)\ \forall v \in X\}.$$

Let Ω be a nonempty subset of a Banach space X and consider its distance function, that is, the function $d_\Omega(\cdot) : X \to R$ defined by

$$d_\Omega(x) = \inf\{\|x - c\| : c \in \Omega\}.$$

The Clarke tangent cone to Ω at $\bar{x}$ is defined by

$$T(\bar{x},\Omega) := \{v \in X \mid d^0_\Omega(\bar{x};v) = 0\}$$

and the Clarke normal cone to Ω at $\bar{x}$ is defined by polarity with $T(\bar{x},\Omega)$:

$$N(\bar{x},\Omega) := \{x^* \in X^* | \langle x^*, v\rangle \le 0 \quad \forall v \in T(\bar{x},\Omega)\}.$$

DEFINITION 2.4 (hypertangent). *Let X be a Banach space. A vector v in X is said to be hypertangent to the set $\Omega \subseteq X$ at the point $x \in \Omega$ if for some $\epsilon > 0$,*

$$y + tw \in \Omega \quad \forall y \in (x+\epsilon B)\cap\Omega, w \in v + \epsilon B, t \in (0,\epsilon).$$

It follows easily that any vector v hypertangent to Ω at x belongs to $T(x,\Omega)$. It is possible to have no hypertangents at all. However, it is clear that when Ω is a convex set with nonempty interior, then any vector $x^* - x$ with $x^* \in \text{int}\Omega$ is hypertangent to Ω at x.

It is known that in any Banach space X and for any $\epsilon \ge 0$

$$N^F_\epsilon(\bar{x},\Omega) \subseteq \hat{N}(\bar{x},\Omega) \subseteq N(\bar{x},\Omega),$$
$$\partial^F_\epsilon \varphi(\bar{x}) \subseteq \hat{\partial}\varphi(\bar{x}) \subseteq \partial^\diamond \varphi(\bar{x}) \subseteq \partial\varphi(\bar{x})$$

and in any Asplund space, the following precise relationships hold [17, Theorems 2.9 and 8.11]:

(i) For any closed set $\Omega \subseteq X$ and $\bar{x} \in \Omega$ one has

$$\hat{N}(\bar{x},\Omega) = \limsup_{x\to\bar{x}} N^F(x,\Omega),$$
$$N(\bar{x};\Omega) = \text{cl}^* co\hat{N}(\bar{x},\Omega).$$

(ii) For any function $\varphi : X \to R$ which is Lipschitz near $\bar{x} \in X$, one has

$$\hat{\partial}\varphi(\bar{x}) = \limsup_{x\to\bar{x}} \partial^F\varphi(x),$$
$$\partial\varphi(\bar{x}) = \text{cl}^* co\hat{\partial}\varphi(\bar{x}).$$

We now summarize the sum rules and chain rules for the various subdifferentials in the literature. For convenience, we do not intend to quote the results under the most general assumptions. Instead, we provide the results under the assumptions we need in our paper. For example, since when Y is finite dimensional, a function $\varphi : X \to Y$ is Lipschitz near $\bar{x} \in X$ is strictly Lipschitzian at $\bar{x}$ in the sense of [17]; Propositions 2.5(ii) and 2.6(ii) are special cases of the results in [17].

PROPOSITION 2.5 (sum rules).

(i) (*See, e.g., the proof of* [6, Lemma 2.2].) *Let X be a Banach space and $\bar{x} \in X$. Let $\varphi_1 : X \to R$ be Fréchet differentiable at $\bar{x}$ and $\varphi_2 \to R \cup \{+\infty\}$ be finite and l.s.c. at $\bar{x}$. Then*

$$\partial^F(\varphi_1 + \varphi_2)(\bar{x}) = \nabla\varphi_1(\bar{x}) + \partial^F\varphi_2(\bar{x}).$$

(ii) (*See* [17, Proposition 2.5 and Theorem 4.1].) *Let X be an Asplund space and $\bar{x} \in X$. Let $\varphi_i : X \to R \cup \{+\infty\}, i = 1,2,$ be l.s.c. at $\bar{x}$ and one of these functions is Lipschitz near $\bar{x}$. Then one has*

$$\hat{\partial}(\varphi_1 + \varphi_2)(\bar{x}) \subseteq \hat{\partial}\varphi_1(\bar{x}) + \hat{\partial}\varphi_2(\bar{x}).$$

(iii) (*See* [4, Proposition 2.3.3].) *Let X be a Banach space and $\bar{x} \in X$. Let $\varphi_i : X \to R, i = 1, 2$, be Lipschitz near $\bar{x}$. Then one has*

$$\partial(\varphi_1 + \varphi_2)(\bar{x}) \subseteq \partial\varphi_1(\bar{x}) + \partial\varphi_2(\bar{x}).$$

Proposition 2.6 (chain rules).

(i) (*See* [5, Theorem 2.5].) *Let X be a Banach space and $\bar{x} \in X$. Suppose that $\varphi : X \to R^n$ is Lipschitz near $\bar{x}$ and $f : R^n \to R$ is Lipschitz near $\varphi(\bar{x})$. Then*

$$\partial(f \circ \varphi)(\bar{x}) \subseteq \mathrm{cl}^* co \cup_{y^* \in \partial f(\varphi(\bar{x}))} \partial\langle y^*, \varphi\rangle(\bar{x}).$$

(ii) (*See* [17, Proposition 2.5 and Corollary 6.3].) *Moreover, if X is an Asplund space, then*

$$\hat{\partial}(f \circ \varphi)(\bar{x}) \subseteq \cup_{y^* \in \hat{\partial} f(\varphi(\bar{x}))} \hat{\partial}\langle y^*, \varphi\rangle(\bar{x}).$$

The following exact penalty results given by Clarke in [4, Proposition 2.4.3] will often be used in the paper.

Proposition 2.7. *Let C be a closed subset of X. Assume that f attains a minimum over C at $\bar{x} \in C$ and f is Lipschitz near $\bar{x}$ with constant $L_f > 0$. Then for all $K \geq L_f$, the function $g(y) = f(y) + K d_C(y)$ also attains a minimum over X at $\bar{x}$.*

3. Proof of Theorem 1.1.

We need only to prove the theorem under the assumption that there do not exist scalars $\alpha_i \geq 0 (i \in I(\bar{x})), \beta_j (j = 1, 2, \ldots, J), \gamma_k \geq 0 (k \in K(\bar{x}))$ not all zero such that

$$0 \in \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \partial^\diamond \phi_k(\bar{x}) + N(\bar{x}, \Omega). \tag{3.1}$$

Indeed, if (3.1) is satisfied by some scalars $\alpha_i \geq 0 (i \in I(\bar{x})), \beta_j (j = 1, 2, \ldots, J), \gamma_k \geq 0 (k \in K(\bar{x}))$ that are not all zero, then by taking $\lambda = 0$ we obtain the Fritz John condition.

Case 1, $J \neq 0$.

Since $\nabla h_j(\bar{x}) (j = 1, 2, \ldots, J)$ are linearly independent by assumption (3.1), by the correction theorem of Halkin [9, Theorem F], there exist a neighborhood U of $\bar{x}$ and a continuous mapping ζ from U into X such that $\zeta(\bar{x}) = 0, \nabla\zeta(\bar{x}) = 0$ and

$$h_j(x + \zeta(x)) = \langle \nabla h_j(\bar{x}), x - \bar{x}\rangle \quad \forall x \in U, \ j = 1, 2, \ldots, J. \tag{3.2}$$

We shall now prove that there is no $v \in \mathrm{int}T(\bar{x}, \Omega)$ such that

$$f^\square(\bar{x}; v) < 0, \tag{3.3}$$

$$\langle \nabla g_i(\bar{x}), v\rangle < 0, \quad i \in I(\bar{x}), \tag{3.4}$$

$$\langle \nabla h_j(\bar{x}), v\rangle = 0, \quad j = 1, 2, \ldots, J, \tag{3.5}$$

$$\phi_k^\square(\bar{x}; v) < 0, \quad k \in K(\bar{x}). \tag{3.6}$$

By contradiction, we assume that there exists $v^* \in \mathrm{int}T(\bar{x}, \Omega)$ such that (3.3)–(3.6) hold. Let

$$\theta(t) = \bar{x} + tv^* + \zeta(\bar{x} + tv^*) \ \ \forall t \in [0, 1].$$

Then by virtue of (3.2) and (3.5), for all $\tau \in (0,1]$ small enough, $h_j(\theta(\tau)) = 0$ for all $j = 1, 2, \ldots, J$. Since $\theta(0) = \bar{x}, \nabla\theta(0) = v^*$, by the chain rule,

$$\lim_{t \to 0^+} \frac{g_i(\theta(t)) - g_i(\theta(0))}{t} = \langle \nabla g_i(\bar{x}), v^* \rangle \quad \forall i \in I(\bar{x}).$$

Consequently, by virtue of (3.4),

$$\lim_{t \to 0^+} \frac{g_i(\theta(t)) - g_i(\theta(0))}{t} < 0 \quad \forall i \in I(\bar{x}).$$

That is, for all $\tau \in (0,1]$ small enough,

$$g_i(\theta(\tau)) < 0 \quad \forall i \in I(\bar{x}).$$

Since ϕ_k is Lipschitz near $\bar{x}$,

$$\phi_k^\square(\bar{x}; v^*) = \sup_{w \in X} \limsup_{v' \to v^*, t \downarrow 0} \frac{\phi_k(\bar{x} + t(v' + w)) - \phi_k(\bar{x} + tw)}{t} \quad \forall k \in K(\bar{x}).$$

Consequently, by virtue of (3.6), we have for all $\tau \in (0,1]$ small enough,

$$\frac{\phi_k(\bar{x} + \tau v^* + \zeta(\bar{x} + \tau v^*)) - \phi_k(\bar{x})}{\tau} < 0 \quad \forall k \in K(\bar{x}).$$

That is, for all $\tau \in (0,1]$ small enough,

$$\phi_k(\theta(\tau)) < 0 \quad \forall k \in K(\bar{x}).$$

Similarly since $f^\square(\bar{x}, v) = \langle \nabla f(\bar{x}), v \rangle$ when f is Fréchet differentiable, for all $\tau \in (0,1]$ small enough, $f(\theta(\tau)) < f(\bar{x})$ by virtue of (3.3). By assumption, there exists a hypertangent to Ω at $\bar{x}$. By Rockafellar (see [4, Theorem 2.4.8]), the set of all hypertangents to Ω at $\bar{x}$ coincides with the interior of the Clarke tangent cone to Ω at $\bar{x}$. So for all $\tau \in (0,1]$ small enough,

$$\bar{x} + tv^* + \zeta(\bar{x} + \tau v^*) = \bar{x} + \tau \left[v^* + \frac{\zeta(\bar{x} + \tau v^*)}{\tau} \right] \in \Omega.$$

By the continuity assumptions at $\bar{x}$ for $g_i(i \notin I(\bar{x})), \phi_k(k \notin K(\bar{x}))$, for all $\tau \in (0,1]$ small enough,

$$\begin{aligned} &g_i(\theta(\tau)) < 0 \quad \forall i \notin I(\bar{x}), \\ &\phi_k(\theta(\tau)) < 0 \quad \forall k \notin K(\bar{x}). \end{aligned}$$

Hence there exists $\tau \in (0,1]$ such that

$$\begin{aligned} &f(\theta(\tau)) < f(\bar{x}), \\ &g_i(\theta(\tau)) < 0, \qquad i = 1, 2, \ldots, I, \\ &h_j(\theta(\tau)) = 0, \qquad j = 1, 2, \ldots, J, \\ &\phi_k(\theta(\tau)) < 0, \qquad k = 1, 2, \ldots, K, \\ &\theta(\tau) \in \Omega, \end{aligned}$$

which contradicts the fact that $\bar{x}$ is a local optimal solution of (P).

Since $T(\bar{x}, \Omega)$ is a closed convex cone and $f^{\square}(\bar{x}; v), \phi_k^{\square}(\bar{x}; v)$ are continuous in v (see Proposition 2.3), by virtue of nonexistence of $v \in \text{int}T(\bar{x}, \Omega)$ satisfying (3.3)–(3.6), the nonemptyness of $\text{int}T(\bar{x}, \Omega)$, and Proposition 4.4, $v = 0$ is a solution to the following problem:

$$\begin{aligned} \min \quad & f^{\square}(\bar{x}; v) \\ \text{s.t.} \quad & \langle \nabla g_i(\bar{x}), v \rangle \le 0, \quad i \in I(\bar{x}), \\ & \langle \nabla h_j(\bar{x}), v \rangle = 0, \quad j = 1, 2, \ldots, J, \\ & \phi_k^{\square}(\bar{x}; v) \le 0, \quad k \in K(\bar{x}), \\ & v \in T(\bar{x}, \Omega). \end{aligned}$$

Applying the generalized multiplier rule of Clarke [4, Theorem 6.1.1], there exist scalars $\lambda \ge 0, \alpha_i \ge 0 (i \in I(\bar{x})), \beta_j (j = 1, 2, \ldots, J), \gamma_k \ge 0 (k \in K(\bar{x}))$ not all zero such that

$$\begin{aligned} 0 \in \lambda \partial_v f^{\square}(\bar{x}; 0) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) \\ + \sum_{k \in K(\bar{x})} \gamma_k \partial_v \phi_k^{\square}(\bar{x}; 0) + N(0, T(\bar{x}, \Omega)), \end{aligned}$$

where ∂_v denotes the generalized gradient with respect to v.

By definition, $\xi \in \partial^{\diamond} f(\bar{x})$ if and only if

$$\langle \xi, v \rangle \le f^{\square}(\bar{x}; v) \ \forall v \in X. \tag{3.7}$$

Since $f^{\square}(\bar{x}; v)$ is a convex function of v (see Proposition 2.3) and obviously $f^{\square}(\bar{x}; 0) = 0$, (3.7) holds if and only if $\xi \in \partial_v f^{\square}(\bar{x}; 0)$. Hence, $\partial_v f^{\square}(\bar{x}; 0) = \partial^{\diamond} f(\bar{x})$. Similarly, $\partial_v \phi_k^{\square}(\bar{x}; 0) = \partial^{\diamond} \phi_k(\bar{x})$. Since $\xi \in N(\bar{x}, \Omega)$ if and only if $\langle \xi, v \rangle \le 0$ for all $v \in T(\bar{x}, \Omega)$,

$$N(0, T(\bar{x}, \Omega)) = N(\bar{x}, \Omega).$$

Hence the Fritz John condition holds in this case.

Case 2, $J = 0, I \ne 0$.

We shall now prove that there is no $v \in \text{int}T_{\Omega}(\bar{x})$ such that

$$f^{\square}(\bar{x}; v) < 0, \tag{3.8}$$

$$\langle \nabla g_i(\bar{x}), v \rangle < 0, \quad i \in I(\bar{x}), \tag{3.9}$$

$$\phi_k^{\square}(\bar{x}; v) < 0, \quad k \in K(\bar{x}). \tag{3.10}$$

By contradiction, we assume that there exists $v^* \in \text{int}T(\bar{x}, \Omega)$ such that (3.8)–(3.10) hold. Since $g_i, i \in I(\bar{x})$ are differentiable at $\bar{x}$, for $t > 0$ small enough,

$$g_i(\bar{x} + tv^*) = g_i(\bar{x}) + t\langle \nabla g_i(\bar{x}), v^* \rangle + \alpha_i(\bar{x}, tv^*) t \|v^*\| \quad \forall i \in I(\bar{x}),$$

where $\lim_{t \to 0} \alpha_i(\bar{x}, tv^*) = 0$ for $i \in I(\bar{x})$.

By virtue of (3.9), for $\tau > 0$ small enough,

$$\langle \nabla g_i(\bar{x}), v^* \rangle + \alpha_i(\bar{x}, \tau v^*) \|v^*\| < 0$$

and hence for $\tau > 0$ small enough,

$$g_i(\bar{x} + \tau v^*) < 0, \quad i = 1, 2, \ldots, I.$$

By virtue of (3.10), we have for all $\tau \in (0,1]$ small enough,

$$\frac{\phi_k(\bar{x}+\tau v^*)-\phi_k(\bar{x})}{\tau} < 0 \quad \forall k \in K(\bar{x}).$$

That is, for all $\tau \in (0,1]$ small enough,

$$\phi_k(\bar{x}+\tau v^*) < 0, \quad k = 1,2,\ldots,K.$$

Similarly, we can prove that for all τ small enough,

$$f(\bar{x}+\tau v^*) < f(\bar{x}).$$

Since v^* is a hypertangent to Ω at $\bar{x}$, $\bar{x}+\tau v^* \in \Omega$ for $\tau > 0$ small enough. Hence there exists $\tau > 0$ such that

$$\begin{aligned}
&f(\bar{x}+\tau v^*) < f(\bar{x}),\\
&g_i(\bar{x}+\tau v^*) < 0, \quad i = 1,2,\ldots,I,\\
&\phi_k(\bar{x}+\tau v^*) < 0, \quad k = 1,2,\ldots,K,\\
&\bar{x}+\tau v^* \in \Omega,
\end{aligned}$$

which contradicts the fact that $\bar{x}$ is a local optimal solution of (P).

The remaining proof is similar to Case 1.

4. Constraint qualifications and the KKT conditions.

In this section we introduce four constraint qualifications which ensure the KKT conditions hold and discuss the relationships among them.

The first constraint qualification for the case $L = 0$ follows naturally from the Fritz John necessary optimality condition as in the following proposition.

THEOREM 4.1 (KKT condition for the case $L = 0$ under the (NNAMCQ)). *In addition to the assumptions of Theorem* 1.1*, assume that there is no nonzero abnormal multiplier, i.e.,*

$$\begin{aligned}
(4.1) \qquad &0 \in \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \partial^{\diamond} \phi_k(\bar{x}) + N(\bar{x},\Omega),\\
&\alpha_i \geq 0, \ i \in I(\bar{x}),
\end{aligned}$$

implies that $\alpha_i = 0$ for all $i \in I(\bar{x}), \beta_j = 0$ for all $j = 1,2,\ldots,J, \gamma_k = 0$ for all $k \in K(\bar{x})$. Then $\lambda > 0$ in the conclusion of Theorem 1.1.

Proof. By Theorem 1.1, there exist scalars $\lambda \geq 0, \alpha_i \geq 0 (i \in I(\bar{x})), \beta_j (j = 1,2,\ldots,J), \gamma_k \geq 0 (k \in K(\bar{x}))$ not all zero such that

$$(4.2) \quad 0 \in \lambda \partial^{\diamond} f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \partial^{\diamond} \phi_k(\bar{x}) + N(\bar{x},\Omega).$$

The case $\lambda = 0$ is impossible. Indeed, if $\lambda = 0$ in the above condition, then the inclusion (4.2) coincides with inclusion (4.1). The assumption then rules out this possibility. □

Motivated by the above KKT condition we define the following constraint qualification for the general problem (P).

DEFINITION 4.2. *We say that* (P) *satisfies the* (NNAMCQ) *if*

$$
\begin{aligned}
&0 \in \sum_{i\in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{k\in K(\bar{x})} \gamma_k \partial^{\diamond} \phi_k(\bar{x}) + \sum_{l=1}^{L} \eta_l \partial^{\diamond} \psi_l(\bar{x}) + N(\bar{x}, \Omega),\\
&\alpha_i \geq 0,\ i \in I(\bar{x}), \gamma_k \geq 0,\ k \in K(\bar{x}),
\end{aligned}
$$

implies that $\alpha_i = 0$ *for all* $i \in I(\bar{x}), \beta_j = 0$ *for all* $j = 1, 2, \ldots, J, \gamma_k = 0$ *for all* $k \in K(\bar{x}), \eta_l = 0$ *for all* $l = 1, 2, \ldots, L$.

We now prove that the (NNAMCQ) is closely related to but weaker than the (GMFCQ) defined as follows.

DEFINITION 4.3. *We say that* (P) *satisfies the* (GMFCQ) *at* $\bar{x}$ *if there exists* $d_0 \in \mathrm{int} T(\bar{x}, \Omega)$ *such that*

(i) $\langle \nabla g_i(\bar{x}), d_0 \rangle < 0, \phi_k^{\square}(\bar{x}; d_0) < 0\ \forall i \in I(\bar{x}), k \in K(\bar{x})$,

(ii) $\langle \nabla h_j(\bar{x}), d_0 \rangle = 0, \psi_l^{\square}(\bar{x}; d_0) = 0,\ i = 1, 2, \ldots, J, l = 1, 2, \ldots, L$,

(iii) *for any* $\xi_l \in \partial^{\diamond} \psi_l(\bar{x}), l = 1, \ldots, L$, $\{\nabla h_1(\bar{x}), \ldots, \nabla h_J(\bar{x}), \xi_1, \ldots, \xi_L\}$ *are linearly independent.*

PROPOSITION 4.4. *The* (GMFCQ) *implies* (NNAMCQ). *Under the assumption that* $\mathrm{int} T(\bar{x}, \Omega) \neq \emptyset$, *the* (GMFCQ) *and* (NNAMCQ) *are equivalent.*

Proof. Since the proof of (GMFCQ) implying (NNAMCQ) is exactly similar to the proof in the case $I = J = 0$ [12, Proposition 4.3], we omit the proof.

We now prove the reverse statement under the assumption that $\mathrm{int} T(\bar{x}, \Omega) \neq \emptyset$. Suppose that the (NNAMCQ) holds but not the (GMFCQ). If for some $\xi_l \in \partial^{\diamond} \psi_l(\bar{x}), l = 1, 2, \ldots, L$, $\{\nabla h_1(\bar{x}), \ldots, \nabla h_J(\bar{x}), \xi_1, \ldots, \xi_L\}$ are linearly dependent, then there exist scalars $\beta_j (j = 1, 2, \ldots, J), \eta_l (l = 1, 2, \ldots, L)$ not all zero such that

$$
\begin{aligned}
0 &\in \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{l=1}^{L} \eta_l \partial^{\diamond} \psi_l(\bar{x})\\
&\subseteq \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{l=1}^{L} \eta_l \partial^{\diamond} \psi_l(\bar{x}) + N(\bar{x}, \Omega),
\end{aligned}
$$

which contradicts the fact that there is no nonzero abnormal multiplier for (P). If there is no $d_0 \in \mathrm{int} T(\bar{x}, \Omega)$ satisfying items (i) and (ii), then in the case $I \neq 0$, $d = 0$ must be an optimal solution to the following problem:

$$
\begin{aligned}
\min\quad & \langle \nabla g_{\bar{i}}(\bar{x}), d \rangle\\
\text{s.t.}\quad & \langle \nabla g_i(\bar{x}), d \rangle \leq 0,\ i \in I(\bar{x}) \backslash \{\bar{i}\},\\
& \langle \nabla h_j(\bar{x}), d \rangle = 0,\ j = 1, \ldots, J,\\
& \phi_k^{\square}(\bar{x}; d) \leq 0,\ k \in K(\bar{x}),\\
& \psi_l^{\square}(\bar{x}; d) = 0,\ l = 1, \ldots, L,\\
& d \in T(\bar{x}, \Omega),
\end{aligned}
$$

where $\bar{i} \in I(\bar{x})$ and in the case where $I = 0$ but $K \neq 0$, $d = 0$ must be an optimal solution to the following problem:

$$
\begin{aligned}
\min\quad & \phi_{\bar{k}}^{\square}(\bar{x}; d)\\
\text{s.t.}\quad & \langle \nabla h_j(\bar{x}), d \rangle = 0,\ j = 1, \ldots, J,\\
& \phi_k^{\square}(\bar{x}; d) \leq 0,\ k \in K(\bar{x}) \backslash \{\bar{k}\},
\end{aligned}
$$

$$\psi_l^{\square}(\bar{x}; d) = 0, \ l = 1, \dots, L,$$
$$d \in T(\bar{x}, \Omega),$$

where $\bar{k} \in K(\bar{x})$. Applying the generalized multiplier rule of Clarke completes the proof. □

In Lipschitz optimization, it is well known that the calmness condition is the weakest constraint qualification. We now extend the definition of the calmness condition [4] to our setting.

Definition 4.5 (calmness). *Let $\bar{x}$ be a solution of* (P). (P) *is calm at $\bar{x}$ provided that there exist $\epsilon > 0$ and $\mu > 0$ such that for all $(p, q, u, v) \in \epsilon B_{I+J+K+L}$ and all $x \in \bar{x} + \epsilon B$ satisfying*

$$g(x) + p \leq 0, h(x) + q = 0, \phi(x) + u \leq 0, \psi(x) + v = 0, x \in \Omega \tag{4.3}$$

one has

$$f(\bar{x}) \leq f(x) + \mu \|(p, q, u, v)\|,$$

where B_n denotes the open unit ball in R^n, $g(x) := (g_1(x), g_2(x), \dots, g_I(x))^t$ and $h(x), \phi(x), \psi(x)$ are the vector-valued mappings defined similarly.

We now prove that the calmness condition is also a constraint qualification in our setting. It is interesting to note that unlike the Fritz John type condition (Theorem 1.1) the KKT conditions under either the calmness condition (Theorem 4.2) or the one under the metric regularity condition (Theorems 4.8 and 4.10) hold even for problem (P) with $L \neq 0$. Moreover, under either the calmness condition or the metric regularity condition, the Fréchet differentiable equality constraints do not need to be continuous near the optimal solution.

Theorem 4.6 (KKT condition under calmness CQ). *Let $\bar{x}$ be a solution of* (P). *Suppose that the objective function f is either Fréchet differentiable at $\bar{x}$ or Lipschitz near $\bar{x}$, the constraint functions satisfy assumption* (A), *and there exists a vector that is hypertangent to Ω at $\bar{x}$. If* (P) *is calm at $\bar{x}$, then there exist $\alpha_i \geq 0 (i \in I(\bar{x})), \beta_j (j = 1, 2, \dots, J), \gamma_k \geq 0 (k \in K(\bar{x})), \eta_l (l = 1, 2, \dots, L)$ such that*

$$0 \in \partial^{\diamond} f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \partial^{\diamond} \phi_k(\bar{x})$$
$$+ \sum_{l=1}^{L} \eta_l \partial^{\diamond} \psi_l(\bar{x}) + N(\bar{x}, \Omega).$$

Proof. By the definition of calmness, $(x, p, u) = (\bar{x}, 0, 0)$ is a local solution to

$$\begin{aligned} \min \quad & f(x) + \mu(\|(p, u)\| + \sum_{j=1}^{J} |h_j(x)| + \sum_{l=1}^{L} |\psi_l(x)|) \\ \text{s.t.} \quad & g(x) + p \leq 0, \\ & \phi(x) + u \leq 0, \\ & x \in \Omega. \end{aligned}$$

For a function $g_i(x)$, denote by $g_i^+(x) := \max\{g_i(x), 0\}$. Since $g(x) - g^+(x) \leq 0$,$\phi(x) - \phi^+(x) \leq 0$ and $g(\bar{x}) - g^+(\bar{x}) = 0, \phi(\bar{x}) - \phi^+(\bar{x}) = 0$, taking $p = -g^+(x), u = -\phi^+(x)$, by the calmness condition $\bar{x}$ is also a local solution of the following problem:

$$\begin{aligned}\min \quad & f(x)+\mu\Bigg(\sqrt{I+K}\max\{g_1(x),\dots,g_I(x),\phi_1(x),\dots,\phi_K(x),0\} \\ & \qquad\qquad + \sum_{j=1}^{J}|h_j(x)|+\sum_{l=1}^{L}|\psi_l(x)|\Bigg) \\ \text{s.t.} \quad & x\in\Omega.\end{aligned}$$

That is, $(x,r,s,t)=(\bar{x},0,0,0)$ is a local solution of the following problem:

$$\begin{aligned}\min \quad & f(x)+\mu\left(\sqrt{I+K}r+\sum_{j=1}^{J}s_j+\sum_{l=1}^{L}t_l\right) \\ \text{s.t.} \quad & r\ge g_i(x),\ i=1,2,\dots,I, \\ & r\ge \phi_k(x),\ k=1,2,\dots,K, \\ & r\ge 0, \\ & s_j\ge h_j(x),\ j=1,2,\dots,J, \\ & s_j\ge -h_j(x),\ j=1,2,\dots,J, \\ & t_l\ge \psi_l(x),\ l=1,2,\dots,L, \\ & t_l\ge -\psi_l(x),\ l=1,2,\dots,L, \\ & x\in\Omega.\end{aligned}$$

It is straightforward to verify that the (NNAMCQ) for the above problem is satisfied and the Lagrange multiplier rule with $\lambda=1$ for the original problem follows from applying Theorem 4.1 to the above problem. □

We also extend the notion of metric regularity in smooth and Lipschitz optimization to our setting.

DEFINITION 4.7. *Let C denote the constraint region of* (P) *and $\bar{x}\in C$. C is said to be metrically regular at $\bar{x}$ if there exist positive constants μ,ϵ such that for all $(p,q,u,v)\in\epsilon B$ and all $x\in\bar{x}+\epsilon B$ satisfying* (4.3)*, one has*

$$d_C(x)\le\mu\|(p,q,u,v)\|.$$

As in smooth and Lipschitz optimization, the metric regularity is stronger than the calmness condition in our setting when the objective function is Lipschitz continuous.

THEOREM 4.8 (KKT condition under the metric regularity assumption when the objective function is Lipschitz). *Let $\bar{x}$ be a solution of* (P)*. Assume that the objective function f is Lipschitz near $\bar{x}$, the constraint functions satisfy assumption* (A)*, and there exists a vector that is hypertangent to Ω at $\bar{x}$. If the constraint region is metrically regular at $\bar{x}$, then the KKT condition as stated in the conclusion of Theorem* 4.6 *also holds.*

Proof. Since the objective function f is Lipschitz near $\bar{x}$, by virtue of Proposition 2.7, $\bar{x}$ is a local solution to the following problem:

$$\min \quad f(x)+L_f d_C(x),$$

where L_f denotes the Lipschitz constant of f near $\bar{x}$ and C is the constraint region of (P). By the metric regularity, $(x,p,q,u,v)=(\bar{x},0,0,0,0)$ is a local solution to the following problem:

$$\begin{array}{ll} \min & f(x) + L_f\mu\|(p,q,u,v)\| \\ \text{s.t.} & g(x)+p \le 0, h(x)+q \le 0, \phi(x)+u \le 0, \psi(x)+v=0, x \in \Omega. \end{array}$$

That is, the calmness CQ is satisfied at $\bar{x}$ and hence the conclusion of Theorem 4.6 also holds. □

Unlike the case where the objective function is Lipschitz continuous, when the objective function is only differentiable, the metric regularity of a constraint region may not imply the calmness as illustrated by the following example.

Example. Consider the following optimization problem:

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x = 0, \end{array}$$

where

$$f(x) := \begin{cases} x^2 \sin\frac{1}{x^2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

It is clear that f is differentiable everywhere with

$$f'(x) := \begin{cases} 2x\sin\frac{1}{x^2} - \frac{2}{x}\cos\frac{1}{x^2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Hence f is differentiable at the optimal solution $\bar{x}=0$ but not Lipschitz near $\bar{x}=0$. The constraint region $\{x : x = 0\}$ is metrically regular since the constraint function is linear. However, the problem is not calm at $\bar{x}=0$ since $\bar{x}=0$ is not a solution to the perturbed problem

$$\min \quad f(x) + \mu\|x\|$$

for any $\mu > 0$.

However, although the metric regularity is not stronger than the calmness condition when the objective function is not Lipschitz, it turns out that the metric regularity is still a constraint qualification when the objective function is Fréchet differentiable. In the remainder of this section, we would like to prove the KKT condition under the metric regularity assumption when the objective function is Fréchet differentiable. First we prove the following formula for the Fréchet normal cone to the feasible region C and then we use the result to derive the multiplier rules.

LEMMA 4.9. *Let $\bar{x}$ be a feasible solution of* (P). *Assume that the constraint functions satisfy assumption* (A) *and there exists a vector that is hypertangent to Ω at $\bar{x}$. If C, the feasible region of* (P), *is metrically regular at $\bar{x}$, then*

$$N^F(\bar{x}, C) \subset \left\{ \sum_{i\in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{J} \beta_j \nabla h_j(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \partial^\diamond \phi_k(\bar{x}) \right. \\ \left. + \sum_{l=1}^{L} \eta_l \partial^\diamond \psi_l(\bar{x}) + N(\bar{x}, \Omega) : \alpha_i \ge 0, \gamma_k \ge 0, i \in I(\bar{x}), k \in K(\bar{x}) \right\}.$$

Proof. Let ξ be any element in $N^F_C(\bar{x})$. Then for any $\lambda \downarrow 0$ there exists $\delta > 0$ such that

$$\langle \xi, x' - \bar{x} \rangle \le \lambda \|x' - \bar{x}\| \qquad \forall x' \in C \cap (\bar{x} + \delta B).$$

That is, $\bar{x}$ is a local solution to the following problem:

$$\begin{array}{ll} \min & -\langle \xi, x' \rangle + \lambda \|x' - \bar{x}\| \\ \text{s.t.} & x' \in C. \end{array}$$

Since the objective function of the above problem is Lipschitz continuous, by virtue of Proposition 2.7, $\bar{x}$ is a local solution to the following problem:

$$\min \quad -\langle \xi, x' \rangle + \lambda \|x' - \bar{x}\| + L d_C(x'),$$

where $L \geq \|\xi\| + \lambda$ for all $\lambda > 0$. By the metric regularity, $\bar{x}$ is a local solution to the following problem:

$$\begin{array}{lll} (\mathrm{P}') & \min & -\langle \xi, x' \rangle + \lambda \|x' - \bar{x}\| \\ & & +L\mu(\sqrt{I+K} \max\{g_1(x'), \ldots, g_I(x'), \phi_1(x'), \ldots, \phi_K(x'), 0\} \\ & & +\|h(x')\| + \|\psi(x')\|) \\ & \text{s.t.} & x' \in \Omega. \end{array}$$

Or equivalently, $(x', r, s, t) = (\bar{x}, 0, 0, 0)$ is a local solution to the following problem:

$$\begin{array}{ll} \min & -\langle \xi, x' \rangle + \lambda \|x' - \bar{x}\| + M \left(r + \sum_{j=1}^{J} s_j + \sum_{l=1}^{L} t_l \right) \\ \text{s.t.} & r \geq g_i(x'), \quad i = 1, 2, \ldots, I, \\ & r \geq \phi_k(x'), k = 1, 2, \ldots, K, \\ & r \geq 0, \\ & s_j \geq h_j(x'), \quad j = 1, 2, \ldots, J, \\ & s_j \geq -h_j(x'), \quad j = 1, 2, \ldots, J, \\ & t_l \geq \psi_l(x'), \quad l = 1, 2, \ldots, L, \\ & t_l \geq -\psi_l(x'), \quad l = 1, 2, \ldots, L, \\ & x' \in \Omega, \end{array}$$

with $M = L\mu\sqrt{I+K}$. One can easily verify that the (NNAMCQ) for the above problem is satisfied. Applying Theorem 4.1, there exist $\alpha_i^\lambda (i \in I(\bar{x}))$, $\beta_j^\lambda (j = 1, 2, \ldots, J)$, $\gamma_k^\lambda (k \in K(\bar{x}))$, $\eta_l^\lambda (l = 1, 2, \ldots, L)$, such that

$$\begin{aligned} 0 \in -\xi + \lambda B^* + \sum_{i \in I(\bar{x})} \alpha_i^\lambda \nabla g_i(\bar{x}) + \sum_{j=1}^{L} \beta_j^\lambda \nabla h_j(\bar{x}) \\ + \sum_{k \in K(\bar{x})} \gamma_k^\lambda \partial^\diamond \phi_k(\bar{x}) + \sum_{l=1}^{L} \eta_l^\lambda \partial^\diamond \psi_l(\bar{x}) + N(\bar{x}, \Omega). \end{aligned}$$

Since the (NNAMCQ) holds for problem (P'), $\{(\alpha^\lambda, \beta^\lambda, \gamma^\lambda, \eta^\lambda)\}$ must be bounded. Without loss of generality, we may assume that $\{(\alpha^\lambda, \beta^\lambda, \gamma^\lambda, \eta^\lambda)\}$ converges. The proof of the lemma is completed after taking limits as $\lambda \to 0$, by virtue of the weak* compactness of the Michel–Penot subdifferentials (see Proposition 2.3). □

THEOREM 4.10 (KKT condition under the metric regularity CQ when the objective function is Fréchet differentiable). *Let $\bar{x}$ be a local optimal solution of* (P).

Assume that f is Fréchet differentiable at $\bar{x}$, the constraint functions satisfy assumption (A) *, and there exists a vector that is hypertangent to Ω at $\bar{x}$. If C is metrically regular at $\bar{x}$, then there exist scalars $\alpha_i \geq 0 (i \in I(\bar{x})), \beta_j (j = 1, \ldots, J), \gamma_k \geq 0 (k \in K(\bar{x})), \eta_l (l = 1, 2, \ldots, L)$ such that*

$$0 \in \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^{L} \beta_j \nabla h_j(\bar{x}) + \sum_{k \in K(\bar{x})} \gamma_k \partial^{\diamond} \phi_k(\bar{x}) + \sum_{l=1}^{L} \eta_l \partial^{\diamond} \psi_l(\bar{x}) + N(\bar{x}, \Omega).$$

Proof. Since f is Fréchet differentiable at $\bar{x}$, we have

$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle}{\|x - \bar{x}\|} = 0.$$

Since $\bar{x}$ is a local solution to (P), one has

$$\begin{aligned} \limsup_{x \to \bar{x}, x \in C} \frac{-\langle \nabla f(\bar{x}), x - \bar{x} \rangle}{\|x - \bar{x}\|} &\leq \limsup_{x \to \bar{x}, x \in C} \frac{f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle}{\|x - \bar{x}\|} \\ &\leq \lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle}{\|x - \bar{x}\|} \\ &= 0. \end{aligned}$$

That is, $-\nabla f(\bar{x}) \in N^F(\bar{x}, C)$. The proof of the theorem follows by applying Lemma 4.9, the expression of the Fréchet normal cone to the constraint region. □

5. Multiplier rules for the case $I = J = 0$.

In this section we consider problem (P) in the case where all constraint functions are Lipschitz and the objective function f is Fréchet differentiable. Under this assumption, we derive multiplier rules without requiring the existence of a hypertangent to the abstract constraint set Ω. Note that in Asplund space, the results are sharper since the limiting subdifferentials and the limiting normal cones instead of the Clarke generalized gradients and the Clarke normal cones are used.

First we prove the following formula for the Fréchet normal cone to the feasible region with $I = J = 0$ and then we use the result to derive the multiplier rules.

LEMMA 5.1. *Let $\bar{x}$ be a feasible solution of* (P) *with $I = J = 0$. Assume that $\phi_k (k \in K(\bar{x})), \psi_l (l = 1, 2, \ldots, L)$ are Lipschitz near $\bar{x}$ and $\phi_k (k \notin K(\bar{x}))$ are continuous at $\bar{x}$. If C is metrically regular at $\bar{x}$, then*

$$N^F(\bar{x}, C) \subset \left\{ \sum_{k \in K(\bar{x})} \gamma_k \partial \phi_k(\bar{x}) + \sum_{l=1}^{L} \eta_l \partial \psi_k(\bar{x}) + N(\bar{x}, \Omega) : \gamma_k \geq 0, k \in K(\bar{x}) \right\},$$

where C denotes the feasible region of (P) *with $I = J = 0$.*

Moreover, if X is an Asplund space, then

$$N^F(\bar{x}, C) \subset \left\{ \sum_{k=1}^{K} \gamma_k \hat{\partial} \phi_k(\bar{x}) + \hat{\partial} \langle \eta, \psi \rangle(\bar{x}) + \hat{N}(\bar{x}, \Omega) : \gamma_k \geq 0, k \in K(\bar{x}) \right\}.$$

Proof. Let ξ be any element in $N_C^F(\bar{x})$. Then for any $\lambda_\nu \downarrow 0$, there exists $\delta > 0$ such that

$$\langle \xi, x' - \bar{x} \rangle \le \lambda_\nu \|x' - \bar{x}\| \qquad \forall x' \in C \cap (\bar{x} + \delta B).$$

That is, $\bar{x}$ is a local solution to the following problem:

$$\begin{array}{ll} \min & -\langle \xi, x' \rangle + \lambda_\nu \|x' - \bar{x}\| \\ \text{s.t.} & x' \in C. \end{array}$$

Since the objective function of the above problem is Lipschitz continuous, by virtue of Proposition 2.7, $\bar{x}$ is a local solution to the following problem:

$$\min \quad -\langle \xi, x' \rangle + \lambda_\nu \|x' - \bar{x}\| + L d_C(x'),$$

where $L \ge \|\xi\| + \lambda_\nu$ for all $\nu = 1, 2, \ldots$. By metrical regularity, $\bar{x}$ is a local solution to the following problem:

$$\begin{array}{ll} \min & -\langle \xi, x' \rangle + \lambda_\nu \|x' - \bar{x}\| + L\mu \left(\sqrt{K} \max\limits_{k \in K(\bar{x})} \{\phi_k(x'), 0\} + \|\psi(x')\| \right) \\ \text{s.t.} & x' \in \Omega. \end{array}$$

Or equivalently, $\bar{x}$ is a local solution to the following problem:

$$\min -\langle \xi, x' \rangle + \lambda_\nu \|x' - \bar{x}\| + M \left(\max_{k \in K(\bar{x})} \{\phi_k(x'), 0\} + \|\psi(x')\| \right) + \tilde{L} d_\Omega(x'),$$

with $M = L\mu\sqrt{K}$ and $\tilde{L}$ being the Lipschitz constant of the objective function of the previous optimization problem.

If X is an Asplund space, then by the sum rule for limiting subdifferentials (Proposition 2.5(ii)),

$$0 \in -\xi + \lambda_\nu B^* + M \hat{\partial} \varphi \circ (\phi, \psi)(\bar{x}) + \hat{N}(\bar{x}, \Omega),$$

where $\varphi(u, v) = \max_{k \in K(\bar{x})} \{u_k, 0\} + \|v\|$. By the chain rule,

$$\xi \in \lambda_\nu B^* + M \cup_{(\gamma,\eta) \in \hat{\partial}\varphi(\phi(\bar{x}), \psi(\bar{x}))} \hat{\partial} \langle (\gamma, \eta), (\phi, \psi) \rangle (\bar{x}) + \hat{N}(\bar{x}, \Omega).$$

That is, there exists $(\gamma_\nu, \eta_\nu) \in \hat{\partial}\varphi(\phi(\bar{x}), \psi(\bar{x}))$ such that

$$\xi \in \lambda_\nu B^* + M \hat{\partial} \langle (\gamma_\nu, \eta_\nu), (\phi, \psi) \rangle (\bar{x}) + \hat{N}(\bar{x}, \Omega).$$

Since φ is Lipschitz, by virtue of Proposition 2.3, (γ_ν, η_ν) is a bounded sequence in R^{K+L} and one can assume that $(\gamma_\nu, \eta_\nu) \rightarrow (\gamma, \eta)$ for some $(\gamma, \eta) \in \hat{\partial}\varphi(\phi(\bar{x}), \psi(\bar{x}))$. Hence,

$$\begin{aligned} \xi &\in \lambda_\nu B^* + M \hat{\partial} \langle (\gamma_\nu, \eta_\nu), (\phi, \psi) \rangle (\bar{x}) + \hat{N}(\bar{x}, \Omega) \\ &\subseteq \lambda_\nu B^* + M [\hat{\partial} \langle (\gamma, \eta), (\phi, \psi) \rangle (\bar{x}) + \hat{\partial} \langle (\gamma_\nu, \eta_\nu) - (\gamma, \eta), (\phi, \psi) \rangle (\bar{x})] + \hat{N}(\bar{x}, \Omega) \\ &\subseteq \lambda_\nu B^* + M \hat{\partial} \langle (\gamma, \eta), (\phi, \psi) \rangle (\bar{x}) + M \|(\gamma_\nu, \eta_\nu) - (\gamma, \eta)\| L_{(\phi,\psi)} B^* + \hat{N}(\bar{x}, \Omega). \end{aligned}$$

Taking limits as $\nu \to \infty$, by virtue of the weak* sequential closedness of limiting subdifferentials, one has

$$\xi \in M \hat{\partial} \langle (\gamma, \eta), (\phi, \psi) \rangle (\bar{x}) + \hat{N}(\bar{x}, \Omega)$$

for some

$$\begin{aligned}(\gamma,\eta) &\in \hat{\partial}\varphi(\phi(\bar{x}),\psi(\bar{x}))\\ &= \left\{(\gamma,\eta) : \sum_{k\in K(\bar{x})}\gamma_k = 1, \gamma_k \geq 0 k \in K(\bar{x}), \eta \in B_L\right\}.\end{aligned}$$

The case where X is a general Banach space can be proved similarly. □

THEOREM 5.2 (KKT condition when $I = I = 0$ under the metric regularity CQ). *Let $\bar{x}$ be a local optimal solution of* (P) *with $I = J = 0$. Assume that f is Fréchet differentiable at $\bar{x}$, $\phi_k(k \in K(\bar{x})), \psi_l(l = 1, 2, \ldots, L)$ are Lipschitz near $\bar{x}$ and $\phi_k(k \notin K(\bar{x}))$ are continuous at $\bar{x}$. If C is metrically regular at $\bar{x}$, then there exist $\gamma_k \geq 0(k \in K(\bar{x})), \eta_l(l = 1, 2, \ldots, L)$ such that*

$$0 \in \nabla f(\bar{x}) + \sum_{k\in K(\bar{x})} \gamma_k\partial\phi_k(\bar{x}) + \sum_{l=1}^{L} \eta_l\partial\psi_l(\bar{x}) + N(\bar{x}, \Omega).$$

Moreover, if X is a Asplund space and C is metrically regular at $\bar{x}$, then there exist $\gamma_k \geq 0(k \in K(\bar{x})), \eta_l \in R(l = 1, 2, \ldots, L)$ such that

$$0 \in \nabla f(\bar{x}) + \sum_{k\in K(\bar{x})} \gamma_k\hat{\partial}\phi_k(\bar{x}) + \hat{\partial}\langle\eta, \psi\rangle(\bar{x}) + \hat{N}(\bar{x}, \Omega).$$

Proof. Since f is Fréchet differentiable at $\bar{x}$, as in the proof of Theorem 4.10,

$$-\nabla f(\bar{x}) \in N^F(\bar{x}, C).$$

The proof of the theorem follows by applying Lemma 5.1, the expression of the Fréchet normal cone to the constraint region. □

Remark 2. Sufficient conditions for metrical regularity in the case $I = J = 0$ include the following:

(i) (see [10, Theorem 3].) The constraint region is defined by a system of linear equalities and inequalities, i.e.,

$$C := \{x \in X : \langle x_k^*, x\rangle = 0, k = 1, \ldots, K, \langle y_l^*, x\rangle \leq 0 l = 1, \ldots, L\}$$

for some $x_k^* \in X^*(k = 1, \ldots, K), y_l^* \in X^*(l = 1, \ldots, L)$.

(ii) In Banach space [4, Theorem 6.6.1], the (NNAMCQ) in the Clarke generalized gradient form is satisfied, i.e.,

$$\begin{aligned}&0 \in \sum_{k\in K(\bar{x})} \gamma_k\partial\phi_k(\bar{x}) + \sum_{l=1}^{L} \eta_l\partial\psi_k(\bar{x}) + N(\bar{x}, \Omega),\\ &\gamma_k \geq 0 \ \forall k \in K(\bar{x})\end{aligned}$$

implies that $\gamma_k = 0, k \in K(\bar{x}), \eta_l = 0, l = 1, 2, \ldots, L$. In Asplund space [18, Corollary 6.2], the (NNAMCQ) in the limiting subdifferential form is satisfied, i.e.,

$$\begin{aligned}&0 \in \sum_{k\in K(\bar{x})} \gamma_k\hat{\partial}\phi_k(\bar{x}) + \hat{\partial}\langle\eta, \psi\rangle(\bar{x}) + \hat{N}(\bar{x}, \Omega),\\ &\gamma_k \geq 0 \ \forall k \in K(\bar{x})\end{aligned}$$

implies that $\gamma_k = 0, k \in K(\bar{x}), \eta_l = 0, l = 1, 2, \ldots, L$.

THEOREM 5.3 (KKT condition when $I = J = 0$ under the calmness CQ). *Let $\bar{x}$ be a local optimal solution of* (P) *with $I = J = 0$. Assume that f is Fréchet differentiable at $\bar{x}$, $\phi_k(k \in K(\bar{x})), \psi_l(l = 1, 2, \ldots, L)$ are Lipschitz near $\bar{x}$ and $\phi_k(k \notin K(\bar{x}))$ are continuous at $\bar{x}$. If* (P) *is calm at $\bar{x}$, then the conclusions of Theorem* 5.2 *hold.*

Proof. By the definition of calmness, $(x, u) = (\bar{x}, 0)$ is a local solution to

$$\begin{aligned} \min \quad & f(x) + \mu(\|u\| + \|\psi(x)\|) \\ \text{s.t.} \quad & \phi(x) + u \le 0, \\ & x \in \Omega. \end{aligned}$$

Since $\phi(x) - \phi^+(x) \le 0$ and $\phi(\bar{x}) - \phi^+(\bar{x}) = 0$, $\bar{x}$ is also a local solution of the following problem:

$$\begin{aligned} \min \quad & f(x) + \mu\left(\sqrt{K} \max\{\phi_1(x), \ldots, \phi_K(x), 0\} + \|\psi(x)\|\right) \\ \text{s.t.} \quad & x \in \Omega. \end{aligned}$$

Case 1. X is a general Banach space. It is easy to see that $(x, r, s) = (\bar{x}, 0, 0)$ is a local solution to the following problem:

$$\begin{aligned} \min \quad & f(x) + \mu\left(\sqrt{K} r + \sum_{l=1}^{L} s_l\right) \\ \text{s.t.} \quad & r \ge \phi_k(x), \ k = 1, \ldots, K, \\ & r \ge 0, \\ & s_l \ge \psi_l(x), \quad l = 1, 2, \ldots, L, \\ & s_l \ge -\psi_l(x), \quad l = 1, 2, \ldots, L, \\ & x \in \Omega. \end{aligned}$$

It is straightforward to verify that the (NNAMCQ) for the above problem is satisfied and the KKT condition follows from Theorem 5.2 and in Remark 2(ii).

Case 2. X is an Asplund space. Equivalently, $\bar{x}$ is a local solution to the following problem:

$$\min \quad f(x) + \mu\left(\sqrt{K} \max_{k \in K(\bar{x})}\{\phi_k(x), 0\} + \|\psi(x)\|\right) + \delta_\Omega(x),$$

where $\delta_\Omega(x)$ is the indicator function of a set Ω defined by

$$\delta_\Omega(x) := \begin{cases} 0 & \text{if } x \in \Omega, \\ +\infty & \text{if } x \notin \Omega. \end{cases}$$

Since f is Fréchet differentiable and $G(x) := \mu(\sqrt{K} \max_{k \in K(\bar{x})}\{\phi_K(x), 0\} + \|\psi(x)\|)$ is Lipschitz near $\bar{x}$, one has

$$\begin{aligned} 0 &\in \nabla f(\bar{x}) + \partial^F(G + \delta_\Omega)(\bar{x}) \quad \text{(by Proposition 2.5(i))} \\ &\subseteq \nabla f(\bar{x}) + \hat{\partial}(G + \delta_\Omega)(\bar{x}) \\ &\subseteq \nabla f(\bar{x}) + \hat{\partial} G(\bar{x}) + \hat{N}(\bar{x}, \Omega) \quad \text{(Proposition 2.5(ii))}. \end{aligned}$$

The remaining proof follows by using the sum rules and the chain rules as in the proof of Lemma 5.1. □

Proof of Theorem 1.2. Suppose X is a Banach space. If the (NNAMCQ) in the Clarke generalized gradient form does not hold, then the Fritz John condition holds with $\lambda = 0$. Otherwise if the (NNAMCQ) in the Clarke generalized gradient form holds, then by Remark 2 and Theorem 5.2, the Fritz John condition holds with $\lambda = 1$.

Similarly suppose that X is an Asplund space. If the (NNAMCQ) in the limiting subdifferential form as in Remark 2 does not hold, then the required Fritz John condition holds with $\lambda = 0$. Otherwise if the (NNAMCQ) in the limiting subdifferential form holds, then by Remark 2 and Theorem 5.2, the required Fritz John condition holds with $\lambda = 1$.

Acknowledgments. The author would like to thank Jay Treiman for his suggestions on replacing the Clarke generalized gradient by the Michel–Penot subdifferential in Theorem 1.1 and those in section 4 in an earlier version.

REFERENCES

[1] M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming Theory and Algorithms*, 2nd ed., John Wiley & Sons, New York, 1993.

[2] J.M. Borwein and Q. Zhu, *A survey of subdifferential calculus with applications*, Nonlinear Anal., 38 (1999), pp. 687–773.

[3] F.H. Clarke, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.

[4] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[5] F.H. Clarke, Yu. S. Ledyaev, R.J. Stern, and P.R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer, New York, 1998.

[6] F.H. Clarke, R. J. Stern, and P.R. Wolenski, *Subgradient criteria for monotonicity, the Lipschitz condition, and convexity*, Canad. J. Math., 45 (1993), pp. 1167–1183.

[7] S. Di, *Classical optimality conditions under weaker assumptions*, SIAM J. Optim., 6 (1996), pp. 178–197.

[8] L.A. Fernández, *On the limits of the Lagrange multiplier rule*, SIAM Rev., 39 (1997), pp. 292–297.

[9] H. Halkin, *Implicit functions and optimization problems without continuous differentiability of the data*, SIAM J. Control, 12 (1974), pp. 229–236.

[10] A.D. Ioffe, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.

[11] A.D. Ioffe, *A Lagrange multiplier rule with small convex-valued subdifferentials for nonsmooth problems of mathematical programming involving equality and nonfunctional constraints*, Math. Programming, 58 (1993), pp. 137–145.

[12] A. Jourani, *Constraint qualifications and Lagrange multipliers in nondifferentiable programming problems*, J. Optim. Theory Appl., 81 (1994), pp. 533–548.

[13] P.D. Loewen, *Optimal Control via Nonsmooth Analysis*, CRM Proc. Lecture Notes 2, AMS, Providence, RI, 1993.

[14] O.L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969; reprint, SIAM, Philadelphia, 1994.

[15] P. Michel and J.-P. Penot, *Calcul sous-différentiel pour des fonctions lipschitziennes et non lipschitziennes*, C.R. Acad. Sci. Paris Sér. I Math, 12 (1984), pp. 269–272.

[16] B.S. Mordukhovich, *On necessary conditions for an extremum in nonsmooth optimization*, Soviet Math. Dokl., 283 (1985), pp. 215–220.

[17] B.S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.

[18] B.S. Mordukhovich and Y. Shao, *Stability of set-valued mappings in infinite dimensions: Point criteria and applications*, SIAM J. Control Optim., 35 (1997), pp. 285–314.

[19] J.S. Treiman, *Shrinking generalized gradients*, Nonlinear Anal., 12 (1988), pp. 1429–1450.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1461–1484

THE LOCAL SOLVABILITY OF A HAMILTON–JACOBI–BELLMAN PDE AROUND A NONHYPERBOLIC CRITICAL POINT*

ARTHUR J. KRENER†

Abstract. We show the existence of a local solution to a Hamilton–Jacobi–Bellman (HJB) PDE around a critical point where the corresponding Hamiltonian ODE is not hyperbolic, i.e., it has eigenvalues on the imaginary axis. Such problems arise in nonlinear regulation, disturbance rejection, gain scheduling, and linear parameter varying control. The proof is based on an extension of the center manifold theorem due to Aulbach, Flockerzi, and Knobloch. The method is easily extended to the Hamilton–Jacobi–Isaacs (HJI) PDE. Software is available on the web to compute local approximtate solutions of HJB and HJI PDEs.

Key words. parametrized optimal control, nonlinear regulation, nonlinear disturbance rejection, gain scheduling, linear parameter varying control, H_∞ regulation

AMS subject classification. 49C05

PII. S0363012999361081

1. Introduction.

Consider a smooth optimal control problem of minimizing

$$\frac{1}{2}\int_0^\infty \|e\|^2 + \|u\|^2 dt \tag{1.1}$$

subject to

$$\begin{array}{rcl} \dot{x} & = & f(x,u) = Ax + Bu + O(x,u)^2, \\ e & = & h(x,u) = Cx + Du + O(x,u)^2. \end{array} \tag{1.2}$$

The optimal cost $\pi(x)$ and the optimal feedback $\kappa(x)$ satisfy the Hamiliton–Jacobi–Bellman (HJB) PDE

$$0 = \frac{\partial \pi}{\partial x}(x) f(x, \kappa(x)) + l(x, \kappa(x)), \tag{1.3}$$

$$\kappa(x) = \text{argmin}_u \left\{ \frac{\partial \pi}{\partial x}(x) f(x,u) + l(x,u) \right\}, \tag{1.4}$$

where

$$\begin{aligned} l(x,u) &= \frac{1}{2}(\|e\|^2 + \|u\|^2) \\ &= \frac{1}{2}\left(x'Qx + 2x'Su + u'Ru\right) + O(x,u)^3 \end{aligned} \tag{1.5}$$

and $Q = C'C,\ S = C'D,\ R = I + D'D$.

The HJB PDE may not admit a smooth global solution but under suitable conditions there does exist a viscosity solution. We refer the reader to [10], [11] for details. It is well known [22] that the HJB PDE admits a smooth solution locally around

*Received by the editors September 9, 1999; accepted for publication (in revised form) August 15, 2000; published electronically January 19, 2001. A preliminary version of this paper appeared as *The existence of optimal regulators*, in Proceedings of the 1998 CDC, Tampa, FL, 1998, pp. 3081–3086. This research was supported in part by AFOSR-49620-95-1-0409.

http://www.siam.org/journals/sicon/39-5/36108.html

†Department of Mathematics and Institute of Theoretical Dynamics, University of California, Davis, CA 95616-8633 (ajkrener@ucdavis.edu).

$x = 0$ under suitable conditions. We briefly review these conditions and the method of proof.

Consider first the linear quadratic part of the above problem, minimizing

$$\tfrac{1}{2}\int_0^\infty (x'Qx + 2x'Su + u'Ru) \tag{1.6}$$

subject to

$$\dot{x} \quad = \quad Ax + Bu. \tag{1.7}$$

If they exist, the optimal cost is quadratic, $\frac{1}{2}x'Px$, and the optimal feedback is linear, $u = Kx$. Moreover, P satisfies the algebraic Riccati equation

$$0 = A'P + PA + Q - (PB + S)R^{-1}(PB + S)'$$

and

$$K = -R^{-1}(PB + S)'.$$

It is well known [3] that if the pair A, B is stabilizable and the pair C, A is detectable, then there is a unique nonnegative definite solution to the algebraic Riccati equation and the resulting feedback is asymptotically stabilizing. If C, A is observable, then P is positive definite.

The $2n$ dimensional Hamiltonian system associated with this problem is linear,

$$\begin{bmatrix} \dot{x} \\ \dot{\lambda}' \end{bmatrix} = \mathbf{H} \begin{bmatrix} x \\ \lambda' \end{bmatrix}, \tag{1.8}$$

where

$$\mathbf{H} = \begin{bmatrix} A - BR^{-1}S' & -BR^{-1}B' \\ -Q + SR^{-1}S' & -A' + SR^{-1}B' \end{bmatrix}. \tag{1.9}$$

If A, B is stabilizable and C, A is detectable, then this system is hyperbolic, i.e., none of the eigenvalues of $\mathbf{H}$ lie on the imaginary axis. Since $\mathbf{H}$ is Hamiltonian, this implies that n eigenvalues lie in the open left half plane and n eigenvalues lie in the open right half plane. In fact, the n dimensional stable subspace of $\mathbf{H}$ is the graph of the gradient of the unique nonnegative definite solution to the algebraic Riccati equation,

$$\lambda = x'P.$$

In other words, the stable subspace is the span of the columns of

$$\begin{bmatrix} I \\ P \end{bmatrix}. \tag{1.10}$$

We return to the nonlinear problem (1.1), (1.2). The associated Hamiltonian system is nonlinear,

$$\begin{aligned} \dot{x}' &= \frac{\partial H}{\partial \lambda}(\lambda, x, \kappa(\lambda, x)), \\ \dot{\lambda} &= -\frac{\partial H}{\partial x}(\lambda, x, \kappa(\lambda, x)), \end{aligned} \tag{1.11}$$

where the Hamiltonian is

$$\begin{array}{rcl} H(\lambda,x,u) & = & \lambda f(x,u)+l(x,u) \\ & = & \lambda\,(Ax+Bu) \\ & & +\frac{1}{2}\,(x'Qx+2x'Su+u'Ru) \\ & & +O(\lambda,x,u)^3 \end{array} \tag{1.12}$$

and the optimal control as determined by the Pontryagin maximum principle satisfies

$$u=\kappa(\lambda,x)=\mathrm{argmin}_u H(\lambda,x,u). \tag{1.13}$$

The linearization of this system (1.11) around the origin is the linear Hamiltonian system above (1.8). Hence if A,B is stabilizable and C,A is detectable, then there is an n dimensional local stable manifold around the origin [15]. Moreover, this submanifold is the graph of the gradient of the optimal cost,

$$\lambda=\frac{\partial\pi}{\partial x}(x).$$

Hence the HJB PDE (1.3), (1.4) is locally solvable. The details can be found in Lukes [22].

In this paper we show that the HJB PDE is locally solvable in certain situations where the linear part of the system is not stabilizable or not detectable. Such systems arise naturally in the problems of nonlinear regulation, disturbance rejection, gain scheduling, and linear parameter varying control. In these problems there tends to be certain modes of the linearized system at the origin which are neutrally stable, uncontrollable, and/or unobservable. But fortunately these modes tend to be sufficiently separated from the others or can be made so by feedforward from the exosystem state so that an extension of the stable manifold theorem can be used to prove the local solvability of the HJB PDE. We proved this extension, which we call the stable and partial center manifold theorem, only to learn that a similar result had already been shown by Aulbach, Flockerzi, and Knobloch [6] and Aulbach and Flockerzi [7]. Since their result is not well known and may not be readily available, we include our proof. We also prove an additional result that the Taylor series of the stable and partial center manifold can be computed term-by-term. This justifies the term-by-term solution of the HJB PDE in these situations in the spirit of Al'brecht [2].

The rest of the paper is organized as follows. In the next section we introduce the problems of nonlinear regulation, disturbance rejection, gain scheduling, and linear parameter varying control and discuss when they can be transformed so that a local solution of the HJB equation exists. In section 4 we state and prove two theorems, the stable and partial center manifold theorem and a theorem on its term-by-term development. In section 5 we show how these theorems can be used to prove the local solvability of the HJB PDE and to construct approximate solutions. In the last section we discuss the local solvability of HJI PDEs and how they arise in H_∞ extensions of the above problems.

2. Nonlinear regulation and related problems.

Consider a smooth nonlinear plant

$$\begin{array}{rcl} \dot{x} & = & f(x,u,\bar{x}) \\ & = & Ax+Bu+F\bar{x} \\ & & +f^{[2]}(x,u,\bar{x})+O(x,u,\bar{x})^3, \\ e & = & h(x,u,\bar{x}) \\ & = & Cx+Du+H\bar{x} \\ & & +h^{[2]}(x,u,\bar{x})+O(x,u,\bar{x})^3 \end{array} \tag{2.1}$$

which is perturbed by a smooth nonlinear exosystem

$$\begin{aligned} \dot{\bar{x}} &= \bar{f}(\bar{x}) \\ &= \bar{A}\bar{x} + \bar{f}^{[2]}(\bar{x}) + O(\bar{x})^3, \end{aligned} \tag{2.2}$$

where superscript $[d]$ denotes terms composed of homogeneous polynomials of degree d. The dimensions of $x, u, \bar{x}, e$ are $n, m, \bar{n}, p$, respectively.

The goal of regulation is to use a combination of feedforward and feedback control $u = \alpha(x, \bar{x})$ so that the output of the plant asymptotically goes to 0,

$$e(t) \longrightarrow 0$$

for every $x(0)$, $\bar{x}(0)$. The plant should also be internally stable.

The exosystem could be a system whose output we wish the plant to track (regulation), a noise source whose disturbances we wish the plant to reject (disturbance rejection), or static and/or dynamic parameters to be used for scheduling the controller of the plant (gain scheduling).

A linear parameter varying (LPV) system,

$$\begin{aligned} \dot{x} &= A(\bar{x})x + B(\bar{x})u \\ &= \left(A^{[0]} + A^{[1]}(\bar{x}) + \cdots\right) x + \left(B^{[0]} + B^{[1]}(\bar{x}) + \cdots\right) u, \\ e &= C(\bar{x})x + D(\bar{x})u \\ &= \left(C^{[0]} + C^{[1]}(\bar{x}) + \cdots\right) x + \left(D^{[0]} + D^{[1]}(\bar{x}) + \cdots\right) u, \end{aligned}$$

falls into the last category.

We make the reasonable assumptions that the linear part of the plant is stabilizable and detectable when $\bar{x} = 0$ and the linear part of the exosystem is stable. Most plants are designed to be linearly stabilizable and detectable. If the exosystem was unstable then it would probably be impossible to overcome its effect on the plant. The combined system (2.1), (2.2) is not linearly stabilizable because we have no control over the stable modes of the exosystem and some of these might not be linearly detectable.

The solution of the regulator problem is in two steps. The first is to use feedforward from the exosystem state to insure exact tracking when the initial conditions of the plant and the exosystem permit this. We are assuming that the state of the exosystem is available for measurement. The more general case, when it is not measurable, was treated in [14] and [8]. Even when the state of the exosystem is not measurable, one must find the feedforward control law that would insure exact tracking if it were measurable. We discuss only the case when x, $\bar{x}$ are measurable.

The linear version of the problem was solved by Francis [12] and its nonlinear generalization is due to Isidori and Byrnes [14]. One seeks $\theta(\bar{x})$, $\beta(\bar{x})$ satisfying the Francis–Byrnes–Isidori (FBI) PDE

$$\begin{aligned} f(\theta(\bar{x}), \beta(\bar{x}), \bar{x}) &= \frac{\partial \theta}{\partial \bar{x}}(\bar{x}) \bar{f}(\bar{x}), \\ h(\theta(\bar{x}), \beta(\bar{x}), \bar{x}) &= 0. \end{aligned} \tag{2.3}$$

If the FBI PDE is solvable, then the control $u = \beta(\bar{x})$ makes $x = \theta(\bar{x})$ an invariant manifold of the combined system consisting of plant and exosystem. On this manifold, exact tracking occurs, $e = 0$.

One can attempt to solve the FBI equations term-by-term. Suppose

$$\theta(\bar{x}) = T\bar{x} + \theta^{[2]}(\bar{x}) + O(\bar{x})^3,$$
$$\beta(\bar{x}) = L\bar{x} + \beta^{[2]}(\bar{x}) + O(\bar{x})^3.$$

The linear part of the FBI equations are the Francis equations

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} T \\ L \end{bmatrix} - \begin{bmatrix} T \\ 0 \end{bmatrix} \bar{A} = - \begin{bmatrix} F \\ H \end{bmatrix}.$$

These equations are solvable for any F, H iff no output zero of the plant is a pole of the exosystem [13, 16, 17]. In other words, the exosystem should not excite those frequencies that the plant cannot produce.

The output zeros of the plant are those complex numbers s for which there exist complex n and p row vectors ξ and ζ such that

$$\begin{bmatrix} \xi & \zeta \end{bmatrix} \begin{bmatrix} A - sI & B \\ C & D \end{bmatrix} = \begin{bmatrix} 0 & 0 \end{bmatrix}.$$

There may be a finite or infinite number of output zeros. For example, if $m = p$, then there are either n zeros or every s is a zero. The poles $\lambda_1, \dots, \lambda_{\bar{n}}$ of the exosystem are the eigenvalues of $\bar{A}$.

If there is a resonance between a pole and zero, the equations will still be solvable for some F, H. The solvability depends on the direction ξ, ζ of the zero and the eigenvector of the pole.

The higher degree equations are linear and depend on the solutions of the lower degree equations. They are solvable for arbitrary higher degree terms iff the harmonics of the exosystem don't resonate with the zeros of the plant [13, 16, 17].

For example, the degree two equations are

$$\begin{aligned} &A\theta^{[2]}(\bar{x}) + B\beta^{[2]}(\bar{x}) - \frac{\partial \theta}{\partial \bar{x}}(\bar{x}) \left(\bar{A}\bar{x}\right) \\ &\qquad = -f^{[2]}(T\bar{x}, L\bar{x}, \bar{x}) + T\bar{f}^{[2]}(\bar{x}), \\ &C\theta^{[2]}(\bar{x}) + D\beta^{[2]}(\bar{x}) \\ &\qquad = -h^{[2]}(T\bar{x}, L\bar{x}, \bar{x}). \end{aligned}$$

These are solvable for arbitrary $f^{[2]}, h^{[2]}$ iff no output zero of the plant equals the sum of two poles of the exosystem, $\lambda_k \neq s_i + s_j$. If there is a resonance, they are solvable for some $f^{[2]}, h^{[2]}$. Matlab-based software is available on the web to compute the solution of the FBI PDE to any degree using the function `fbi.m` of the Nonlinear Systems Toolbox [20].

Now suppose that the FBI equations have been solved. The second step is to use additional feedforward and feedback to insure that the closed loop system converges to the tracking manifold $x = \theta(\bar{x})$ where $e = 0$. This can be achieved locally by linear pole placement techniques [14], but an alternative approach is to use optimal control methods to achieve a nonlinear solution [16, 17]. Define transverse coordinates z, v by

$$\begin{aligned} z &= x - \theta(\bar{x}) = x - T\bar{x} - \theta^{[2]}(\bar{x}) + O(\bar{x})^3, \\ v &= u - \beta(\bar{x}) = u - L\bar{x} - \beta^{[2]}(\bar{x}) + O(\bar{x})^3. \end{aligned} \tag{2.4}$$

In these coordinates the plant and exosystem are of the form

$$\begin{aligned} \dot{z} &= \tilde{f}(z,v,\bar{x}) = Ax + Bu + \tilde{f}^{[2]}(z,v,\bar{x}) + O(z,v,\bar{x})^3, \\ \dot{\bar{x}} &= \bar{f}(\bar{x}) = \bar{A}\bar{x} + \bar{f}^{[2]}(\bar{x}) + O(\bar{x})^3, \\ e &= \tilde{h}(z,v,\bar{x}) = Cz + Dv + \tilde{h}^{[2]}(z,v,\bar{x}) + O(z,v,\bar{x})^3, \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} \tilde{f}(z,v,\bar{x}) &= f(z+\theta(\bar{x}), v+\beta(\bar{x}), \bar{x}) - f(\theta(\bar{x}), \beta(\bar{x}), \bar{x}), \\ \tilde{h}(z,v,\bar{x}) &= h(z+\theta(\bar{x}), v+\beta(\bar{x}), \bar{x}). \end{aligned} \tag{2.6}$$

Notice that the linear part of the z dynamics and the linear part of the output are unaffected by $\bar{x}$. Recall we have assumed that the linear part of the plant is stabilizable and detectable and the linear part of the exosystem is neutrally stable. Furthermore,

$$\begin{aligned} \tilde{f}(0,0,\bar{x}) &= f(\theta(\bar{x}), \beta(\bar{x}), \bar{x}) - f(\theta(\bar{x}), \beta(\bar{x}), \bar{x})) = 0, \\ \tilde{h}(0,0,\bar{x}) &= h(\theta(\bar{x}), \beta(\bar{x}), \bar{x}) = 0. \end{aligned} \tag{2.7}$$

A stabilizing feedback can be found by minimizing

$$\tfrac{1}{2}\textstyle\int_0^\infty \|e\|^2 + \|v\|^2 dt \tag{2.8}$$

subject to the dynamics (2.5). Other cost criterions l can be used as long as they satisfy (2.11). In particular a cost criterion like $x'Qx + u'Ru$ should not be used as then (2.11) will not hold. Intuitively, one should not cost the part of the state and the control that are necessary to achieve exact tracking. Even in the linear case, there is considerable confusion on this point, e.g., [3].

Let $\pi(z,\bar{x})$ denote the optimal cost and $\gamma(z,\bar{x})$ the optimal feedback; then π, γ satisfy the HJB PDE

$$\begin{aligned} 0 &= \frac{\partial \pi}{\partial z}(z,\bar{x})\tilde{f}(z,\gamma(z,\bar{x}),\bar{x}) + \frac{\partial \pi}{\partial \bar{x}}(z,\bar{x})\bar{f}(\bar{x}) \\ &\quad + l(z,\gamma(z,\bar{x}),\bar{x}), \\ 0 &= \frac{\partial \pi}{\partial z}(z,\bar{x})\frac{\partial \tilde{f}}{\partial v}(z,\gamma(z,\bar{x}),\bar{x}) + \frac{\partial l}{\partial v}(z,\gamma(z,\bar{x}),\bar{x}), \end{aligned} \tag{2.9}$$

where

$$\begin{aligned} l(z,v,\bar{x}) &= \tfrac{1}{2}(\|e\|^2 + \|v\|^2) \\ &= \tfrac{1}{2}\left(z'Qz + 2z'Sv + v'Rv\right) \\ &\quad + l^{[3]}(z,v,\bar{x}) + O(z,v,\bar{x})^4 \end{aligned} \tag{2.10}$$

for the matrices $Q = C'C,\ S = C'D,\ R = I + D'D$, and some cubic polynomial $l^{[3]}$.

By generalizing Al'brecht's method [2], we can solve the HJB PDE term-by-term [16]. Since

$$\begin{aligned} \tilde{f}(z,v,\bar{x}) &= O(z,v), \\ \tilde{h}(z,v,\bar{x}) &= O(z,v), \\ l(z,v,\bar{x}) &= O(z,v)^2, \end{aligned} \tag{2.11}$$

we expect that

$$\begin{array}{rcl}\pi(z,\bar{x}) & = & O(z)^2,\\ \gamma(z,\bar{x}) & = & O(z).\end{array} \tag{2.12}$$

In particular, we expect that

$$\begin{aligned}\pi(z,\bar{x}) &= \frac{1}{2}z'Pz + \pi^{[3]}(z,\bar{x}) + O(z,\bar{x})^4,\\ \gamma(z,\bar{x}) &= Kz + \gamma^{[2]}(z,\bar{x}) + O(z,\bar{x})^3.\end{aligned}$$

The lowest degree terms in the HJB equations are the familiar Riccati equation and the formula for the optimal linear feedback

$$\begin{array}{rcl}0 & = & A'P + PA + Q - (PB+S)R^{-1}(PB+S)',\\ K & = & -R^{-1}(PB+S)'.\end{array} \tag{2.13}$$

At each higher degree $d > 1$, the equations are linear in the unknowns $\pi^{[d+1]}$, $\gamma^{[d]}$ and depend on the lower order terms of the solution. They are solvable if the linear part of the plant is stabilizable and the linear part of the exosystem is stable. For example, to find the next terms $\pi^{[3]}(z,\bar{x})$, $\gamma^{[2]}(z,\bar{x})$, one plugs the first two terms of π, γ into HJB equations and collects the next terms (degree 3 from the first HJB equation and degree 2 from the second HJB equation)

$$\begin{aligned}0 = {}&\frac{\partial\pi^{[3]}}{\partial z}(z,\bar{x})(A+BK)z + \frac{\partial\pi^{[3]}}{\partial\bar{x}}(z,\bar{x})(\bar{A}\bar{x})\\ &+z'P\tilde{f}^{[2]}(z,Kz,\bar{x}) + l^{[3]}(z,Kz,\bar{x}),\\ 0 = {}&\frac{\partial\pi^{[3]}}{\partial z}(z,\bar{x})B + z'P\frac{\partial\tilde{f}^{[2]}}{\partial v}(z,Kz,\bar{x})\\ &+\gamma^{[2]}(z,\bar{x})'R + \frac{\partial l^{[3]}}{\partial v}(z,Kz,\bar{x}).\end{aligned} \tag{2.14}$$

Notice that the first equation involves only $\pi^{[3]}$, the other unknown $\gamma^{[2]}$ does not appear. This equation is solvable if $A+BK$ is asymptotically stable and $\bar{A}$ is stable. This follows from the fact that the mapping

$$\pi^{[3]}(z,\bar{x}) \mapsto \frac{\partial\pi^{[3]}}{\partial z}(z,\bar{x})(A+BK)z + \frac{\partial\pi^{[3]}}{\partial\bar{x}}(z,\bar{x})(\bar{A}\bar{x})$$

is a linear operator on cubic polynomials. It is not hard to see that its eigenvalues are the sum of three eigenvalues of $A+BK$ and $\bar{A}$. The operator restricts to a linear operator on the subspace of $\pi^{[3]}(z,\bar{x})$ satisfying (2.12), where its eigenvalues are the sum of three eigenvalues of $A+BK$ or the sum of two eigenvalues of $A+BK$ and one eigenvalue of $\bar{A}$. Since the eigenvalues of $A+BK$ are in the open left half plane and those of $\bar{A}$ are in the closed left half plane, the restricted operator is invertible and the first equation of (2.14) is always solvable. We discuss this further in the proof of Theorem 4.2.

Given the solution $\pi^{[3]}$, we can then solve the second equation for $\gamma^{[2]}$

$$\gamma^{[2]}(z,\bar{x}) = -R^{-1}\left(\frac{\partial\pi^{[3]}}{\partial z}(z,\bar{x})B + z'P\frac{\partial\tilde{f}^{[2]}}{\partial v}(z,Kz,\bar{x}) + \frac{\partial l^{[3]}}{\partial v}(z,Kz,\bar{x})\right)'.$$

The higher degree terms are found in a similar fashion. Matlab-based software is available on the web to compute the solution of the HJB PDE to any degree using the function `hjb.m` in the Nonlinear Systems Toolbox [20]. If one wants to solve the FBI PDE and then the HJB PDE in the transverse coordinates, use the function `mdl_mtch.m`.

Given the solutions of the FBI and HJB equations, the desired feedforward/feedback is

$$\begin{aligned} u &= \alpha(x,\bar{x}) \\ &= \beta(\bar{x}) + \gamma(x-\theta(\bar{x}),\bar{x}). \end{aligned}$$

Of course the above discussion is formal. We shall show using results from [6], [7] that the HJB PDE (2.9) is locally solvable. Furthermore, its Taylor series expansion can be computed term-by-term as described above. To do so we shall use an invariant manifold theorem that we shall discuss in the next section. In section 4 we use this theorem to show the local existence of the solution to the HJB equation (2.9).

Suppose one has computed approximate solutions to the FBI PDE up to degree d and the HJB PDE up to degree $d+1$, and one has the desired $\alpha(x,\bar{x})$ up to degree d. Despite the formal nature of these, one can explicitly verify where it gives the desired solution. The function $\pi(x-\theta(\bar{x},\bar{x}))$ is a potential Lyapunov function for the approximate tracking manifold $x=\theta(\bar{x})$ on which the error $e=O(\bar{x})^{d+1}$. Using this and the true closed loop dynamics, one can estimate the basin of attraction of the approximate tracking manifold.

3. Stable and partial center manifold theorem.

The following theorem was proven by Aulbach, Flockerzi, and Knobloch [6] and Aulbach and Flockerzi [7]. We were unaware of their work and suspected that such a theorem must hold because of the formal discussion of the last section. We present our independent proof because [6] and [7] are not widely known nor readily available. Moreover, Theorem 3.2 is new and its proof depends on the proof of Theorem 3.1.

THEOREM 3.1 (see [6], [7]). *Given an ODE of the form*

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \end{bmatrix}, \tag{3.1}$$

where $x_i \in \mathbf{R}^{\mathbf{n_i}}$, $n=n_1+n_2+n_3$, and $f_i(x)$ is C^k for $k\ge 2$. Suppose that

(3.2) *the eigenvalues of A_1 have negative real part,*

(3.3) *the eigenvalues of A_2 have nonnegative real part,*

(3.4) *the eigenvalues of A_3 have nonpositive real part,*

$$f_i(0,0,0)=0,\ i=1,2,3, \tag{3.5}$$

$$\frac{\partial f_i}{\partial x_j}(0,0,0)=0,\ i,j=1,2,3, \tag{3.6}$$

$$f_i(0,x_2,0)=0,\ i=1,3. \tag{3.7}$$

Then there exists, around $x=(0,0,0)$, a local C^{k-2} invariant manifold

$$x_3=\phi(x_1,x_2), \tag{3.8}$$

where

$$\phi(0,x_2)=0, \tag{3.9}$$

$$\frac{\partial\phi}{\partial x}(0,0)=0 \textit{ if } k>2. \tag{3.10}$$

Remarks. The condition (3.7) implies that $\{x_1=0,\ x_3=0\}$ is an invariant manifold. When the spectrum of A_2 lies on the imaginary axis, we call this a partial center manifold as it corresponds to only some of the eigenvalues on the imaginary axis. We call (3.8) a local stable and partial center manifold because it contains the local stable manifold and part of the local center manifold. The partial center manifold provides the needed gap between the eigenvalues that are associated to the invariant manifold and those that are not. The necessity of the existence of the partial center manifold to the existence of the stable and partial center manifold could be argued as follows. If the stable and partial center manifold exists, then its intersection with the center manifold should yield the partial center manifold. The flaw in this argument is that center manifolds are not necessarily unique and the intersection of manifolds is not necessarily a manifold [5], [9]. Still it is plausible. In the above theorem there is a loss of smoothness, from a C^k dynamics to a C^{k-2} local stable and partial center manifold. This is probably an artifact of the proof. In the stable manifold theorem and the theorem of Aulbach and Flockerzi [7] there is no loss of smoothness, and in the center manifold theorem there is a loss of smoothness from C^k to C^{k-1}.

Proof. The first step to make suitable linear changes of coordinates on each of the three subspaces so that there exist $\alpha\geq 16\beta>0$ such that for all $x_1,\ x_2,\ x_3$

$$\begin{aligned} x_1'A_1x_1 &\leq -\alpha|x_1|^2,\\ x_2'A_2x_2 &\leq \beta|x_2|^2,\\ -x_3'A_3x_3 &\leq \beta|x_3|^2. \end{aligned}$$

This is possible by Lemma 1 of [15].

The next step is to use a cut-off function to redefine f. Let $\nu(x)$ be a scalar valued C^∞ function, $0\leq\nu(x)\leq 1$, $\nu(x)=1$ for $0\leq|x|\leq 1$, and $\nu(x)=0$ for $|x|\geq 2$. For any $\epsilon>0$, define

$$f(x;\epsilon):=f(\nu(x/\epsilon)x).$$

Since $f(x;\epsilon)$ agrees with $f(x)$ for $|x|\leq\epsilon$, it suffices to prove the theorem for some $\epsilon>0$.

Next we show that there exists a continuous function $k(\epsilon)$ with $k(0)=0$ and a constant $K>0$ such that for all $x,\ \bar{x}\in\mathbf{R^n}$ and for $i=1,2,3$

$$|f_i(x;\epsilon)-f_i(\bar{x};\epsilon)|^2\leq k^2(\epsilon)|x-\bar{x}|^2 \tag{3.11}$$

and for $i=1,3$

$$|f_i(x;\epsilon)-f_i(\bar{x};\epsilon)|^2\leq k^2(\epsilon)\left|\begin{array}{c} x_1-\bar{x}_1\\ x_3-\bar{x}_3\end{array}\right|^2+K\left|\begin{array}{c}\bar{x}_1\\ \bar{x}_3\end{array}\right|^2|x_2-\bar{x}_2|^2. \tag{3.12}$$

Note that

$$|f_i(x;\epsilon)-f_i(\bar{x};\epsilon)|^2\leq\left|\frac{\partial f_i}{\partial x}(\xi;\epsilon)\right|^2|x-\bar{x}|^2,$$

where ξ is some point on the line between x and $\bar{x}$. Also for $i = 1, 3$

$$\begin{aligned}|f_i(x;\epsilon) - f_i(\bar{x};\epsilon)|^2 &\le |f_i(x;\epsilon) - f_i(\bar{x}_1, x_2, \bar{x}_3;\epsilon) + f_i(\bar{x}_1, x_2, \bar{x}_3;\epsilon) - f_i(\bar{x};\epsilon)|^2 \\ &\le 2|f_i(x;\epsilon) - f_i(\bar{x}_1, x_2, \bar{x}_3;\epsilon)|^2 + 2|f_i(\bar{x}_1, x_2, \bar{x}_3;\epsilon) - f_i(\bar{x};\epsilon)|^2 \\ &\le 2\left|\frac{\partial f_i}{\partial x}(\xi_1, x_2, \xi_3;\epsilon)\right|^2 \left|\begin{array}{c} x_1 - \bar{x}_1 \\ 0 \\ x_3 - \bar{x}_3 \end{array}\right|^2 \\ &\quad +2\left|\frac{\partial f_i}{\partial x_2}(\bar{x}_1, \xi_2, \bar{x}_3;\epsilon)\right|^2 |x_2 - \bar{x}_2|^2,\end{aligned}$$

where (ξ_1, x_2, ξ_3) is some point on the line between x and $(\bar{x}_1, x_2, \bar{x}_3)$ and $(\bar{x}_1, \xi_2, \bar{x}_3)$ is some point on the line between $(\bar{x}_1, x_2, \bar{x}_3)$ and $\bar{x}$. Furthermore,

$$\left|\frac{\partial f_i}{\partial x_2}(\bar{x}_1, \xi_2, \bar{x}_3;\epsilon)\right|^2 \le \left|\frac{\partial^2 f_i}{\partial x \partial x_2}(\xi;\epsilon)\right|^2 \left|\begin{array}{c} \bar{x}_1 \\ 0 \\ \bar{x}_3 \end{array}\right|^2,$$

where ξ is some point on the line between $(0, \xi_2, 0)$ and $(\bar{x}_1, \xi_2, \bar{x}_3)$.

Now $\nu(x)$ and its partials are continuous functions with compact support so there exists a constant M such that

$$\begin{aligned}\left|\frac{\partial \nu}{\partial x}(x)\right| &\le M, \\ \left|\frac{\partial^2 \nu}{\partial x^2}(x)\right| &\le M\end{aligned}$$

for all x. Since f_i satisfies (3.6) we can choose M large enough so that

$$\begin{aligned}\left|\frac{\partial f_i}{\partial x}(x)\right| &\le M|x|, \\ \left|\frac{\partial^2 f_i}{\partial x^2}(x)\right| &\le M\end{aligned}$$

for all $|x| \le 1$. Then for $0 < \epsilon < 1/2$

$$\begin{aligned}\left|\frac{\partial f_i}{\partial x}(x;\epsilon)\right| &\le \left|\frac{\partial f_i}{\partial x}(\nu(x/\epsilon)x)\right| \left|\nu(x/\epsilon) + \frac{\partial \nu}{\partial x}(x/\epsilon)\frac{x}{\epsilon}\right| \\ &\le 2M(1+2M)\epsilon,\end{aligned}$$

$$\begin{aligned}\left|\frac{\partial^2 f_i}{\partial x^2}(x;\epsilon)\right| &\le \left|\frac{\partial^2 f_i}{\partial x^2}(\nu(x/\epsilon)x)\right| \left|\nu(x/\epsilon) + \frac{\partial \nu}{\partial x}(x/\epsilon)\frac{x}{\epsilon}\right|^2 \\ &\quad + \left|\frac{\partial f_i}{\partial x}(\nu(x/\epsilon)x)\right| \left|\frac{\partial \nu}{\partial x}(x/\epsilon)\frac{2}{\epsilon} + \frac{\partial^2 \nu}{\partial x^2}(x/\epsilon)\frac{x}{\epsilon^2}\right| \\ &\le M(1+2M)^2 + 8M^2.\end{aligned}$$

Let

$$\begin{aligned}k^2(\epsilon) &= 2\left(2M(1+2M)\epsilon\right)^2, \\ K &= 2\left(M(1+2M)^2 + 8M^2\right)^2;\end{aligned}$$

then $k(\epsilon)$ is continuous and goes to 0 as ϵ goes to 0.

Henceforth we suppress the ϵ and write $f(x)$ for $f(x;\epsilon)$.

Let $k_1,\ k_2$ be any positive constants and X denote the space of all Lipschitz continuous functions $\phi(x_1,x_2)$ defined for $|x_1|<\epsilon$ and any x_2 such that

$$\phi(0,x_2)=0, \tag{3.13}$$

$$|\phi(x_1,x_2)-\phi(\bar{x}_1,\bar{x}_2)|^2 \le k_1|x_1-\bar{x}_1|^2+k_2|\bar{x}_1|\,|x_2-\bar{x}_2|^2. \tag{3.14}$$

Taking $\bar{x}_1=0$, these imply that

$$|\phi(x_1,x_2)|^2\le k_1|x_1|^2$$

so we can define

$$\|\phi\|^2=\sup\left\{\frac{|\phi(x_1,x_2)|^2}{|x_1|}:|x_1|<\epsilon\right\}. \tag{3.15}$$

With this norm, X is a complete space.

For $|x_1|<\epsilon,\ x_2\in\mathbf{R^{n_2}}$, and $\phi\in X$, define

$$\xi_i(t)=\xi_i(t;x_1,x_2,\phi)$$

for $i=1,2$ to be the solution of

$$\dot{\xi}_i=A_i\xi_i+f_i(\xi_1,\xi_2,\phi(\xi_1,\xi_2)), \tag{3.16}$$

$$\xi_i(0)=x_i. \tag{3.17}$$

Define a mapping T on X as follows:

$$(T\phi)(x_1,x_2)=\int_\infty^0 e^{-A_3s}f_3(\xi_1(s),\xi_2(s),\phi(\xi_1(s),\xi_2(s)))\,ds. \tag{3.18}$$

We would like to show that for ϵ sufficiently small, T is a contraction on X.

Suppose $x_1=0$; then $\xi_1(t)=0$ because of (3.7) and for the same reason

$$(T\phi)(0,x_2)=0$$

so $T\phi$ satisfies (3.13).

Suppose $\phi,\ \bar{\phi}\in X$; then for any $|x_1|<\epsilon,\ x_2\in\mathbf{R^{n_2}},\ |\bar{x}_1|<\epsilon,\ \bar{x}_2\in\mathbf{R^{n_2}}$, and $x_3=\phi(x_1,x_2),\ \bar{x}_3=\bar{\phi}(\bar{x}_1,\bar{x}_2)$, then by the above for $i=1,2,3$

$$\begin{aligned}|f_i(x)-f_i(\bar{x})|^2&\le 2|f_i(x)-f_i(\bar{x}_1,\bar{x}_2,\phi(\bar{x}_1,\bar{x}_2))|^2\\&\quad+2|f_i(\bar{x}_1,\bar{x}_2,\phi(\bar{x}_1,\bar{x}_2),)-f_i(\bar{x})|^2\\&\le 2k^2(\epsilon)\left((1+k_1)|x_1-\bar{x}_1|^2+(1+k_2\epsilon)\,|x_2-\bar{x}_2|^2+\epsilon\|\phi-\bar{\phi}\|^2\right)\end{aligned} \tag{3.19}$$

and for $i=1,3$

$$\begin{aligned}|f_i(x)-f_i(\bar{x})|^2&\le 2|f_i(x)-f_i(\bar{x}_1,\bar{x}_2,\phi(\bar{x}_1,\bar{x}_2))|^2\\&\quad+2|f_i(\bar{x}_1,\bar{x}_2,\phi(\bar{x}_1,\bar{x}_2))-f_i(\bar{x})|^2\\&\le l_1(\epsilon)|x_1-\bar{x}_1|^2+l_2(\epsilon)|\bar{x}_1|\,|x_2-\bar{x}_2|^2+l_3(\epsilon)|\bar{x}_1|\,\|\phi-\bar{\phi}\|^2,\end{aligned} \tag{3.20}$$

where the functions

$$\begin{aligned} l_1(\epsilon) &= 2k^2(\epsilon)(1+k_1), \\ l_2(\epsilon) &= 2\left(k^2(\epsilon)k_2 + K(1+k_1)\epsilon\right), \\ l_3(\epsilon) &= 2k^2(\epsilon) \end{aligned}$$

go to 0 as $\epsilon \to 0$.

Suppose $\xi_i(t),\ \bar{\xi}_i(t)$ for $i = 1, 2$ are the solutions of

$$\begin{aligned} \dot{\xi}_i &= A_i\xi_i + f_i(\xi_1, \xi_2, \phi(\xi_1, \xi_2)), \\ \xi_i(0) &= x_i, \end{aligned}$$

$$\begin{aligned} \dot{\bar{\xi}}_i &= A_i\bar{\xi}_i + f_i(\bar{\xi}_1, \bar{\xi}_2, \bar{\phi}(\bar{\xi}_1, \bar{\xi}_2)), \\ \bar{\xi}_i(0) &= \bar{x}_i \end{aligned}$$

and $\xi_3(t) = \phi(\xi_1(t), \xi_2(t)),\ \bar{\xi}_3(t) = \bar{\phi}(\bar{\xi}_1(t), \bar{\xi}_2(t))$.

Then since $2ab \le a^2 + b^2$ and $(a+b)^2 \le 2a^2 + 2b^2$

$$\begin{aligned} &\frac{d}{dt}\frac{|\xi_1 - \bar{\xi}_1|^2 + |\xi_2 - \bar{\xi}_2|^2}{2} \\ &\le (\xi_1 - \bar{\xi}_1)'\left(A_1(\xi_1 - \bar{\xi}_1) + f_1(\xi) - f_1(\bar{\xi})\right) \\ &\quad + (\xi_2 - \bar{\xi}_2)'\left(A_2(\xi_2 - \bar{\xi}_2) + f_2(\xi) - f_2(\bar{\xi})\right) \\ &\le -\alpha|\xi_1 - \bar{\xi}_1|^2 + \beta|\xi_2 - \bar{\xi}_2|^2 + k(\epsilon)\left(|\xi_1 - \bar{\xi}_1| + |\xi_2 - \bar{\xi}_2|\right) \\ &\quad \left[2(1+k_1)|\xi_1 - \bar{\xi}_1|^2 + 2(1+k_2\epsilon)|\xi_2 - \bar{\xi}_2|^2 + 2\epsilon\|\phi - \bar{\phi}\|^2\right]^{\frac{1}{2}} \\ &\le \left(-\alpha + k(\epsilon)(2+k_1)\right)|\xi_1 - \bar{\xi}_1|^2 + \left(\beta + k(\epsilon)(2+k_2\epsilon)\right)|\xi_2 - \bar{\xi}_2|^2 \\ &\quad + k(\epsilon)\epsilon\|\phi - \bar{\phi}\|^2. \end{aligned}$$

We assume ϵ is small enough so that

$$\begin{aligned} -\alpha + k(\epsilon)(2+k_1) &\le 2\beta \\ k(\epsilon)(2+k_2\epsilon) &\le \beta \\ k(\epsilon)\epsilon &\le 2\beta; \end{aligned}$$

then

$$\frac{d}{dt}\left(|\xi_1 - \bar{\xi}_1|^2 + |\xi_2 - \bar{\xi}_2|^2\right) \le 4\beta\left(|\xi_1 - \bar{\xi}_1|^2 + |\xi_2 - \bar{\xi}_2|^2 + \|\phi - \bar{\phi}\|^2\right)$$

and by Gronwall's inequality

$$|\xi_1(t) - \xi_1(t)|^2 + |\xi_2(t) - \xi_2(t)|^2 \le e^{4\beta t}\left(|x_1 - \bar{x}_1|^2 + |x_2 - \bar{x}_2|^2 + \|\phi - \bar{\phi}\|^2\right). \tag{3.21}$$

With this inequality in hand we can obtain a stricter one by using (3.20) instead of (3.19).

$$\begin{aligned} \frac{d}{dt}\frac{|\xi_1 - \bar{\xi}_1|^2}{2} &\le -\alpha|\xi_1 - \bar{\xi}_1|^2 + |\xi_1 - \bar{\xi}_1| \\ &\quad \left[l_1(\epsilon)|\xi_1 - \bar{\xi}_1|^2 + l_2(\epsilon)|\bar{\xi}_1|\ |\xi_2 - \bar{\xi}_2|^2 + l_3(\epsilon)|\bar{\xi}_1|\ \|\phi - \bar{\phi}\|^2\right]^{\frac{1}{2}}. \end{aligned} \tag{3.22}$$

Now suppose that $\bar{x}_1 = 0,\ \bar{x}_2 = 0$ so that $\bar{\xi}_1 = 0,\ \bar{\xi}_2 = 0$; then

$$\frac{d}{dt}\frac{|\xi_1|^2}{2} \le \left(-\alpha + l_1^{\frac{1}{2}}(\epsilon)\right)|\xi_1|^2.$$

If ϵ is small enough so that

$$l_1^{\frac{1}{2}}(\epsilon) \le \frac{\alpha}{2},$$

then by Gronwall

$$|\xi_1(t)|^2 \le e^{-\alpha t}|x_1|^2. \tag{3.23}$$

For $a,\ b,\ c \ge 0$ we have $\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$ so (3.22) becomes

$$\begin{aligned}\frac{d}{dt}\frac{|\xi_1 - \bar{\xi}_1|^2}{2} &\le \left(-\alpha + l_1^{\frac{1}{2}}(\epsilon)\right)|\xi_1 - \bar{\xi}_1|^2\\ &\quad + l_2^{\frac{1}{2}}(\epsilon)|\xi_1 - \bar{\xi}_1|\ |\bar{\xi}_1|^{\frac{1}{2}}|\xi_2 - \bar{\xi}_2|\\ &\quad + l_2^{\frac{1}{2}}(\epsilon)|\xi_1 - \bar{\xi}_1|\ |\bar{\xi}_1|^{\frac{1}{2}}\|\phi - \bar{\phi}\|\\ &\le \left(-\alpha + l_1^{\frac{1}{2}}(\epsilon) + l_2^{\frac{1}{2}}(\epsilon) + l_3^{\frac{1}{2}}(\epsilon)\right)|\xi_1 - \bar{\xi}_1|^2\\ &\quad + l_2^{\frac{1}{2}}(\epsilon)|\bar{\xi}_1|\ |\xi_2 - \bar{\xi}_2|^2\\ &\quad + l_3^{\frac{1}{2}}(\epsilon)|\bar{\xi}_1|\ \|\phi - \bar{\phi}\|^2.\end{aligned}$$

Assume ϵ is small enough so that

$$l_1^{\frac{1}{2}}(\epsilon) + l_2^{\frac{1}{2}}(\epsilon) + l_3^{\frac{1}{2}}(\epsilon) \le \frac{\alpha}{2},$$

then using (3.23), (3.21), this becomes

$$\begin{aligned}\frac{d}{dt}\frac{|\xi_1 - \bar{\xi}_1|^2}{2} &\le -\frac{\alpha}{2}|\xi_1 - \bar{\xi}_1|^2\\ &\quad + l_2^{\frac{1}{2}}(\epsilon)e^{(4\beta - \frac{\alpha}{2})t}|\bar{x}_1|\left(|x_1 - \bar{x}_1|^2 + |x_2 - \bar{x}_2|^2 + \|\phi - \bar{\phi}\|^2\right)\\ &\quad + l_3^{\frac{1}{2}}(\epsilon)e^{-\frac{\alpha}{2}t}|\bar{x}_1|\ \|\phi - \bar{\phi}\|^2.\end{aligned}$$

Since $16\beta \le \alpha$,

$$\begin{aligned}\frac{d}{dt}|\xi_1 - \bar{\xi}_1|^2 &\le -\alpha|\xi_1 - \bar{\xi}_1|^2\\ &\quad + 2l_2^{\frac{1}{2}}(\epsilon)e^{-\frac{\alpha}{4}t}|\bar{x}_1|\left(|x_1 - \bar{x}_1|^2 + |x_2 - \bar{x}_2|^2\right)\\ &\quad + 2(l_2^{\frac{1}{2}}(\epsilon) + l_3^{\frac{1}{2}}(\epsilon))e^{-\frac{\alpha}{4}t}|\bar{x}_1|\ \|\phi - \bar{\phi}\|^2\end{aligned}$$

so by Gronwall

$$\begin{aligned}|\xi_1(t) - \bar{\xi}_1(t)|^2 &\le e^{-\alpha t}|x_1 - \bar{x}_1|^2\\ &\quad + \frac{8}{3}l_2^{\frac{1}{2}}(\epsilon)e^{-\frac{\alpha}{4}t}|\bar{x}_1|\left(|x_1 - \bar{x}_1|^2 + |x_2 - \bar{x}_2|^2\right)\\ &\quad + \frac{8}{3}(l_2^{\frac{1}{2}}(\epsilon) + l_3^{\frac{1}{2}}(\epsilon))e^{-\frac{\alpha}{4}t}|\bar{x}_1|\ \|\phi - \bar{\phi}\|^2.\end{aligned} \tag{3.24}$$

Next we use (3.20) to estimate

$$\begin{aligned}\left|(T\phi)(x_1,x_2)-(T\bar{\phi})(\bar{x}_1,\bar{x}_2)\right|^2 &= \left|\int_0^\infty e^{-A_3 s}\left(f_3(\xi(s))-f_3(\bar{\xi}(s))\right)\ ds\right|^2\\ &\le \int_0^\infty e^{2\beta s}\left|f_3(\xi(s))-f_3(\bar{\xi}(s))\right|^2\ ds\\ &\le \int_0^\infty e^{2\beta s}\ \big[l_1(\epsilon)|\xi_1(s)-\bar{\xi}_1(s)|^2\\ &\qquad +l_2(\epsilon)|\bar{\xi}_1(s)|\ |\xi_2(s)-\bar{\xi}_2(s)|^2\\ &\qquad +l_3(\epsilon)|\bar{\xi}_1(s)|\ \|\phi-\bar{\phi}\|^2\big]\ ds.\end{aligned}$$

From (3.21), (3.23), (3.24) and $16\beta\le\alpha$

$$\begin{aligned}&\left|(T\phi)(x_1,x_2)-(T\bar{\phi})(\bar{x}_1,\bar{x}_2)\right|^2 \le \int_0^\infty e^{2\beta s}\Big[l_1(\epsilon)\,(\ e^{-\alpha s}|x_1-\bar{x}_1|^2\\ &\quad+\frac{8}{3}l_2^{\frac{1}{2}}(\epsilon)e^{-\frac{\alpha}{4}s}|\bar{x}_1|\left(|x_1-\bar{x}_1|^2+|x_2-\bar{x}_2|^2\right)\\ &\quad+\frac{8}{3}\left(l_2^{\frac{1}{2}}(\epsilon)+l_3^{\frac{1}{2}}(\epsilon)\right)e^{-\frac{\alpha}{4}s}|\bar{x}_1|\ \|\phi-\bar{\phi}\|^2\ )\\ &\quad+l_2(\epsilon)e^{-\frac{\alpha}{2}s}|\bar{x}_1|\ e^{4\beta s}\left(|x_1-\bar{x}_1|^2+|x_2-\bar{x}_2|^2+\|\phi-\bar{\phi}\|^2\right)\\ &\quad+l_3(\epsilon)e^{-\frac{\alpha}{2}s}|\bar{x}_1|\ \|\phi-\bar{\phi}\|^2\Big]\ ds\\ &\quad\le m_1(\epsilon)|x_1-\bar{x}_1|^2+m_2(\epsilon)|\bar{x}_1|\ |x_2-\bar{x}_2|^2\\ &\quad+m_3(\epsilon)|\bar{x}_1|\ \|\phi-\bar{\phi}\|^2,\end{aligned}\tag{3.25}$$

where

$$\begin{aligned}m_1(\epsilon)&=l_1(\epsilon)\left(\frac{8}{7\alpha}+\frac{64}{3\alpha}l_2^{\frac{1}{2}}(\epsilon)\epsilon\right)+\frac{8}{\alpha}l_2(\epsilon)\epsilon,\\ m_2(\epsilon)&=\frac{64}{3\alpha}l_1(\epsilon)l_2^{\frac{1}{2}}(\epsilon)+\frac{8}{\alpha}l_2(\epsilon),\\ m_3(\epsilon)&=\frac{64}{3\alpha}l_1(\epsilon)\left(l_2^{\frac{1}{2}}(\epsilon)+l_3^{\frac{1}{2}}(\epsilon)\right)+\frac{8}{\alpha}l_2(\epsilon)+\frac{8}{3\alpha}l_3(\epsilon).\end{aligned}$$

Notice that $m_i(\epsilon)\to 0$ as $\epsilon\to 0$.

By letting $\bar{\phi}=\phi$ we see that

$$\begin{aligned}\left|(T\phi)(x_1,x_2)-(T\phi)(\bar{x}_1,\bar{x}_2)\right|^2 &\le 2m_1(\epsilon)|x_1-\bar{x}_1|^2\\ &\quad+2m_2(\epsilon)|\bar{x}_1|\ |x_2-\bar{x}_2|^2,\end{aligned}$$

so $(T\phi)(x_1,x_2)$ satisfies (3.14) for ϵ sufficiently small and T maps X to X.

By letting $\bar{x}=x$ we see that

$$\left|(T\phi)(x_1,x_2)-(T\bar{\phi})(x_1,x_2)\right|^2\le m_3(\epsilon)|\bar{x}_1|\ \|\phi-\bar{\phi}\|^2,$$

so $T : X \to X$ is a contraction for ϵ sufficiently small. Hence there exists a unique $\phi \in X$ such that

$$\phi = T\phi.$$

Let $\xi_i(t)$ satisfy (3.16), (3.17) for $i = 1, 2$ and $\xi_3(t) = \phi(\xi_1(t), \xi_2(t))$. By the definition of T (3.18),

$$\begin{aligned} \xi_3(t) &= (T\phi)(\xi_1(t), \xi_2(t)) \\ &= \int_\infty^0 e^{-A_3 s} f_3(\xi(t+s))\, ds \\ &= \int_\infty^t e^{-A_3(t-s)} f_3(\xi(s))\, ds, \end{aligned}$$

so $\xi(t)$ is a solution of the differential equation (3.1) and (3.8) defines a C^0 invariant manifold.

Now suppose $k > 2$. We wish to show that the invariant manifold (3.8) is C^1. Consider the dynamics tangent to (3.1),

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} g_1(x,z) \\ g_2(x,z) \\ g_3(x,z) \end{bmatrix}, \tag{3.26}$$

where

$$g_i(x,z) = \frac{\partial f_i}{\partial x}(x) z.$$

The combined system (3.1), (3.26) satisfies the hypothesis of Theorem 3.1, so a C^0 invariant manifold

$$\begin{bmatrix} x_3 \\ z_3 \end{bmatrix} = \begin{bmatrix} \phi(x_1, x_2, z_1, z_2) \\ \psi(x_1, x_2, z_1, z_2) \end{bmatrix} \tag{3.27}$$

can be found by the extension of the above contraction, call it S. Suppose $\phi(x_1, x_2)$ is a C^1 element of X; define

$$\psi(x_1, x_2, z_1, z_2) = \frac{\partial \phi}{\partial(x_1, x_2)}(x_1, x_2) \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}. \tag{3.28}$$

If

$$(\bar{\phi}(x_1, x_2, z_1, z_2), \bar{\psi}(x_1, x_2, z_1, z_2)) = S(\phi(x_1, x_2), \psi(x_1, x_2, z_1, z_2)),$$

then it is straightforward to verify that

$$\bar{\phi}(x_1, x_2, z_1, z_2) = T(\phi(x_1, x_2)),$$

so $\bar{\phi}(x_1, x_2, z_1, z_2) = \bar{\phi}(x_1, x_2) \in X$ and

$$\bar{\psi}(x_1, x_2, z_1, z_2) = \frac{\partial \bar{\phi}}{\partial(x_1, x_2)}(x_1, x_2) \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Hence if we start the contraction at $\phi,\ \psi$ satisfying (3.28), it will converge to a $\phi,\ \psi$ satisfying (3.28), so the invariant manifold (3.8) is C^1. By repeated application of this technique we can show it is C^{k-2}.

It remains to verify (3.10) if $k > 2$. Clearly (3.9) implies

$$\frac{\partial \phi}{\partial x_2}(0,0) = 0.$$

Since (3.8) defines a C^{k-2} invariant manifold, we can take its time derivative to obtain the PDE

$$\begin{aligned}&A_3\phi(x_1,x_2) + f_3(x_1,x_2,\phi(x_1,x_2))\\ &= \sum_{j=1}^{2} \frac{\partial \phi_i}{\partial x_j}(x_1,x_2)\left(A_j x_j + f_j(x_1,x_2,\phi(x_1,x_2))\right). \end{aligned} \tag{3.29}$$

Taking the linear terms from both sides, we obtain a homogeneous linear equation

$$A_3\frac{\partial \phi}{\partial x_1}(0,0) - \frac{\partial \phi}{\partial x_1}(0,0)A_1 = 0. \tag{3.30}$$

The eigenvalues of the linear mapping

$$B \mapsto A_3B - BA_1$$

have positive real part because they are of the form $\lambda_3 - \lambda_1$, where $\lambda_3,\ \lambda_1$ are eigenvalues of $A_3,\ A_1$, respectively, and so the real part of $\lambda_3 - \lambda_1$ is positive. Hence (3.30) is nonsingular and

$$\frac{\partial \phi}{\partial x_1}(0,0) = 0. \qquad \square$$

The next theorem gives a term-by-term approximation of the stable and partial center manifold.

Theorem 3.2. *Suppose the hypothesis of Theorem* 3.1 *holds for* $k > 3$ *and let* $\phi(x_1,x_2)$ *define the* C^{k-2} *stable and partial center manifold. Suppose* $\psi(x_1,x_2)$ *is a* C^{k-2} *function satisfying* (3.9) *and the PDE* (3.29) *through terms of degree* $k-3$*,*

$$\begin{aligned}&A_3\psi(x_1,x_2) + f_3(x_1,x_2,\psi(x_1,x_2))\\ &= \sum_{j=1}^{2} \frac{\partial \psi}{\partial x_j}(x_1,x_2)\left(A_j x_j + f_j(x_1,x_2,\psi(x_1,x_2))\right) + O(x_1,x_2)^{k-2}. \end{aligned} \tag{3.31}$$

Then ϕ *and* ψ *agree to degree* $k-3$*,*

$$\psi(x_1,x_2) = \phi(x_1,x_2) + O(x_1,x_2)^{k-2}. \tag{3.32}$$

Proof. The theorem holds because the Taylor series coefficients to degree $k-3$ of any ψ satisfying (3.31) are uniquely determined. To see this we use induction on r. Assume that ϕ and ψ satisfy (3.31) to degree r and their Taylor series coefficients agree up to $r-1$.

Assume that $A_1,\ A_2,\ A_3$ are semisimple so that there exist bases of left and right eigenvectors. If they are not semisimple, the argument is essentially the same involving bases of left and right generalized eigenvectors but the details are messier, and hence are ommitted. Suppose for $j = 1, 2$, $k = 1, \dots, n_i$, and $l = 1, \dots, n_j$

$$w_{j,l}A_j = \lambda_{j,l}w_{j,l}, \tag{3.33}$$

$$A_3 v^{3,k} = \lambda_{3,k}v^{3,k}. \tag{3.34}$$

Let $\psi^{[r]}(x_1, x_2)$ be the part of $\psi(x_1, x_2)$ that is a homogeneous polynomial of degree r. A basis for the n_3-vector fields homogeneous of degree r in x_1, x_2 consists of the vector fields

$$\phi^{3,k}_{j_1,l_1;\dots;j_r,l_r}(x) = v^{3,k}(w_{j_1,l_1}x_{j_1})\dots(w_{j_r,l_r}x_{j_r}), \tag{3.35}$$

where $k = 1, \dots, n_i$, $j_s = 1, 2$, $l_s = 1, \dots, n_{j_s}$ and the pairs $(j_1, l_1) \le \dots \le (j_r, l_r)$ are in lexographic order.

Thus

$$\psi^{[r]}(x_1, x_2) = \sum_{k;j_1,l_1;\dots;j_r,l_r} \gamma^{j_1,l_1;\dots;j_r,l_r}_{3,k} \psi^{3,k}_{j_1,l_1;\dots;j_r,l_r}(x_1, x_2) \tag{3.36}$$

with

$$\gamma^{j_1,l_1;\dots;j_r,l_r}_{3,k} = 0 \tag{3.37}$$

if $j_1 = \dots = j_r = 2$ so that (3.9) is satisfied.

If we extract the degree r terms from (3.31), we obtain

$$A_3\psi^{[r]}3(x_1, x_2) - \sum_{j=1}^{2} \frac{\partial \psi^{[r]}_i}{\partial x_j}(x_1, x_2)A_j x_j = h^{[r]}(x_1, x_2), \tag{3.38}$$

where $h^{[r]}(x_1, x_2)$ depends only on the ODE (3.1) and the lower degree part of $\psi(x_1, x_2)$ which has been determined by (3.31). Now

$$\begin{aligned} A_3\psi^{3,k}_{j_1,l_1;\dots;j_r,l_r}(x_1, x_2) &- \sum_{j=1}^{2} \frac{\partial \psi^{3,k}_{j_1,l_1;\dots;j_r,l_r}(x_1, x_2)}{\partial x_j}(x_1, x_2)A_j x_j \\ &= (\lambda_{3,k} - \lambda_{j_1,l_1} - \dots - \lambda_{j_r,l_r})\, \psi^{3,k}_{j_1,l_1;\dots;j_r,l_r}(x_1, x_2). \end{aligned}$$

The real part of $\lambda_{3,k}$ is nonnegative and the real parts of λ_{j_s,l_s} are negative if $j_s = 1$ and are nonpositive if $j_s = 2$. Because of (3.37), we can restrict our attention to $\psi^{3,k}_{j_1,l_1;\dots;j_r,l_r}$, where at least one $j_s = 1$ so such $\gamma^{j_1,l_1;\dots;j_r,l_r}_{3,k}$ are uniquely determined by (3.38). □

Notice that if (3.7) is not satisfied to degree r, then (3.38) might not be solvable for then $h^{[r]}(x_1, x_2)$ might contain terms of the form $\psi^{3,k}_{j_1,l_1;\dots;j_r,l_r}(x_1, x_2)$, where $j_1 = \dots = j_r = 2$ and $\lambda_{3,k} - \lambda_{2,l_1} - \dots - \lambda_{2,l_r}$ might be zero.

4. Local solvability of the HJB PDE. The principle theorem of this paper is the following.

THEOREM 4.1. *Suppose the plant* (2.1) *and exosystem* (2.2) *are C^k, the linear part of the plant is stabilizable and detectable when $\bar{x} = 0$, the linear part of the exosystem is stable, the FBI PDE* (2.3) *has a C^k solution in some neighborhood of* 0 *in $\bar{x}$ space. Then in some neighborhood of* $0, 0$ *in $x, \bar{x}$ space there exists a C^{k-2} solution to HJB PDE* (2.9) *satisfying* (2.12).

Proof. The proof generalizes the standard approach [22] to showing the existence of local solutions to HJB PDEs. The graph of gradient of the solution π of the HJB PDE (2.9) is an invariant manifold of the associated Hamiltonian system of ODEs. In the standard case, the Hamiltonian ODEs have a hyperbolic fixed point at the origin and the invariant manifold is the stable manifold of this fixed point. But in this case, the Hamiltonian ODEs do not have a hyperbolic fixed point at the origin and the desired invariant manifold is a stable and partial center manifold.

Consider the Hamiltonian associated to the optimal control problem (2.8),

$$\begin{aligned} H(\lambda,\mu,z,\bar{x},v) &= \lambda\tilde{f}(z,v,\bar{x})+\mu\bar{f}(\bar{x})+l(z,v,\bar{x}) \\ &= \lambda\big(Az+Bv+\tilde{f}^{[2]}(z,v,\bar{x})\big) \\ &\quad +\mu\left(\bar{A}\bar{x}+\bar{f}^{[2]}(\bar{x})\right) \\ &\quad +\tfrac{1}{2}\left(z'Qz+2z'Sv+v'Rv\right) \\ &\quad +l^{[3]}(z,v,\bar{x})+O(\lambda,\mu,z,\bar{x},v)^4. \end{aligned} \tag{4.1}$$

The Pontryagin maximum principle asserts that the optimal control is

$$v=\gamma(\lambda,\mu,z,\bar{x})=\arg\min_{v} H(\lambda,\mu,z,\bar{x},v). \tag{4.2}$$

For small $\lambda,\mu,z,\bar{x}$ this is given by solving

$$\frac{\partial H}{\partial v}(\lambda,\mu,z,\bar{x},v)=0,$$

which yields

$$\begin{aligned} \gamma = -R^{-1}&\left(B'\lambda'+S'z+\left(\frac{\partial\tilde{f}^{[2]}}{\partial v}\right)'\lambda'+\left(\frac{\partial\tilde{l}^{[3]}}{\partial v}\right)'\right) \\ &+O(\lambda,\mu,z,\bar{x})^3. \end{aligned}$$

The HJB PDE (2.9) can be expressed in terms of the Hamiltonian as

$$H\left(\frac{\partial\pi}{\partial z},\frac{\partial\pi}{\partial\bar{x}},z,\bar{x},\gamma\left(\frac{\partial\pi}{\partial z},\frac{\partial\pi}{\partial\bar{x}},z,\bar{x}\right)\right)=0. \tag{4.3}$$

The Hamiltonian ODEs are

$$\begin{aligned} \dot{z}' &= \frac{\partial H}{\partial\lambda}(\lambda,\mu,z,\bar{x},\gamma(\lambda,\mu,z,\bar{x})), \\ \dot{\lambda} &= -\frac{\partial H}{\partial z}(\lambda,\mu,z,\bar{x},\gamma(\lambda,\mu,z,\bar{x})), \\ \dot{\bar{x}}' &= \frac{\partial H}{\partial\mu}(\lambda,\mu,z,\bar{x},\gamma(\lambda,\mu,z,\bar{x})), \\ \dot{\mu} &= -\frac{\partial H}{\partial\bar{x}}(\lambda,\mu,z,\bar{x},\gamma(\lambda,\mu,z,\bar{x})) \end{aligned} \tag{4.4}$$

and these are C^{k-1} since the Hamiltonian is C^k.

The linearization of this system around $0,0,0,0$ is

$$\left[\begin{array}{c} \dot{z} \\ \dot{\lambda}' \\ \hline \dot{\bar{x}} \\ \dot{\mu}' \end{array}\right]=\left[\begin{array}{cc} \mathbf{H_{11}} & \mathbf{H_{12}} \\ \mathbf{H_{21}} & \mathbf{H_{22}} \end{array}\right]\left[\begin{array}{c} z \\ \lambda' \\ \hline \bar{x} \\ \mu' \end{array}\right], \tag{4.5}$$

where

$$\mathbf{H} = \left[\begin{array}{cc|cc} A - BR^{-1}S' & -BR^{-1}B' & 0 & 0 \\ -Q + SR^{-1}S' & -A' + SR^{-1}B' & 0 & 0 \\ \hline 0 & 0 & \bar{A} & 0 \\ 0 & 0 & 0 & -\bar{A}' \end{array}\right]. \tag{4.6}$$

The column span of

$$\left[\begin{array}{cc} I_{n\times n} & 0 \\ P & 0 \\ 0 & I_{\bar{n}\times\bar{n}} \\ 0 & 0 \end{array}\right] \tag{4.7}$$

is an $n + \bar{n}$ dimensional stable and partial center subspace of the linear Hamiltonian system (4.5), where P is the unique nonnegative definite solution of the algebraic Riccati equation (2.13). We know that such a solution exists because the linear part of the plant was assumed to be stabilizable and detectable [3]. Half of the eigenvalues of the upper left $2n \times 2n$ block $\mathbf{H_{11}}$ lie in the open left half plane and half lie in the open right half plane. The asymptotically stable subspace is spanned by the first n columns of (4.7). As for the lower right $2\bar{n} \times 2\bar{n}$ block $\mathbf{H_{22}}$, by assumption the eigenvalues of $\bar{A}$ are in the closed left half plane and hence those of $-\bar{A}'$ are in the closed right half plane. A stable subspace is spanned by the last $\bar{n}$ columns of (4.7). Furthermore, the submanifold $z = 0,\ \lambda = 0,\ \mu = 0$ is an invariant submanifold of the nonlinear Hamiltonian system (4.4), so the conditions of the stable and partial center manifold Theorem are satisfied. There exists an $n + \bar{n}$ dimensional stable and partial center manifold in the $2(n + \bar{n})$ dimensional $z, \lambda, \bar{x}, \mu$ space which is tangent to the column span of (4.7) at $0, 0, 0, 0$. Hence this manifold is given by

$$\begin{array}{rcl} \lambda & = & \phi(z, \bar{x}), \\ \mu & = & \psi(z, \bar{x}), \end{array} \tag{4.8}$$

where ϕ, ψ are C^{k-3} and

$$\begin{array}{rcl} \phi(0, \bar{x}) & = & 0, \\ \psi(0, \bar{x}) & = & 0. \end{array} \tag{4.9}$$

This submanifold is Lagrangian, i.e., a maximal dimension submanifold on which the canonical two form

$$\omega = d\lambda\ dz + d\mu\ d\bar{x}$$

vanishes [1], [4]. To see that it vanishes we note that ω is invariant under the Hamiltonian flow (4.4) and this flow is converging to the $\bar{n}$ dimensional submanifold $z = 0,\ \lambda = 0,\ \mu = 0$, where ω clearly vanishes. The submanifold (4.8) is of maximal dimension, $n + \bar{n}$, in $2(n + \bar{n})$ variables.

Hence the one form

$$\phi(z, \bar{x})\ dz + \psi(z, \bar{x})\ d\bar{x}$$

is closed locally around $0, 0$ in $z, \bar{x}$ space and so there exists a C^{k-2} function $\pi(z, \bar{x})$ such that

$$\frac{\partial \pi}{\partial z}(z, \bar{x}) = \phi(z, \bar{x}),$$

$$\begin{aligned}
\frac{\partial \pi}{\partial \bar{x}}(z,\bar{x}) &= \psi(z,\bar{x}),\\
\pi(0,\bar{x}) &= 0,\\
\frac{\partial \pi}{\partial z}(0,\bar{x}) &= 0,\\
\frac{\partial \pi}{\partial \bar{x}}(0,\bar{x}) &= 0.
\end{aligned}$$

Note that π satisfies (2.12).

Differentiating (4.8) with respect to t along the Hamiltonian flow (4.4) yields

$$\begin{aligned}
\frac{\partial H}{\partial \lambda}\frac{\partial^2 \pi}{\partial z^2} + \frac{\partial H}{\partial \mu}\frac{\partial^2 \pi}{\partial z \partial \bar{x}} + \frac{\partial H}{\partial z} &= 0,\\
\frac{\partial H}{\partial \lambda}\frac{\partial^2 \pi}{\partial z \partial \bar{x}} + \frac{\partial H}{\partial \mu}\frac{\partial^2 \pi}{\partial \bar{x}^2} + \frac{\partial H}{\partial z} &= 0
\end{aligned}$$

or equivalently

$$\begin{aligned}
\frac{\partial}{\partial z} H\left(\frac{\partial \pi}{\partial z}, \frac{\partial \pi}{\partial \bar{x}}, z, \bar{x}, \gamma\left(\frac{\partial \pi}{\partial z}, \frac{\partial \pi}{\partial \bar{x}}, z, \bar{x}\right)\right) &= 0,\\
\frac{\partial}{\partial \bar{x}} H\left(\frac{\partial \pi}{\partial z}, \frac{\partial \pi}{\partial \bar{x}}, z, \bar{x}, \gamma\left(\frac{\partial \pi}{\partial z}, \frac{\partial \pi}{\partial \bar{x}}, z, \bar{x}\right)\right) &= 0.
\end{aligned}$$

Clearly π satisfies the HJB PDE (4.3) at $z = 0,\ \bar{x} = 0$, so it satisfies it in a neighborhood of this point. Moreover π is of the form

$$\pi(z,\bar{x}) = \frac{1}{2} z' P z + O(z,\bar{x})^3 \qquad \square \tag{4.10}$$

The next theorem shows that the solution to the HJB PDE (2.9) can be computed term-by-term.

THEOREM 4.2. *Suppose the hypotheses of Theorem* 4.1 *hold for* $k > 3$ *and let* $\pi(z,\bar{x})$ *be the* C^{k-2} *solution of the HJB PDE* (2.9) *satisfying* (2.12). *Suppose* $\phi(x_1,x_2)$ *is a* C^{k-2} *function satisfying the HJB PDE through terms of degree* $k-3$ *and satisfying* (2.12). *Then* π *and* ψ *agree to degree* $k-3$,

$$\pi(z,\bar{x})) = \psi(z,\bar{x}) + O(z,\bar{x})^{k-2} \tag{4.11}$$

Proof (sketch). Clearly $\pi(z,\bar{x})$ satisfies the term-by-term equations, so the result follows if we can show that these equations have unique solutions satisfying (2.12). We showed above that the quadratic terms agree, and as for the cubic terms, consider (2.14). The first equation is a linear equation for $\pi^{[3]}$. For simplicity assume that $A + BK$ and $\bar{A}$ have bases of left eigenvectors

$$\begin{array}{lcll}
\xi_i(A+BK) & = & \lambda_i \xi_i, & i = 1,\ldots,n,\\
\zeta_j \bar{A} & = & \mu_j \xi_i, & j = 1,\ldots,\bar{n},
\end{array} \tag{4.12}$$

otherwise we use bases of generalized eigenvectors. Since the linear part of the plant is stabilizable and detectable, the linear part of the closed loop system is asymptotically stable, Re $\lambda_i < 0$, and by assumption the linear part of the exosystem is stable, Re $\mu_j \leq 0$.

Now any cubic polynomial $\pi^{[3]}(z,\bar{x})$ satisfying (2.12) can be expressed as

$$\begin{aligned}\pi^{[3]}(z,\bar{x}) = &\sum c_{i_1,i_2,i_3}\ \xi_{i_1}z\ \xi_{i_2}z\ \xi_{i_3}z\\ &+\sum d_{i_1,i_2,j_3}\ \xi_{i_1}z\ \xi_{i_2}z\ \zeta_{j_3}\bar{x}\end{aligned}$$

and

$$\begin{aligned}&\frac{\partial\pi^{[3]}}{\partial z}(z,\bar{x})(A+BK)z+\frac{\partial\pi^{[3]}}{\partial\bar{x}}(z,\bar{x})(\bar{A}\bar{x})\\ &=\sum c_{i_1,i_2,i_3}\ \ (\lambda_{i_1}+\lambda_{i_2}+\lambda_{i_3})\,\xi_{i_1}z\ \xi_{i_2}z\ \xi_{i_3}z\\ &\quad+\sum d_{i_1,i_2,j_3}\ \ (\lambda_{i_1}+\lambda_{i_2}+\mu_{j_3})\,\xi_{i_1}z\ \xi_{i_2}z\ \zeta_{j_3}\bar{x}.\end{aligned}$$

It follows from (2.11) that

$$z'P\tilde{f}^{[2]}(z,Kz,\bar{x})+l^{[3]}(z,Kz,\bar{x})=O(z,\bar{x})^2,$$

so

$$\begin{aligned}&z'P\tilde{f}^{[2]}(z,Kz,\bar{x})+l^{[3]}(z,Kz,\bar{x})\\ &\quad=\sum k_{i_1,i_2,i_3}\ \xi_{i_1}z\ \xi_{i_2}z\ \xi_{i_3}z\\ &\qquad+\sum l_{i_1,i_2,j_3}\ \xi_{i_1}z\ \xi_{i_2}z\ \zeta_{j_3}\bar{x}\end{aligned}$$

for some k,l's. Hence there is a unique $\pi^{[3]}$ satisfying (2.14) and (2.12) given by

$$\begin{aligned}c_{i_1,i_2,i_3}&=-\frac{k_{i_1,i_2,i_3}}{\lambda_{i_1}+\lambda_{i_2}+\lambda_{i_3}},\\ d_{i_1,i_2,j_3}&=-\frac{l_{i_1,i_2,j_3}}{\lambda_{i_1}+\lambda_{i_2}+\mu_{j_3}}\end{aligned}$$

because the denominators are not zero, Re $\lambda_i<0$, and Re $\mu_j\le 0$. The higher degree terms are handled in a similar fashion. □

5. H_∞ regulation.

One can also use nonlinear H_∞ control techniques to stabilize the transverse dynamics in a robust fashion. Consider a smooth plant

$$\begin{aligned}\dot{x}&=f(x,u,\bar{x})+g(x,\bar{x},w)\\ &=Ax+Bu+F\bar{x}+Gw\\ &\quad+f^{[2]}(x,u,\bar{x})+g^{[2]}(x,\bar{x},w)+O(x,u,\bar{x},w)^3\\ e&=h(x,u,\bar{x})\\ &=Cx+Du+H\bar{x}\\ &\quad+h^{[2]}(x,u,\bar{x})+O(x,u,\bar{x})^3\end{aligned}\tag{5.1}$$

which is perturbed by an unknown noise $w(t)$ and by a smooth nonlinear exosystem

$$\begin{aligned}\dot{\bar{x}}&=\bar{f}(\bar{x},w)\\ &=\bar{A}\bar{x}+\bar{B}w+\bar{f}^{[2]}(\bar{x},w)+O(\bar{x},w)^3.\end{aligned}\tag{5.2}$$

Notice that there is no direct interaction between the control and the noise in the dynamics and the noise w does not directly affect the error e.

The goal is as before, to find a feedforward and feedback control $u = \alpha(x, \bar{x})$ to drive $e(t)$ as close to zero as possible for any $x(0)$, $\bar{x}(0)$ despite the unknown noise. More precisely, for any choice of α the closed loop system defines a map from the initial conditions $x(0)$, $\bar{x}(0)$ and the noise $w(t)$ to the variables that we want to keep small, $u(t)$, $e(t)$. We would like the gain of this mapping to be as small as possible. This is a very difficult problem to solve directly so we settle for a suboptimal solution. Given an attenuation level $\delta > 0$, we seek an $u = \alpha(x, \bar{x})$ so that the map from $x(0)$, $\bar{x}(0)$, $w(t)$ to $u(t)$, $e(t)$ has gain less than δ. This goal needs to be modified because as before we should not penalize those parts of $x(0)$, $u(t)$ that are necessary for exact tracking.

As before we start by solving the FBI equations (2.3) for exact tracking and transform the combined system into transverse coordinates (2.4) to obtain

$$\begin{aligned}
\dot{z} &= \tilde{f}(z, v, \bar{x}) + \tilde{g}(z, \bar{x}, w)\\
&= Ax + Bv + \tilde{G}w + \tilde{f}^{[2]}(z, v, \bar{x}, w) + \tilde{g}^{[2]}(z, \bar{x}, w) + O(z, v, \bar{x}, w)^3\\
\dot{\bar{x}} &= \bar{f}(\bar{x}) = \bar{A}\bar{x} + \bar{B}w + \bar{f}^{[2]}(\bar{x}, w) + O(\bar{x}, w)^3\\
e &= \tilde{h}(z, v, \bar{x}) = Cz + Dv + \tilde{h}^{[2]}(z, v, \bar{x}) + O(z, v, \bar{x})^3,
\end{aligned} \tag{5.3}$$

where

$$\begin{aligned}
\tilde{f}(z, v, \bar{x}) &= f(z + \theta(\bar{x}), v + \beta(\bar{x}), \bar{x}) - \tfrac{\partial \theta}{\partial \bar{x}}(\bar{x})\bar{f}(\theta(\bar{x}), \bar{x})\\
\tilde{g}(z, \bar{x}, w) &= g(z + \theta(\bar{x}), \bar{x}, w)\\
\tilde{h}(z, v, \bar{x}) &= h(z + \theta(\bar{x}), v + \beta(\bar{x}), \bar{x})\\
\tilde{G} &= G - TB.
\end{aligned} \tag{5.4}$$

We wish to find the control $v = \gamma(z, \bar{x})$ that maximizes

$$\pi(z, \bar{x}) = \inf_{w} \frac{1}{2} \int_0^t \delta^2 |w(s)|^2 - |e(s)|^2 - |v(s)|^2 \, ds, \tag{5.5}$$

where the infimum is over all $t \geq 0$, with $w(s)$ generating a trajectory satisfying $z(0) = 0$, $\bar{x}(0) = 0$, $z(t) = z$, $\bar{x}(t) = \bar{x}$. For any control $v = \gamma(z, \bar{x})$, the function $\pi(z, \bar{x})$ is the minimum required net energy that must be supplied to the combined system to go from the origin $0,\ 0$ to $z,\ \bar{x}$. Energy is supplied to the system at the rate $\frac{\delta^2}{2}|w(s)|^2$ and extracted from the system at the rate $\frac{1}{2}(|e(s)|^2 + |v(s)|^2)$. The goal is to supremize the energy necessary to reach any z, $\bar{x}$. See [23] and [18, 19] for more on nonlinear H_∞ control.

An immediate consequence of the definition of $\pi(z, \bar{x})$ is that along any trajectory of the system

$$\pi(z(s), \bar{x}(s))]_{t_1}^{t_2} \leq \frac{1}{2} \int_{t_1}^{t_2} \delta^2 |w(s)|^2 - |e(s)|^2 - |v(s)|^2 \, ds. \tag{5.6}$$

This is called a dissipation inequality; if we view $\pi(z, \bar{x})$ as the energy stored in the combined system when it is in state z, $\bar{x}$ then the change in stored energy over any time interval is less than or equal to the net energy supplied to the system over that time interval.

If there exists a control $v = \gamma(z, w)$ so that $\pi(z, \bar{x}) \geq 0$, then

$$\frac{1}{2} \int_{t_1}^{t_2} |e(s)|^2 + |v(s)|^2 \, ds \leq \pi(z(t_1), \bar{x}(t_1)) + \frac{\delta^2}{2} \int_{t_1}^{t_2} |w(s)|^2 \, ds. \tag{5.7}$$

If this holds, then the energy of the tracking error plus the energy of the control used to reduce it is less than the energy of the initial mismatch between the plant and exosystem, $\pi(z(t_1), \bar{x}(t_1))$, plus the energy of the disturbance.

We can view (5.5) as the cost criterion of a differential game pitting the control v against the noise w. The optimal $\pi,\ v^*,\ w^*$ satisfy the HJI PDE

$$\begin{aligned}0 &= \frac{\partial \pi}{\partial z}(z,\bar{x})\left(\tilde{f}(z,v,\bar{x}) + \tilde{g}(z,\bar{x},w)\right)\\ &\quad+\frac{\partial \pi}{\partial \bar{x}}(z,\bar{x})\bar{f}(\bar{x},w) + l(z,v,\bar{x},w)\\ v^*, w^* &= \arg\ \min_v \max_w \left\{ \frac{\partial \pi}{\partial z}(z,\bar{x})\left(\tilde{f}(z,v,\bar{x}) + \tilde{g}(z,\bar{x},w)\right)\right.\\ &\quad\left.+\frac{\partial \pi}{\partial \bar{x}}(z,\bar{x})\bar{f}(\bar{x},w) + l(z,v,\bar{x},w)\right\},\end{aligned} \tag{5.8}$$

where

$$\begin{aligned}l(z,v,\bar{x},w)) &= \frac{1}{2}\left(|e|^2 + |v|^2\right) - \frac{\gamma^2}{2}|w|^2\\ &= \frac{1}{2}\left(z'Qz + 2z'Sv + v'Rv\right) - \frac{\gamma^2}{2}|w|^2\\ &\quad+l^{[3]}(z,v,\bar{x}) + O(z,v,\bar{x})^4.\end{aligned}$$

Van der Schaft [23] considered the local solvability of the HJI PDE when the plant is stabilizable and detectable and there is no exosystem (so $x = z,\ u = v,\ G = \tilde{G}$). He showed that a local solution exists if the linear quadratic part of the problem admits a stable solution. That is, there exists a $P \geq 0$ satisfying the Riccati equation

$$\begin{aligned}0 &= A'P + PA + Q + \frac{1}{\gamma^2}P\tilde{G}\tilde{G}'P\\ &\quad-(PB+S)R^{-1}(PB+S)'\end{aligned} \tag{5.9}$$

and such that the closed loop spectrum is in the open left half plane,

$$\sigma\left(A - BR^{-1}(B'P + S') + \frac{1}{\gamma^2}\tilde{G}\tilde{G}'P\right) < 0. \tag{5.10}$$

The optimal linear feedback and worst case noise are

$$\begin{aligned}v^* &= -R^{-1}(B'P + S')z,\\ w^* &= \frac{1}{\gamma^2}\tilde{G}'Pz.\end{aligned}$$

Following the approach described in section 4 using the stable and partial center manifold theorem, one can prove the following theorems.

THEOREM 5.1. *Suppose the plant* (5.1) *and exosystem* (5.2) *are* C^k, *the linear part of the plant is stabilizable and detectable when* $\bar{x} = 0$, *the linear part of the exosystem is stable, and the FBI PDE* (2.3) *has a* C^k *solution in some neighborhood of* 0 *in* $\bar{x}$ *space. If there exists a* $P \geq 0$ *satisfying the Riccati equation* (5.9) *and such that the closed loop spectrum is in the open left half plane* (5.10), *then in some neighborhood of* $0, 0$ *in* $x, \bar{x}$ *space there exists a* C^{k-2} *solution to HJB PDE* (5.8) *satisfying* (2.12).

THEOREM 5.2. *Suppose the hypotheses of Theorem* 5.1 *hold for* $k > 3$ *and let* $\pi(z, \bar{x})$ *be the* C^{k-2} *solution of the HJI PDE* (5.8) *satisfying* (2.12). *Suppose* $\phi(x_1, x_2)$ *is a* C^{k-2} *function satisfying the HJI PDE through terms of degree* $k-3$ *and satisfying* (2.12). *Then* π *and* ψ *agree to degree* $k-3$,

$$\pi(z, \bar{x})) = \psi(z, \bar{x}) + O(z, \bar{x})^{k-2}. \tag{5.11}$$

REFERENCES

[1] R. ABRAHAM AND J. E. MARSDEN, *Foundations of Mechanics*, Benjamin/Cummings, Reading, MA, 1978.

[2] E. G. AL'BRECHT, *On the optimal stabilization of nonlinear systems*, PMM-J. Appl. Math. Mech., 25 (1961), pp. 1254–1266.

[3] B. D. O. ANDERSON AND J. B. MOORE, *Optimal Control, Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[4] V. I. ARNOL'D, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.

[5] V. I. ARNOL'D, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, Berlin, 1983.

[6] B. AULBACH, D. FLOCKERZI, AND H. W. KNOBLOCH, *Invariant manifolds and the concept of geometric phase*, Casopis Pro Pestovani Matematiky, 111 (1986), pp. 156–176.

[7] B. AULBACH AND D. FLOCKERZI, *An existence theorem for invariant manifolds*, J. Appl. Math Phys., 38 (1987), pp. 151–171.

[8] C. I. BYRNES, F. DELLI PRISCOLI, A. ISIDORI, AND W. KANG, *Structurally stable output regulation of nonlinear systems*, Automatica, 33 (1997), pp. 369–385.

[9] J. CARR, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1981.

[10] L. C. EVANS, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[11] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions.* Springer-Verlag, New York, 1992.

[12] B. A. FRANCIS, *The linear multivariable regulator problem*, SIAM J. Control Optim., 15 (1977), pp. 486–505.

[13] J. HUANG AND W. J. RUGH, *An approximation method for the nonlinear servomechanism problem*, IEEE Trans. Automat. Control, 37 (1992), pp. 1395–1398.

[14] A. ISIDORI AND C. I. BYRNES, *Output regulation of nonlinear systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 131–140.

[15] A. KELLEY, *The stable, center-stable, center, center-unstable and unstable manifolds*, J. Differential Equations, 3 (1967), pp. 546–570.

[16] A. J. KRENER, *The construction of optimal linear and nonlinear regulators*, in Systems, Models and Feedback: Theory and Applications, A. Isidori and T. J. Tarn, eds., Birkhäuser, Boston, 1992, pp. 301–322.

[17] A. J. KRENER, *Optimal model matching controllers for linear and nonlinear systems*, in Nonlinear Control System Design 1992, M. Fliess, ed., Pergamon Press, Oxford, 1993, pp. 209–214.

[18] A. J. KRENER, *Necessary and sufficient conditions for nonlinear worst case (H_∞) control and estimation*, summary and electronic publication, J. Math. Systems, Estim. Control, 4 (1994), pp. 485–488.

[19] A. J. KRENER, *Necessary and sufficient conditions for nonlinear worst case (H_∞) control and estimation*, summary and electronic publication, J. Math. Systems, Estim. Control, 7 (1997), pp. 81–106.

[20] A. J. KRENER, *Nonlinear Systems Toolbox V.* 1.0, the Math Works, Natick, MA, 1997. MATLAB based toolbox available by ftp from scad.utdallas.edu.

[21] A. J. KRENER, *The existence of optimal regulators*, in Proceedings of the 1998 IEEE Conference on Decision and Control, Tampa, FL, IEEE, Piscataway, NJ, 1998, pp. 3081–3086.

[22] D. L. LUKES, *Optimal regulation of nonlinear dynamical systems*, SIAM J. Control, 7 (1969), pp. 75–100.

[23] A. J. VAN DER SCHAFT, *On a state space approach to nonlinear H_∞ control*, Systems Control Lett., 16 (1991), pp. 1–81.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1485–1498

VALUE-FUNCTIONS FOR DIFFERENTIAL GAMES AND CONTROL SYSTEMS WITH DISCONTINUOUS TERMINAL COST[*]

SŁAWOMIR PLASKACZ[†] AND MARC QUINCAMPOIX[‡]

Abstract. This paper deals with Mayer's problem for control systems and differential games with discontinuous terminal cost. There are two main results in the paper. The first one says that the value function for control systems can be characterized as the unique solution—in suitable sense—to the Hamilton–Jacobi–Bellman equation without any regularity assumptions on the terminal cost. For differential games satisfying Isaacs's minmax condition, the second main result says that the value function is the unique solution to the Hamilton–Jacobi–Isaacs equation when the terminal cost is semicontinuous. This allows to prove the existence of the value under Isaacs's condition. This paper extends some results already well known in the continuous case.

Key words. viscosity solution, Hamilton–Jacobi equations, Mayer's problem

AMS subject classifications. 49J35, 49L25, 49L05, 93C15, 90D25, 35D05, 35F10, 49N55

PII. S0363012998340387

1. Introduction.

We consider a differential game in which dynamics is given by $x'(t) = f(t,x,u,v)$, where the state variable x belongs to $\mathbb{R}^n$, and the controls $u : [t_0,T] \mapsto U$ and $v : [t_0,T] \mapsto V$ are measurable functions. By $x(\cdot;\, t_0,x_0,u,v)$ we denote the solution to the Cauchy problem

$$\left\{ \begin{array}{l} x'(t) = f(t,x(t),u(t),v(t)), \\ x(t_0) = x_0. \end{array} \right. \tag{1}$$

We are interested in a differential game with a terminal cost $g : \mathbb{R}^n \mapsto \mathbb{R}$. Namely, the player acting on the control u tries to minimize the terminal cost $g(x(T))$ while the other player tries to maximize it. These optimal behaviors of the two players are modelized in the framework of nonanticipative strategies introduced by Varayia–Elliot–Kalton–Roxin.

Let us denote by $\mathcal{U}(t_0)$ and $\mathcal{V}(t_0)$ the set of measurable controls on $[t_0,T]$, by $\alpha : \mathcal{V}(t_0) \to \mathcal{U}(t_0)$ a nonanticipative strategy of the first player and by $\beta : \mathcal{U}(t_0) \to \mathcal{V}(t_0)$ a nonanticipative strategy of the second player. The set $\Gamma(t_0)$ (respectively, $\Delta(t_0)$) denotes the set of all nonanticipative strategies of the first (respectively, the second) player.

This leads to the definition of two value functions

$$\left\{ \begin{array}{l} V_g^-(t_0,x_0) := \inf_{\alpha\in\Gamma(t_0)} \sup\{g(x) :\, x \in \mathrm{cl}\ (A_\alpha(t_0,x_0))\}, \\ V_g^+(t_0,x_0) := \sup_{\beta\in\Delta(t_0)} \inf\{g(x) :\, x \in \mathrm{cl}\ (B_\beta(t_0,x_0))\}, \end{array} \right. \tag{2}$$

where cl means closure and $A_\alpha(t_0,x_0) = \{x(T;\, t_0,x_0,\alpha(v),v) :\ v \in \mathcal{V}(t_0)\}$, $B_\beta(t_0,x_0) = \{x(T;\, t_0,x_0,u,\beta(u)) :\ u \in \mathcal{U}(t_0)\}$ denote the reachable sets. Let us notice that when g is continuous, we can skip the closure in the definition (2) of value functions. In this paper we provide an example with a discontinuous g showing that the two

[*] Received by the editors June 15, 1998; accepted for publication (in revised form) August 2, 2000; published electronically January 25, 2001.
http://www.siam.org/journals/sicon/39-5/34038.html

[†] Department of Mathematics and Computer Science, Nicholas Copernicus University, Chopina 12/18, 87-100 Toruń, Poland (plaskacz@mat.uni.torun.pl).

[‡] Département de Mathématiques, Université de Bretagne Occidentale, 6 avenue Victor Le Gorgeu, BP 809, F-29285 Brest cedex, France (nicolle@univ-brest.fr).

value functions V_g^+ and V_g^- are not equal when we do not take the closure in the definition (2).

The main question we address here is the existence of a value of the game, namely, the question of equality between V_g^+ and V_g^-. This problem has been studied—and solved—in [13] in the case of Lipschitz terminal cost g. The present paper deals with discontinuous terminal cost.

Before explaining how we solve the discontinuous Mayer problem for games, let us recall the approach of Evans and Souganidis [13] in the case of a Lipschitz continuous terminal cost. They prove that the two value functions are the unique continuous viscosity solutions of two Hamilton–Jacobi equations. Under the assumption that the two Hamiltonians coincide—the so-called Isaacs condition—these two value functions are hereby equal. In the context of discontinuous terminal cost, the value functions are discontinuous and standard uniqueness results for viscosity solutions of PDE cannot be used.

Here our approach consists of proposing a definition for solutions to Hamilton–Jacobi equations and to prove a uniqueness result for the Hamilton–Jacobi equations coming from differential games with Isaacs's condition. Our main aim is not to add a new concept to the already numerous notions of generalized solutions of PDEs but to prove the existence of a value of the game in the discontinuous case using the definition of solution.

Our second main interest is optimal control systems with discontinuous end-cost which leads to the following definition of the value function:[1]

$$W_g(t_0, x_0) = \inf_{u \in \mathcal{U}(t)} g(x(T;\, t_0, x_0, u)), \tag{3}$$

where $x(\cdot;\, t_0, x_0, u)$ is solution to the Cauchy problem

$$x'(t) = f(t, x(t), u(t)), \quad x_0 = x(t_0). \tag{4}$$

Our main result for control says that W_g is the unique solution to the corresponding Hamilton–Jacobi–Bellman equation for arbitrary discontinuous terminal cost g.

The proposed definition of solutions of Hamilton–Jacobi equations and our methods of proof are motivated by three already known approaches we described below.

First, Frankowska has observed that some invariance property of the epigraph and/or hypograph of the value function for control can be used to define a notion of solution to some Hamilton–Jacobi equation [14]. The author has used this fact to characterize the value function of the Mayer problem in the lower semicontinuous case in [15]. Such type of monotonicity properties of the value function along trajectories (or equivalently invariance or viability [1] of the epigraph and/or hypograph) has also been studied in the spirit of [15] in some works; among them we quote [6], [7], [8], [9], [10], [11], [16].

Second, we consider the comparison principle [12] and Barles–Perthame stability result [3] for viscosity super- and subsolutions. Since the beginning of the theory where viscosity solutions were (bounded uniformly) continuous, there was, until now, a constant effort to develop notions of discontinuous viscosity solutions. Without doing an exhaustive history of these theories, let us mention the Ishii solution (well exposed in [3]) based on semicontinuous envelopes of functions, Barron–Jensen semicontinuous

[1] One can easily check that the definition (3) of W_g does not require the use of the closure of the reachable set used in (2).

solutions [5], [2] for convex Hamiltonians, and envelope[2] solutions [4] which are related to Subbotin solution.

Third, we consider the minimax solution introduced by Subbotin (cf. [21] and its bibliography; see also [19]). This concept of solution is related to an interpretation of the Hamilton–Jacobi equation through a dynamic game with Isaacs's condition. In [21], the existence and the uniqueness of a semicontinuous minmax solution to the Hamilton–Jacobi–Isaacs equation are proved; this allows us to deduce the existence of the value of the game in the context of positional strategy [17].

The novelty of the results presented here are mainly the existence of the value for differential games with nonanticipative strategies and semicontinuous terminal cost, the characterization of the semicontinuous game value function as the unique solution of the Hamilton–Jacobi–Isaacs equation, and the characterization of "fully" discontinuous value function as the unique solution of the Hamilton–Jacobi–Bellman equation for optimal control.

From the point of view of PDEs, which is not our main topic, one can consider our work as existence and uniqueness results for PDEs with convex Hamiltonian and "fully" discontinuous boundary condition and as existence and uniqueness results for PDEs with nonnecessary convex Hamiltonian and semicontinuous boundary condition. In fact, at the end of the paper [2], there is an illuminating example showing for a PDE with nonconvex Hamiltonian that many Ishii solutions can exist; we discuss this example and prove the existence and uniqueness of solution in the context of the notion of solution used in the present paper. Also our work says that g is a semiresolutive function for the Hamilton–Jacobi–Bellman equation in the meaning of [18].

Let us explain how the paper is organized. In the first section, we introduce some preliminaries and we state a result concerning the discontinuous Mayer problem for control. In the second section, we present our main results concerning games. The last section is devoted to an appendix with the technical proofs of our claims.

2. Preliminaries. Consider the measurable functions $u : [t_0, T] \mapsto U$ and $v : [t_0, T] \mapsto V$. Let U and V be compact metric spaces. Let us denote by $\mathcal{U}(t_0)$ and $\mathcal{V}(t_0)$ the set of such measurable controls. We say that a map $\alpha : \mathcal{V}(t_0) \to \mathcal{U}(t_0)$ is a nonanticipative strategy of the first player if for every control v_1, $v_2 \in \mathcal{V}(t_0)$ such that

$$v_1(s) = v_2(s) \text{ for almost all } s \in [t_0, \tau],$$

we have

$$\alpha(v_1)(s) = \alpha(v_2)(s) \text{ for almost all } s \in [t_0, \tau].$$

We say that a map $\beta : \mathcal{U}(t_0) \to \mathcal{V}(t_0)$ is a nonanticipative strategy of the second player if for every controls u_1, $u_2 \in \mathcal{U}(t_0)$ such that

$$u_1(s) = u_2(s) \text{ for almost all } s \in [t_0, \tau],$$

we have

$$\beta(u_1)(s) = \beta(u_2)(s) \text{ for almost all } s \in [t_0, \tau].$$

[2] We thank the anonymous referee for pointing out that the notion of solution used in the present paper is similar to envelope solution.

Let $\Gamma(t_0)$ $(\Delta(t_0))$ denote the set of all such nonanticipative strategies of the first (second) player. Setting $A_\alpha(t_0,x_0) = \{x(T;\, t_0,x_0,\alpha(v),v) :\ v \in \mathcal{V}(t_0)\}$, $B_\beta(t_0,x_0) = \{x(T;\, t_0,x_0,u,\beta(u)) :\ u \in \mathcal{U}(t_0)\}$ the reachable sets, this enables us to define the value functions V_g^+ and V_g^- by relation (2). We assume that $f : [0,T]\times\mathbb{R}^n\times U\times V \to \mathbb{R}^n$ satisfies the following:

$$(5)\qquad \left\{\begin{array}{l} f(\,\cdot,\,\cdot\,,u,v) \text{ is Lipschitz continuous,}\\ f(t,x,\,\cdot,\,\cdot\,) \text{ is continuous,}\\ f \text{ has a linear growth, i.e.,}\\ \sup_{(t,u,v)} \|f(t,x,u,v)\| \le a(1+\|x\|)\\ \text{for some given } a>0. \end{array}\right.$$

For convenience, we do not repeat the same assumption in the control case viewed as a particular differential game where f does not depend on v.

Let us recall the Isaacs condition

$$(6)\qquad \left\{\begin{array}{c} \min_{u\in U}\max_{v\in V}\langle f(t,x,u,v),p\rangle = \max_{v\in V}\min_{u\in U}\langle f(t,x,u,v),p\rangle\\ \text{for every } t,\ x, \text{ and } p\in\mathbb{R}^n. \end{array}\right.$$

Throughout the paper, we assume that

$$(7)\qquad f(t,x,U,v) \ \text{ is convex for every } t,\ x,\ v$$

and

$$(8)\qquad f(t,x,u,V) \ \text{ is convex for every } t,\ x,\ u$$

hold true.

Let $g : \mathbb{R}^n \mapsto \mathbb{R}$ be a terminal cost.

If the terminal cost g is discontinuous, then so is the value function. To describe the value function as a unique solution to a corresponding Hamilton–Jacobi equation we introduce the following.

Definition 1. *Let $H : [0,T]\times\mathbb{R}^{2n} \to \mathbb{R}$ be a Hamiltonian. The function $(t,x)\mapsto u(t,x)$ is a solution to the following Hamilton–Jacobi equation with terminal condition:*

$$(9)\qquad \left\{\begin{array}{l} \dfrac{\partial u}{\partial t} + H\left(t,x,\dfrac{\partial u}{\partial x}\right) = 0,\\ u(T,x) = g(x),\ x\in\mathbb{R}^n, \end{array}\right.$$

if and only if

$$(10)\qquad \left\{\begin{array}{ll} \text{(i)} & u \textit{ is the supremum on the set of subsolutions}\\ & \phi \textit{ such that } \phi(T,x)\le g(x)\,\forall\, x\in\mathbb{R}^n,\\ \text{(ii)} & u \textit{ is the infimum on the set of supersolutions}\\ & \psi \textit{ such that } \psi(T,x)\ge g(x)\,\forall\, x\in\mathbb{R}^n. \end{array}\right.$$

The above meaning of solution is similar to envelope solution introduced in [4].

Here we call supersolution[3] any lower semicontinuous function $\psi : (0,T]\times\mathbb{R}^n \to \mathbb{R}$ such that

$$\forall (t,x) \in\ (0,T)\times\mathbb{R}^n,\ \forall\ (p_t,p_x)\in\ \partial_-\psi(t,x),\ p_t + H(t,x,p_x) \le 0,$$

[3] The definitions of the subdifferential $\partial_-\psi(t,x)$ and the superdifferential $\partial_+\phi(t,x)$ can be found in the appendix.

and we call subsolution any upper semicontinuous function $\phi : (0,T] \times \mathbb{R}^n \to \mathbb{R}$ such that

$$\forall (t,x) \in \ (0,T) \times \mathbb{R}^n,\ \forall\ (p_t, p_x) \in \partial_+\phi(t,x),\ p_t + H(t,x,p_x) \geq 0.$$

3. Discontinuous Mayer problem for control.

3.1. Main result for control.

We state the result obtained for the control case.

THEOREM 2. *Let $g : \mathbb{R}^n \mapsto \mathbb{R}$ be a bounded function. Assume that $f : [0,T] \times \mathbb{R}^n \times U \to \mathbb{R}^n$ satisfies* (5) *and* (7). *Then the value function $W_g : (0,T] \times \mathbb{R}^n \to \mathbb{R}$ given by*

$$W_g(t,x) = \inf_{u \in \mathcal{U}(t)} g(x(T;\, t,x,u))$$

is the unique generalized solution to the Hamilton–Jacobi–Bellmann equation (18) *where*

$$H(t,x,p) := \min_{u \in\ U} \langle f(t,x,u), p\rangle. \tag{11}$$

To avoid repetition of arguments, we postpone the proof until the appendix and we shall obtain this proof using results stated in the differential game context.

Remark 1. Theorem 2 is in fact the existence and uniqueness result for Hamilton–Jacobi equation (18) with Hamiltonian given by (11) and arbitrary terminal condition g. A uniqueness result in the case of lower semicontinuous g has been obtained in [5], [15] in the framework of different definitions of solutions. If g is lower semicontinuous, then the solutions in the meaning of Definition 1 as well as in the meaning of [5], [15] are equal to the value function W, so they coincide. We give an example of nonsemicontinuous g.

Example 1. Let $g : \mathbb{R} \to \mathbb{R}$ be the characteristic function of rationals. The dynamics $x' = f(t,x)$ of a system is given by a right-hand side that depends neither on u nor on v and satisfies (5). In this case the value $V(t_0,x_0) = g(x(T;\, t_0,x_0))$ is discontinuous at every point. In spite of this, by Theorem 2, V is the unique solution (in the sense of Definition 1) of the corresponding problem (18). Let us remark that the concepts of solution from [15] and [21] do not apply to the example.

3.2. Application to Mayer control problems with state constraints.

Let K be a closed subset of $\mathbb{R}^n$. We are interested in the characterization of the value function $W_g^K : [0,T] \times K \to R$

$$W_g^K(t_0,x_0) = \inf_{\left\{\begin{array}{l} u \in U(t_0), \\ x(t;t_0,x_0,u) \in K \text{ for } t \in [t_0,T) \end{array}\right.} g(x(T)) \tag{12}$$

as a unique solution to a Hamilton–Jacobi equation. In the literature there are many attempts to solve this problem (see [20], [16]). The minimal requirement guaranteeing that the function W_g^K is well defined by (12) is[4]

$$\left\{\begin{array}{l} \text{for any initial condition } (t_0,x_0) \in [0,T] \times K \text{ there exist} \\ \text{a control } u \in U(t_0) \text{ such that the solution } x(t;t_0,x_0,u) \\ \text{remains in set of constraints } K \text{ for every } t \in [t_0,T]. \end{array}\right. \tag{13}$$

[4] Property (12), called the viability property, can also be characterized by a geometrical condition in terms of Bouligand tangent cones (cf. [1]).

We provide a characterization of the value function W_g^K under assumption (13).

PROPOSITION 3. *Let $K \subset \mathbb{R}^n$ be closed and $g : \mathbb{R}^n \mapsto \mathbb{R}$ be a function bounded by $M > 0$. Assume that $f : [0,T] \times \mathbb{R}^n \times U \to \mathbb{R}^n$ satisfies* (5), (7) *and that* (13) *holds true for f, K. Then*

$$W_g^K(t,x) = U(t,x,0) \quad \text{for } x \in K,$$

where $U : [0,T] \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is the unique solution to

$$\left\{ \begin{array}{l} \frac{\partial U}{\partial t} + \tilde{H}(t,x,y,\frac{\partial U}{\partial x},\frac{\partial U}{\partial y}) = 0, \\ U(T,x,y) = g(x) + (M+1)\chi_{(0,\infty)}(y), \end{array} \right. \tag{14}$$

where $\tilde{H}(t,x,y,p_x,p_y) = \min_{u\in U}\langle f(t,xu),p_x\rangle + d_K(x)p_y$ and $\chi_{(0,\infty)}$ denotes the characteristic function of the open interval $(0,\infty)$.

Proof. We adopt the classical method of adding an extra variable (usually used to reduce a Bolza problem to a Mayer one) and the technique of penalization function. We consider a new control problem

$$\left\{ \begin{array}{l} x'(t) = f(t,x(t),u(t)), \\ y'(t) = d_K(x(t)), \end{array} \right.$$

where $d_K(x)$ denotes the distance from x to K. It is obvious that (5), (7) hold true for the extended control system. By Theorem 2, we obtain that the value function

$$\begin{array}{l} U(t_0,x_0,y_0) \\ = \inf_{u\in U(t_0)} g(x(T;t_0,x_0,u)) + (M+1)\chi_{(0,\infty)}\left(y_0 + \int_{t_0}^T d_K(x(t;t_0,x_0,u))dt\right) \end{array}$$

is the unique generalized solution to (14). On the other hand, just from the very definition, one can easily check that for every $x_0 \in K$ we have

$$W_g^K(t_0,x_0) = U(t_0,x_0,0). \qquad \square$$

4. Mayer problem for differential games.

4.1. First comparison result. Let us state our first result comparing the value functions following.

PROPOSITION 4. *If* (5), (6) *hold true and a terminal cost function g is locally bounded, then*

$$V_g^+(t,x) \le V_g^-(t,x)$$

for every $(t,x) \in [0,T] \times \mathbb{R}^n$.

The proof is a direct conclusion from the following lemma.

LEMMA 5. *Assume that* (5), (6) *hold true. Then*

$$\mathrm{cl}\ (A_\alpha(t_0,x_0)) \cap \mathrm{cl}(B_\beta(t_0,x_0)) \neq \emptyset$$

for each $\alpha \in \Gamma(t_0)$, $\beta \in \Delta(t_0)$.

Proof. Suppose to the contrary that there exist α_0, β_0 such that cl $(A_{\alpha_0}(t_0,x_0)) \cap$ cl $(B_{\beta_0}(t_0,x_0)) = \emptyset$.

Then there exists a Lipschitz continuous function $h : \mathbb{R}^n \to [0,1]$ such that $h(x) = 0$ for $x \in$ cl (A_{α_0}) and $h(x) = 1$ for $x \in$ cl (B_{β_0}). Hence,

$$V_h^-(t_0,x_0) = 0 < 1 = V_h^+(t_0,x_0).$$

This is a contradiction with Theorem 4.1 in [13] stating that if the terminal cost is Lipschitz continuous, then the upper value equals to the lower value. $\square$

To obtain the reverse inequality we shall study more deeply the Isaacs equation.

4.2. Values and Isaacs's equation. In the section we prove that any supersolution to Isaacs's equation is greater than the corresponding upper value and that any subsolution is smaller than the lower value. The main tool in proofs is viability theory (see [1]).

PROPOSITION 6. *Assume that* (5), (7) *hold true. Suppose that* $\psi : (0,T] \times \mathbb{R}^n \to \mathbb{R}$ *is lower semicontinuous and is a supersolution to*

$$\psi_t + H^-(t,x,\psi_x) = 0 \tag{15}$$

on $(0,T) \times \mathbb{R}^n$ *when*

$$H^-(t,x,p) = \max_{v\in V} \min_{u\in U} \langle f(t,x,u,v), p\rangle.$$

Then for every $(t_0, x_0) \in (0,T) \times \mathbb{R}^n$ *there exists a nonanticipative strategy* $\alpha \in \Gamma(t_0)$ *such that*

$$\psi(t_0,x_0) \geq \psi(t, x(t;\, t_0, x_0, \alpha(v), v)) \tag{16}$$

for every $v \in \mathcal{V}(t_0)$ *and* $t \in [t_0, T]$.

The proof of this proposition is postponed until the appendix.

COROLLARY 7. *Under the assumptions of Proposition* 6 *we obtain*

$$\psi(t,x) \geq V_g^-(t,x),$$

where $g(x) := \psi(T,x)$.

Proof. Since g is lower semicontinuous, then for every subset $A \subset \mathbb{R}^n$ we have $\sup\{g(x) :\, x \in A\} = \sup\{g(x) :\, x \in \mathrm{cl}\ (A)\}$. Thus

$$V_g^-(t_0,x_0) = \inf_{\alpha\in\Gamma(t_0)} \sup\{g(x) :\, x \in A_\alpha(t_0,x_0)\}.$$

By Proposition 6 we obtain

$$\psi(t_0,x_0) \geq \inf_{\alpha\in\Gamma(t_0)} \sup\{g(x) :\, x \in A_\alpha(t_0,x_0)\},$$

which gives us the desired inequality. □

PROPOSITION 8. *Assume that* (5), (8) *hold true. Suppose that* $\phi : (0,T] \times \mathbb{R}^n \to \mathbb{R}$ *is upper semicontinuous and is a subsolution to*

$$\phi_t + H^+(t,x,\phi_x) = 0$$

on $(0,T) \times \mathbb{R}^n$ *when*

$$H^+(t,x,p) = \min_{u\in U} \max_{v\in V} \langle f(t,x,u,v), p\rangle.$$

Then for every $(t_0, x_0) \in (0,T) \times \mathbb{R}^n$ *there exists a nonanticipative strategy* $\beta \in \Delta(t_0)$ *such that*

$$\phi(t_0,x_0) \leq \phi(t, x(t;\, t_0, x_0, u, \beta(u)))$$

for every $u \in \mathcal{U}(t_0)$ *and* $t \in [t_0, T]$.

The proof can be done using the same method as in the proof of Proposition 6.

Corollary 9. *Under the assumptions of Proposition* 8 *we obtain*

$$\phi(t,x) \leq V_h^+(t,x),$$

where $h(x) := \phi(T,x)$.

The proof is similar to the proof of Corollary 7.

If Isaacs's condition (6) holds true, then $H^- = H^+ (=: H)$ and previous results can be summarized in the following comparizon result (cf. Théorème d'unicité forte 4.10 in [3])

Proposition 10 (comparison result). *Assume that* (5)–(8) *hold true. Suppose that* $\psi : (0,T] \times \mathbb{R}^n \to \mathbb{R}$ *is lower semicontinuous and is a supersolution to*

$$\psi_t + H(t,x,\psi_x) = 0 \tag{17}$$

on $(0,T) \times \mathbb{R}^n$ *and* $\phi : (0,T] \times \mathbb{R}^n \to \mathbb{R}$ *is upper semicontinuous and is a subsolution to* (17) *on* $(0,T) \times \mathbb{R}^n$. *If* $\psi(T,x) \geq \phi(T,x)$ *for* $x \in \mathbb{R}^n$, *then* $\psi(t,x) \geq \phi(t,x)$ *for* $t \in (0,T]$ *and* $x \in \mathbb{R}^n$.

Proof. By Proposition 4 and Corollaries 7, 9, we have

$$\phi(t,x) \leq V_h^+(t,x) \leq V_h^-(t,x) \leq V_g^-(t,x) \leq \psi(t,x),$$

where $h(x) = \phi(T,x)$ and $g(x) = \psi(T,x)$ for $x \in \mathbb{R}^n$. □

4.3. Main result. In the section we prove the existence of value and characterize it by Isaacs's equation.

Theorem 11. *Assume that* (5)–(8) *hold true and* $g : \mathbb{R}^n \to \mathbb{R}$ *is a locally bounded lower semicontinuous function. Then the game has a value, i.e.,*

$$V_g^+ = V_g^- (=: V).$$

The value function V *is the smallest supersolution to the Hamilton–Jacobi–Isaacs equation*

$$V_t + H(t,x,V_x) = 0 \tag{18}$$

satisfying $V(T,x) \geq g(x)$ *when* $H := H^+ = H^-$.

Moreover, the value function V *is the unique generalized solution to* (18) *satisfying* $V(T,x) = g(x)$.

We have stated the result in the lower semicontinuous case. After typical reformulation it remains valid in the upper semicontinuous case. Before proving this result let us recall—in an adapted version—a result proved in [3] (cf. Theorem 4.1 in [3]), which plays an important role in the proof of our main theorem.

Lemma 12. *Assume that* $H : (0,T) \times \mathbb{R}^{2n} \to \mathbb{R}$ *is a continuous Hamiltonian. If* $w_n : (0,T) \times \mathbb{R}^n \to \mathbb{R}$ *is an increasing sequence of uniformly locally bounded supersolutions of a Hamilton–Jacobi equation*

$$\eta_t + H(t,x,\eta_x) = 0 \tag{19}$$

and $w : (0,T) \times \mathbb{R}^n \to \mathbb{R}$ *is a pointwise limit of* w_n, *then* w *is a supersolution of* (19).

Proof of Theorem 11.[5] We define a sequence $g_n : \mathbb{R}^n \to \mathbb{R}$ by

$$g_n(x) = \inf_{y \in \mathbb{R}^n} g(y) + n\|x-y\|.$$

[5] We would like to thank P. Cardaliaguet, who brought our attention to the inf-convolution method, which simplified the proof.

The inf-convolutions g_n are Lipschitz, $g_n(x) \leq g_{n+1}(x)$ and $\lim_n g_n(x) = g(x)$ for every $x \in \mathbb{R}^n$.

Using Evans–Souganidis characterization of value functions in the case of Lipschitz continuous terminal cost [13, Theorem 4.1], we have $V^+_{g_n} = V^-_{g_n} (:= V_n)$ and V_n is a viscosity solution (i.e., super- and subsolution) to (18).

Denote $W(t,x) = \lim_n V_n(t,x)$. By Lemma 12, W is a supersolution to (18). By Corollary 7, we obtain $W \geq V^-_g$. Since $V^+_g \geq V^+_{g_n}$, we deduce $V^+_g \geq W$. Hence

$$V^+_g \geq V^-_g.$$

Combining it with Proposition 4, we obtain $V^+_g = V^-_g = W$.

If $\psi : (0,T] \times \mathbb{R}^n \to \mathbb{R}$ is a supersolution to (18) and $\psi(T,x) \geq g(x)$, then $\psi \geq V^-_g$. Thus V is the smallest supersolution to (18) satisfying $V(T,x) \geq g(x)$.

Since V_n is a subsolution to (18), $V_n(T,x) \leq g(x)$ and $V = \lim_n V_n$, we obtain that V is a generalized solution to (18), $V(T,\cdot) = g(\cdot)$. □

Remark 2. Due to general properties of monotone approximation, V is also a solution to (18) in the Ishii sense. Namely, upper semicontinuous envelope of V coincides with the upper weak limit of V_n (cf. the exercise on page 91 in [3]), which by Theorem 4.1 in [3] is a subsolution of (18).

The following example with a slight modification is taken from [2]. It served in [2] as a counter-example to uniqueness of discontinuous solution—in the Ishii sense—to a Hamilton–Jacobi equation. Definition 1 is not equivalent to the notion of solution introduced by Ishii. In the example there exists a unique solution in the meaning of Definition 1 and there are several solutions in the Ishii sense [2].

Example 2. Let $U = V = [-1,1]$. We define $f : (-\infty, 0] \times \mathbb{R} \times U \times V \to \mathbb{R}$ by

$$f(t,x,u,v) = \chi_{(x \leq t)}(x-t)v + \chi_{(x \geq t)}(x-t)u.$$

It is easy to check that f satisfies (5)–(8) and the corresponding Hamiltonian is given by

$$H(t,x,p) = (x-t)|p|.$$

To define a terminal cost function $g : \mathbb{R} \to \mathbb{R}$, we fix $t_0 = x_0 < 0$. Let $b = x(0; t_0, x_0, u_1, v)$, $a = x(1; t_0, x_0, u_{-1}, v)$, where $u_1(t) = 1$, $u_{-1}(t) = -1$ for $t \in [t_0, 0]$, v is an arbitrary control. We define

$$g(x) = \begin{cases} 1 & \text{if } x \in (a,b), \\ -1 & \text{elsewhere.} \end{cases}$$

We set the terminal time T to be zero.

By Theorem 11, the value V for this game exists and is the *unique* solution to the corresponding Hamilton–Jacobi equation

$$\begin{cases} V_t + (x-t)|V_x| = 0, \\ V(0,x) = g(x) \text{ for every } x \in \mathbb{R}. \end{cases} \quad \square$$

Remark 3. The assumptions (7) and (8) concerning the convexity of the right-hand side are crucial for obtaining $V^+_g \geq V^-_g$ because we used a viability approach which requires convexity. We recall that thanks to Proposition 4, inequality $V^+_g \leq V^-_g$ holds true.

4.4. On the definition of the values of the game. In the definition of upper and lower values (2) we have used the closure of reachable sets. They can be defined as well without closure:

$$\begin{cases} U_g^-(t_0,x_0) := \inf_{\alpha\in\Gamma(t_0)} \sup\{g(x) : x \in A_\alpha(t_0,x_0)\}, \\ U_g^+(t_0,x_0) := \sup_{\beta\in\Delta(t_0)} \inf\{g(x) : x \in B_\beta(t_0,x_0)\}. \end{cases}$$

We shall exhibit an example where $U_g^- \neq U_g^+$.

Repeating the same arguments we can prove Corollaries 7, 9 for values $U^{+/-}$ instead of $V^{+/-}$ and show that $U_g^+ \geq U_g^-$. To obtain the reverse inequality it would be enough to know that $A_\alpha \cap B_\beta \neq \emptyset$ (see Proposition 4). The following example shows that in general it is not true. Thus we can define a terminal cost function g in such a way that $U_g^+ > U_g^-$.

Example 3. We provide an example of a differential game where $U_g^+ > U_g^-$. For doing this we construct a pair of nonanticipative strategies (α,β) such that

$$A_\alpha \cap B_\beta = \emptyset.$$

(Let us notice that this implies that neither A_α nor B_β are closed by Proposition 4.) We consider the following differential game on $\mathbb{R}^2$:

$$\begin{cases} x'(t) = u, \\ y'(t) = v, \end{cases}$$

where $U = V = [0,1]$. We set $x_0 = 0$, $t_0 = 0$ and $T = 1$. We denote by x_u (y_v) the solution to the Cauchy problem $x'(t) = u(t)$, $x(0) = 0$ (respectively, $y'(t) = v(t)$, $y(0) = 0$). We define the constant controls $u_0(t) = v_0(t) = 0$ and $u_1(t) = v_1(t) = 1$ for $t \in [0,1]$. For measurable functions w, $z : [0,1] \to [0,1]$ we define an (ultrametric) distance

$$\rho(w,z) = 1 - \max\{t \in [0,1] : \ w(s) = z(s) \text{ for almost everywhere } s \in [0,t]\}.$$

(This ultrametric distance has been used in [8] and [11] to prove Lemma 14.)

Set $B = \{u \in \mathcal{U}(0) : \ \rho(u,u_0) < 1\}$ and $S = \{u \in \mathcal{U}(0) : \ \rho(u,u_0) = 1\}$. Define two nonanticipative strategies α, β as follows:

$$\alpha(v) = \begin{cases} u_0 & \text{if} \quad v \in B, \\ u_1 & \text{if} \quad v \in S, \end{cases}$$

$$\beta(u) = \begin{cases} v_1 & \text{if} \quad u \in B, \\ v_0 & \text{if} \quad u \in S. \end{cases}$$

If $u \in S$, then $x_u(1) > 0$. If $p \in (0,1]$, then there exists a control $u \in S$ such that $x_u(1) = p$.

If $u \in B$, then $x_u(1) < 1$. If $p \in [0,1)$, then there exists a control $u \in B$ such that $x_u(1) = p$.

We have

$$\begin{aligned} A_\alpha &= \{(x_{\alpha(v)}(1), y_v(1)) : \ v \in B\} \cup \{(x_{\alpha(v)}(1), y_v(1)) : \ v \in S\} \\ &= \{0\} \times [0,1) \cup \{1\} \times (0,1] \end{aligned}$$

and

$$\begin{aligned} B_\beta &= \{(x_u(1), y_{\beta(u)}(1)) : \ u \in B\} \cup \{(x_u(1), y_{\beta(u)}(1)) : \ u \in S\} \\ &= [0,1) \times \{1\} \cup (0,1] \times \{0\}. \end{aligned}$$

Setting $g = \chi_{B_\beta}$ we obtain $U_g^+(0,0) = 1 > 0 = U_g^-(0,0)$. □

We did not succeed in finding an example where g is semicontinuous. Hence the question of knowing if $U_g^- = U_g^+$ (so, it would be equal to $V_g^- = V_g^+$) for semicontinuous g remains an open problem.

Appendix. We recall some necessary notions and facts from nonsmooth analysis.

Let $K \subset \mathbb{R}^n$ be a nonempty subset and $x_0 \in K$. The contingent cone to K at x_0, $T_K(x_0)$, is defined by

$$v \in T_K(x_0) \iff \liminf_{h \to 0^+} \frac{dist(x_0 + hv, K)}{h} = 0.$$

A polar cone T^- to a subset $T \subset \mathbb{R}^n$ is defined by

$$T^- = \{p \in \mathbb{R}^n : \ \forall v \in T, \ \langle p, v \rangle \leq 0\}.$$

Let $\Omega \subset \mathbb{R}^n$ be an open subset and $w : \Omega \to \mathbb{R}$ be a lower semicontinuous function. The subdifferential of w at $x_0 \in \Omega$ is given by

$$\partial_- w(x_0) = \left\{ p \in \mathbb{R}^n : \ \liminf_{x \to x_0} \frac{w(x) - w(x_0) - \langle p, x - x_0 \rangle}{\|x - x_0\|} \geq 0 \right\}.$$

It is well known that

$$p \in \partial_- w(x_0) \iff (p, -1) \in \left[T_{Epi(w)}(x_0, w(x_0)) \right]^-,$$

where *Epi* stands for the epigraph. For an upper semicontinuous function w we define a superdifferential by

$$\partial_+ w(x_0) = \left\{ p \in \mathbb{R}^n : \ \limsup_{x \to x_0} \frac{w(x) - w(x_0) - \langle p, x - x_0 \rangle}{\|x - x_0\|} \leq 0 \right\}$$

and we have

$$p \in \partial_+ w(x_0) \iff (-p, 1) \in \left[T_{Hypo(w)}(x_0, w(x_0)) \right]^-,$$

where *Hypo* stands for the hypograph.

The following Rockafellar result (see [15]) gives more information about the connection between subgradients and normals to epigraph.

LEMMA 13. *Consider a lower semicontinuous function* $w : \Omega \to \mathbb{R}$ *and* $x_0 \in \Omega$. *If* $(p, 0) \in \left[T_{Epi(w)}(x_0, w(x_0)) \right]^-$, *then there exist* $x_n \to x_0$, $p_n \to p$, $q_n \to 0$, $q_n < 0$ *such that*

$$(p_n, q_n) \in \left[T_{Epi(w)}(x_n, w(x_n)) \right]^-.$$

The following result can be deduced from [8, Theorem 2.1].

LEMMA 14. *Assume that* $f : \mathbb{R}^n \times U \times V \to \mathbb{R}^n$ *satisfies* (5), (7) *and a subset* $D \subset \mathbb{R}^n$ *is closed. If for every* $z \in D$ *we have*

$$\forall p \in \left[T_D(z) \right]^-, \ \max_{v \in V} \min_{u \in U} \langle f(z, u, v), p \rangle \leq 0, \tag{20}$$

then for every $z_0 \in D$, $t_0 < T$, *there exists a nonanticipative strategy* $\alpha : \mathcal{V}(t_0) \to \mathcal{U}(t_0)$ *such that for every* $v \in \mathcal{V}(t_0)$ *we have*

$$z(t; t_0, z_0, \alpha(v), v) \in D$$

for $t \in [t_0, T]$*, where* $z(\cdot;\, t_0, z_0, u, v)$ *denotes the solution to the problem*

$$\left\{ \begin{array}{l} z' = f(z, u(t), v(t)), \\ z(t_0) = z_0. \end{array} \right.$$

Proof of Proposition 6. Fix $t_0 \in (0, T)$. We set

$$D_\psi := \mathrm{cl}(\{(t, x, r) :\, t \in (0, T],\ x \in \mathbb{R}^n,\ r \geq \psi(t, x)\}) \cup [T, +\infty) \times \mathbb{R}^n,$$

$$\tilde{f}(t, x, r, u, v) = \left\{ \begin{array}{lll} 0 & \text{if} & t < 0, \\ \frac{t}{t_0}(1, f(t, x, u, v), 0) & \text{if} & t \in [0, t_0], \\ (1, f(t, x, u, v), 0) & \text{if} & t \in (t_0, T], \\ (1, f(T, x, u, v), 0) & \text{if} & t > T. \end{array} \right.$$

We show that (20) holds true for $\tilde{f}$, D_ψ.

Let $z_0 = (s_0, x_0, r_0) \in D_\psi$. If $s_0 = 0$, then $\tilde{f} = 0$. Obviously, (20) holds true. If $s_0 \geq T$ and $(p_s, p_x, p_r) \in \left[T_{D_\psi}(s_0, x_0, r_0)\right]^-$, then $p_s \leq 0$, $p_x = 0$, $p_r = 0$. Hence, (20) holds true.

It remains to consider the case $s_0 \in (0, T)$. We have $\left[T_{D_\psi}(s_0, x_0, r_0)\right]^- \subset \left[T_{D_\psi}(s_0, x_0 \psi(s_0, x_0))\right]^-$. Let $(p_s, p_x, p_r) \in \left[T_{D_\psi}(s_0, x_0 \psi(s_0, x_0))\right]^-$. If $p_r < 0$, then $(\frac{p_s}{-p_r}, \frac{p_x}{-p_r}) \in \partial_- \psi(s_0, x_0)$. Since ψ is a supersolution to (15) we have

$$\frac{p_s}{-p_r} + \max_{v \in V} \min_{u \in U} \left\langle f(s_0, x_0, u, v), \frac{p_x}{-p_r} \right\rangle \leq 0.$$

Hence,

$$\max_{v \in V} \min_{u \in U} \langle \tilde{f}(s_0, x_0, r_0, u, v), (p_s, p_x, p_r) \rangle \leq 0.$$

Now, we consider the case $p_v = 0$. By Lemma 13, there exist $s_n \to s_0$, $x_n \to x_0$, $p_{sn} \to p_s$, $p_{xn} \to p_x$, $p_{rn} \to 0$, $p_{rn} < 0$ such that

$$(p_{sn}, p_{xn}, p_{rn}) \in \left[T_{Epi(\psi)}(s_n, x_n, \psi(t_n, x_n))\right]^-.$$

Since $p_{rn} < 0$, from what we already have proved, we obtain

$$\max_{v \in V} \min_{u \in U} \langle \tilde{f}(s_n, x_n, \psi(s_n, x_n), u, v), (p_{sn}, p_{xn}, p_{rn} \rangle \leq 0.$$

Since f is continuous and U is compact, we have

$$\max_{v \in V} \min_{u \in U} \langle \tilde{f}(s_0, x_0, r_0, u, v), (p_s, p_x, p_r \rangle \leq 0.$$

By Theorem 14, there exists $\alpha \in \Gamma(t_0)$ such that for every $v \in \mathcal{V}(t_0)$

$$z(s;\, t_0, z_0, \alpha(v), v) \in D_\psi \tag{21}$$

for every $s \in [t_0, T]$, where $z(s;\, t_0, z_0, u, v)$ denotes the solution to the Cauchy problem

$$\left\{ \begin{array}{l} z'(s) = \tilde{f}(z(s), u(s), v(s)), \\ z(t_0) = z_0. \end{array} \right.$$

Let $z(s, t_0, z_0, \alpha(v), v) = (t(s), x(s), r(s))$. By the definition of $\tilde{f}$, we have $t(s) = s$, $x(s) = x(s;\, t_0, x_0, \alpha(v), v)$, $r(s) = \psi(t_0, x_0)$. It yields (16) for $t \in [t_0, T)$. Since ψ is lower semicontinuous, we obtain (16) for $t = T$. $\square$

Let us deduce now our result concerning the control problem.

Proof of Theorem 2. Fix $(t_0, x_0) \in (0, T] \times \mathbb{R}^n$. Let $\varepsilon > 0$. There exists $u_\varepsilon \in \mathcal{U}(t_0)$ such that $g(x(T;\, t_0, x_0, u_\varepsilon)) < W_g(t_0, x_0) + \varepsilon$. We define $h : \mathbb{R}^n \to \mathbb{R}$ by

$$h(x) = \begin{cases} g(x) & \text{for} \quad x = x(T;\, t_0, x_0, u_\varepsilon), \\ M & \text{for} \quad x \neq x(T;\, t_0, x_0, u_\varepsilon), \end{cases}$$

where M is a bound of $\|g\|$. Obviously, h is lower semicontinuous. By Theorem 11, the value W_h is a supersolution to (18). We have $W_h(t_0, x_0) < W_g(t_0, x_0) + \varepsilon$. Hence,

$$W_g(t_0, x_0) = \inf\{\psi(t_0, x_0) : \ \psi \text{ is a supersolution to (18)}, \ \psi(T, \cdot) \geq g(\cdot)\}.$$

We define $l : \mathbb{R}^n \to \mathbb{R}$ by

$$l(x) = \begin{cases} W_g(t_0, x_0) & \text{if} \quad x \in \{x(T;\, t_0, x_0, u) : \ u \in \mathcal{U}(t_0)\}, \\ -M & \text{if} \quad x \notin \{x(T;\, t_0, x_0, u) : \ u \in \mathcal{U}(t_0)\}. \end{cases}$$

By (5), (7), the reachable set $\{x(T;\, t_0, x_0, u) : \ u \in \mathcal{U}(t_0)\}$ is closed. Thus, l is upper semicontinuous. Obviously, we have $W_g(t_0, x_0) = W_l(t_0, x_0)$. By Theorem 11 (in a version for upper semicontinuous terminal cost), we obtain that W_l is a subsolution to (18). Hence,

$$W_g(t_0, x_0) = \sup\{\phi(t_0, x_0) : \ \phi \text{ is a subsolution to (18)}, \ \phi(T, \cdot) \leq g(\cdot)\}. \qquad \square$$

REFERENCES

[1] J.-P. Aubin, *Viability Theory*, Birkhäuser, Boston, 1992.

[2] G. Barles, *Discontinuous viscosity solutions of first-order Hamilton-Jacobi equations: A guided visit*, Nonlinear Anal., 20 (1993), pp. 1123–1134.

[3] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, Springer, Paris, 1994.

[4] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[5] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonian*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[6] L. Berkovitz, *Characterizations of the values of differential games*, Appl. Math. Optim., 17 (1988), pp. 177–183.

[7] L. Berkovitz, *The existence of value and saddle point in games of fixed duration*, SIAM J. Control Optim., 23 (1985), pp. 172–196.

[8] P. Cardaliaguet, *A differential game with two players and one target*, SIAM J. Control Optim., 34 (1996), pp. 1441–1460.

[9] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Numerical methods for optimal control and numerical games*, in Stochastic and Differential Games, Theory and Numerical Methods (Annals of International Society of Dynamical Games), M. Bardi and R. Parthasarathy, eds., Birkhäuser, Boston, 1999, pp. 177–249.

[10] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Minimal times for constrained nonlinear control problems without local controllability*, Appl. Math. Optim, 36 (1997), pp. 21–42.

[11] P. Cardaliaguet and S. Plaskacz, *Invariant solutions for differential games*, in Topology in Nonlinear Analysis, Banach Center Publ. 35, Polish Acad. Sci., Warsaw, 1996, pp. 149–158.

[12] Crandall M., H. Ishii, and P.-L. Lions, *User's guide to the viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1992), pp. 487–502.

[13] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[14] H. Frankowska, *Equations d'Hamilton-Jacobi contingentes*, C. R. Acad. Sci. Paris Sér. I Math., 304 (1987), pp. 295–298.

[15] H. Frankowska, *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[16] H. Frankowska and S. Plaskacz, *Semicontinuous solutions of Hamilton-Jacobi-Bellman equations with state constraints*, in Differential Inclusions and Optimal Control, Lecture Notes in Nonlinear Analysis, Vol. 2, Juliusz Schauder Center for Nonlinear Studies, Torun, Poland, 1998, pp. 145–161.

[17] N. N. Krasovskii and A. I. Subbotin, *Game-Theoretical Control Problems*, Springer, New York, 1988.

[18] M. Ramaswamy and S. Ramaswamy, *Perron's method and barrier functions for the viscosity solutions of the Dirichlet problem for some non-linear partial differential equations*, Z. Anal. Anwendungen 13 (1994), pp. 199–207.

[19] M. A. Rozyev and A. I. Subbotin, *Semicontinuous solutions of Hamilton-Jacobi equations*, J. Appl. Math. Mech., 52 (1988), pp. 141–146; translation from Prikl. Mat. Mekh., 52 (1988), pp. 179–185.

[20] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[21] A.I. Subbotin, *Generalized Solutions of First-Order PDEs. The Dynamical Optimization Perspective*, Birkhäuser, Boston, 1995.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1499–1519

ADAPTIVE LQG CONTROL OF INPUT-OUTPUT SYSTEMS—A COST-BIASED APPROACH*

M. PRANDINI† AND M. C. CAMPI†

Abstract. In this paper, we consider linear systems in input-output form and introduce a new adaptive linear quadratic Gaussian (LQG) control scheme which is shown to be self-optimizing. The identification algorithm incorporates a cost-biasing term, which favors the parameters with smaller LQG optimal cost and a second term that aims at moderating the time-variability of the estimate. The corresponding closed-loop scheme is proven to be stable and to achieve an asymptotic LQG cost equal to the one obtained under complete knowledge of the true system (self-optimization).

The results of this paper extend in a nontrivial way previous results established along the cost-biased approach in other settings.

Key words. LQG adaptive control, least squares identification, cost-biased identification, self-optimality

AMS subject classifications. 93E20, 93E15, 93E24, 49L20

PII. S0363012999366369

1. Introduction.

Since the appearance of the original contribution of Aström and Wittenmark [1], the analysis of self-tuning control systems has constituted a challenging topic for theorists working in the area of adaptive control. The first significant convergence results were obtained in the late 1970s for minimum-variance control schemes. In particular, a global convergence result for an adaptive control system based on the stochastic gradient algorithm has been established in [13]. Extensions to the least squares (LS) algorithm are dealt with in [32] by introducing a suitable modification to the standard recursive least squares algorithm. Such a modification is in fact not necessary in order to achieve optimality [20].

The common result of all the above-mentioned contributions is that a minimum-variance self-tuning control system obtains under various operating conditions the same performance as the one achievable under complete knowledge of the true plant (*self-optimization*). It is important, however, to emphasize that the minimum-variance control law calls for the restrictive—and often unrealistic—assumption that the plant is minimum-phase. Extending these results to more general control techniques suitable for nonminimum-phase plants has attracted much attention in the last decade. The corresponding analysis, however, is far more complex.

It is by now well known (see, e.g., [21, 22, 18, 25, 33]) that the self-optimization result does not hold true for general control laws based on the minimization of multistep performance indexes. As a matter of fact, the interplay between identification and control in a certainty equivalence adaptive control scheme may result in the convergence of the parameter estimate to a parameterization different from the true one in absence of suitable excitation conditions (see, e.g., [5, 18, 2, 6]). When a cost criterion other than the output variance is considered, this identifiability problem results in a strictly suboptimal performance. In particular, the identifiability problem

*Received by the editors December 20, 1999; accepted for publication July 28, 2000; published electronically January 25, 2001. Research supported by MURST under the projects on "Identification and Control of Industrial Systems" and "Synthesis of Adaptive and Robust Controllers."

http://www.siam.org/journals/sicon/39-5/36636.html

†Dipartimento di Elettronica per l'Automazione, Università degli Studi di Brescia, Via Branze 38, 25123 Brescia, Italy (prandini@ing.unibs.it, campi@ing.unibs.it).

is significant in infinite-horizon linear quadratic Gaussian (LQG) control and, in fact, in [33] it is proven that for a state space system subject to Gaussian noise the set of the parameterizations leading to optimality of LQG control is strictly contained in the set of the potential convergence points.

A first approach to achieve optimality consists of securing the parameter consistency by introducing suitable probing signals in the control system. The probing signals should be sufficiently exciting so that consistency is achieved, and—at the same time—mild enough in order not to degrade the control system performance. In [8, 9, 10, 14, 28], this is obtained by a careful selection of an asymptotically vanishing dither noise. This approach is useful only in the case when noise injection is feasible.

A second approach—adopted in this paper—is based on the so-called *cost-biased method* originally introduced in [21]. In order to better focus on the basic idea underlying this approach and to highlight the main contributions given in the present paper, we proceed as follows: first we introduce the dynamic systems we consider; then we outline the cost-biased approach with specific reference to our class of systems; finally we put our results into perspective with the other existing results obtained along the cost-biased approach.

We consider dynamic systems in input-output form described by the following equation:

$$\mathcal{A}(\vartheta^\circ; q^{-1})\, y_t = \mathcal{B}(\vartheta^\circ; q^{-1})\, u_{t-1} + n_t, \tag{1.1}$$

where $\mathcal{A}(\vartheta^\circ; q^{-1}) = 1 - \sum_{i=1}^{n} a_i^\circ q^{-i}$ and $\mathcal{B}(\vartheta^\circ; q^{-1}) = \sum_{i=1}^{m} b_i^\circ q^{-i+1}$ are polynomials in the unit-delay operator q^{-1} and $\vartheta^\circ = [\,a_1^\circ\ a_2^\circ \dots a_n^\circ\ b_1^\circ\ b_2^\circ \dots b_m^\circ\,]^T$ is the system parameter vector. The control objective is to minimize the quadratic cost

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} [\, y_t^2 + \beta\, u_t^2], \tag{1.2}$$

where the control weighting coefficient β is strictly positive.

The basic idea of the cost-biased approach can be outlined as follows.

Suppose a standard LS algorithm is used for the identification of system (1.1) and let $\hat{\vartheta}_t^{LS}$ be the corresponding LS estimate at time t. According to the certainty equivalence principle, the control action is obtained by the relation $u_t = u_t^\star(\hat{\vartheta}_t^{LS})$, where $u_t^\star(\vartheta)$ indicates the optimal LQG control law for system (1.1) with parameter ϑ. For ease of reference, let us introduce the symbol $S(\vartheta_1, \vartheta_2)$ for the control system formed by system (1.1) with parameter ϑ_1 with the loop closed by $u_t = u_t^\star(\vartheta_2)$. Since the identification is performed in closed-loop, it is expected that the behavior of $S(\vartheta^\circ, \hat{\vartheta}_t^{LS})$ will be the same, at least in the long run, as the one of $S(\hat{\vartheta}_t^{LS}, \hat{\vartheta}_t^{LS})$. Then, the LQG cost for $S(\vartheta^\circ, \hat{\vartheta}_t^{LS})$—i.e., the incurred cost—will be the same as the LQG cost for $S(\hat{\vartheta}_t^{LS}, \hat{\vartheta}_t^{LS})$. However, one should note that the latter configuration is optimal for the estimated model, whereas the incurred cost obviously cannot be lower than the optimal cost for the true system. From this, one concludes that the least squares algorithm has a natural tendency to return estimates with an optimal cost that is not smaller than the optimal cost associated with the true system and that, when it is strictly larger, the adaptive scheme attains a suboptimal performance.

In the cost-biased approach an extra term that favors parameters with smaller optimal cost is added to the LS identification cost. This extra term is selected with a twofold objective. On the one hand, it should be strong enough so that the optimal LQG cost associated with the estimated model is asymptotically not larger than the optimal cost for the true system. If, on the other hand, it is so mild that the closed-loop identification property $S(\vartheta^\circ, \hat{\vartheta}_t^{LS}) = S(\hat{\vartheta}_t^{LS}, \hat{\vartheta}_t^{LS})$ is preserved, then one still has that the incurred cost is equal to the cost for $S(\hat{\vartheta}_t^{LS}, \hat{\vartheta}_t^{LS})$. From this, optimality of the adaptive control scheme is achieved.

The cost-biased approach has been successfully applied in a number of different settings. Controlled Markov chains with a finite parameter set are considered in [21]. The results of this paper have been extended to Markov chains with an infinite parameter set in [24] and to systems with a general state space but still with a finite parameter set in [19].

Linear systems in a state space representation are dealt with in [18] and [7]. In these papers, the restrictive assumption that the state is fully accessible is made. Moreover, it is assumed that the noise system affects all state variables. This assumption is crucial for the correct functioning of the proposed identification procedure. As a matter of fact, the presence of a full-range noise sheds light on the existing difference between the true system and the estimated model and this helps the identification task. In the paper [7], it is in fact shown that this mechanism is effective enough so as to counteract the biasing effect of the cost-biasing term thus guaranteeing the closed-loop identification property. Unfortunately, the assumption that the noise is full-range is so restrictive that it cannot be applied to many situations of interest. In particular, a state space realization of the input-output system (1.1) does not satisfy this condition.

In the present paper, an optimal adaptive control scheme for system (1.1) still based on the cost-biasing idea is presented. Extending the cost-biased approach to systems as (1.1) is important in that input-output systems are largely used in adaptive control applications. Moreover, assuming only the input and output measurability is much more realistic than assuming full state accessibility. As a side remark we also note that, in contrast with [18] and [7], our approach does not require the noise to be Gaussian.

The paper is organized as follows: in section 2, we describe the cost-biased adaptive LQG control scheme and recall some relevant properties of the standard LS estimates. The study of the cost-biased identification algorithm is presented in section 3. Section 4 is devoted to the analysis of the closed-loop stability and the characterization of the self-tuning LQG control performance. Finally, section 5 presents conclusions and suggestions for future research.

2. The cost-biased adaptive LQG control system.

2.1. The LQG optimal control problem. In this section, we summarize some known facts on infinite-horizon LQG control relevant for the subsequent developments. This is also useful in order to introduce the assumptions and the notations we shall use throughout the paper.

Consider the discrete time single input, single output (SISO) system (1.1) where signal n_t is a stochastic disturbance precisely described in the following.

ASSUMPTION 2.1. *$\{n_t\}$ is a martingale difference sequence with respect to a filtration $\{\mathcal{F}_t\}$, satisfying the following conditions:*

1. $\sup_t E[|n_t|^p/\mathcal{F}_{t-1}] < \infty$ *almost surely (a.s)* $\forall p > 0$;
2. $\lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} n_t^2 = \sigma^2 > 0$ *a.s.*

Note that Assumption 2.1 is satisfied when $\{n_t\}$ is an independently and identically distributed (i.i.d.) Gaussian sequence, but it includes many other situations.

We make the assumption on system (1.1) that $n > 0$ (nontrivial autoregressive part). Note that if $n = 0$, the trivial control law $u_t = 0$, $t \geq 0$, is obviously optimal irrespective of the value of ϑ°.

We further assume that system (1.1) belongs to a known set of stabilizable models according to the following.

ASSUMPTION 2.2. *$\vartheta^\circ \in \Theta$, where Θ is a compact set such that $\Theta \subset \{\vartheta \in \Re^{n+m} : q^s\mathcal{A}(\vartheta; q^{-1})$ and $q^{s-1}\mathcal{B}(\vartheta; q^{-1})$ have no unstable pole-zero cancellations$\}$, $s = \max\{n, m\}$ being the order of the system.*

System (1.1) is initialized at time $t = 0$ with $y_t = u_{t-1} = 0$, $t \leq 0$.

For the determination of an optimal control law, it is convenient to represent system (1.1) in a state space form such that the state is accessible and then apply the well-known solution to the optimal LQG control problem for full state accessible state space systems (see, e.g., [10], [3]).

Defining $x_t := [y_t \; y_{t-1} \ldots y_{t-(n-1)} \; u_{t-1} \; u_{t-2} \ldots u_{t-(m-1)}]^T$, system (1.1) can be given the following state space representation of order $\bar{s} := n + m - 1$

$$\left\{ \begin{array}{l} x_{t+1} = A(\vartheta^\circ)x_t + B(\vartheta^\circ)u_t + Cn_{t+1}, \quad x_0 = [0\,0\ldots 0]^T, \\ y_t = Hx_t \end{array} \right. \tag{2.1}$$

with matrices

$$A(\vartheta) = \left[\begin{array}{cccc|cccc} a_1 & \cdots & a_{n-1} & a_n & b_2 & \cdots & b_{m-1} & b_m \\ 1 & 0 & \cdots & & 0 & \cdots & & 0 \\ & \ddots & \ddots & & & \ddots & & 0 \\ & & 1 & 0 & & & & 0 \\ \hline 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 1 & 0 & & \\ & \ddots & \ddots & & & \ddots & \ddots & \\ & & 0 & 0 & & & 1 & 0 \end{array} \right],$$

$$B(\vartheta) = \left[\begin{array}{c} b_1 \\ 0 \\ \vdots \\ 0 \\ \hline 1 \\ 0 \\ \vdots \\ 0 \end{array} \right], \quad C = \left[\begin{array}{c} 1 \\ 0 \\ \vdots \\ 0 \\ \hline 0 \\ 0 \\ \vdots \\ 0 \end{array} \right], \quad H = \left[\begin{array}{cccc|ccc} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{array} \right].$$

In this way, the LQG regulation problem for the system in input-output representation (1.1) is reformulated as a complete state information control problem where the performance index to be minimized is given by $\limsup_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} [x_t^T T x_t + \beta u_t^2]$, where $T = H^T H \geq 0$ and $\beta > 0$.

Note that, in the case when $n > 1$ and $m > 1$, the state space representation (2.1) of system (1.1) is nonminimal (the order of system (1.1) is $s = \max\{n, m\}$, whereas the dimension of matrix $A(\vartheta^\circ)$ is $\bar{s} = n + m - 1$). However, from the block triangular matrix structure of $A(\vartheta^\circ)$ it is easily seen that the added eigenvalues are identically equal to zero. Hence from Assumption 2.2 it follows that $(A(\vartheta^\circ), B(\vartheta^\circ))$

is stabilizable and $(A(\vartheta^\circ), H)$ is detectable and the standard approach based on the solution to a Riccati equation can be used to determine the control law.

Specifically, the solution to the original LQG control problem has the following expression [10]:

$$u_t = \mathcal{S}(\vartheta^\circ; q^{-1})\, y_t + \mathcal{R}(\vartheta^\circ; q^{-1})\, u_t, \tag{2.2}$$

where $\mathcal{S}(\vartheta^\circ; q^{-1}) = \sum_{i=0}^{n-1} s_i(\vartheta^\circ) q^{-i}$ and $\mathcal{R}(\vartheta^\circ; q^{-1}) = \sum_{i=1}^{m-1} r_i(\vartheta^\circ) q^{-i}$, and coefficients $\{s_i(\vartheta^\circ)\}$ and $\{r_i(\vartheta^\circ)\}$ are computed as follows.

Set $L(\vartheta^\circ) := [\, s_0(\vartheta^\circ)\ s_1(\vartheta^\circ) \dots s_{n-1}(\vartheta^\circ)\ r_1(\vartheta^\circ) \dots r_{m-1}(\vartheta^\circ)\,]$. Then

$$L(\vartheta^\circ) = -(B(\vartheta^\circ)^T P(\vartheta^\circ) B(\vartheta^\circ) + \beta)^{-1} B(\vartheta^\circ)^T P(\vartheta^\circ) A(\vartheta^\circ), \tag{2.3}$$

where $P(\vartheta^\circ)$ is the unique positive semidefinite solution to the discrete time algebraic Riccati equation

$$P = A(\vartheta^\circ)^T \left[P - PB(\vartheta^\circ)(B(\vartheta^\circ)^T P B(\vartheta^\circ) + \beta)^{-1} B(\vartheta^\circ)^T P \right] A(\vartheta^\circ) + H^T H.$$

Moreover, the optimal LQG cost is given by $J^\star(\vartheta^\circ) = \sigma^2 trace(P(\vartheta^\circ) C C^T)$, a.s.

Remark 2.3. *Since the positive semidefinite solution $P(\vartheta)$ to*

$$P = A(\vartheta)^T \left[P - PB(\vartheta)(B(\vartheta)^T P B(\vartheta) + \beta)^{-1} B(\vartheta)^T P \right] A(\vartheta) + H^T H \tag{2.4}$$

is analytic as a function of the parameter vector ϑ in the set $\mathcal{C} = \{\vartheta \in \Re^{n+m} : q^s \mathcal{A}(\vartheta; q^{-1})$ and $q^{s-1} \mathcal{B}(\vartheta; q^{-1})$ have no unstable pole-zero cancellations$\}$ (see [12]*), it is easily seen that $s_i(\vartheta)$, $r_i(\vartheta)$, and $J^\star(\vartheta)$ are analytic functions of ϑ, $\vartheta \in \mathcal{C}$, as well.*

2.2. The cost-biased identification algorithm.

Introducing the observation vector $\varphi_t := [\, y_t \dots y_{t-(n-1)}\ u_t \dots u_{t-(m-1)}]^T$, system (1.1) can be given the regression-like form

$$y_t = \varphi_{t-1}^T \vartheta^\circ + n_t, \tag{2.5}$$

and the LS identification index for the estimate of ϑ° is [26]

$$V_t(\vartheta) = \sum_{s=1}^{t} (y_s - \varphi_{s-1}^T \vartheta)^2. \tag{2.6}$$

In the theorem below, we recall a fundamental result for the LS estimate proven in [23, Theorem 1].

Theorem 2.4. *Suppose that u_t is $\mathcal{F}_t$-measurable. Then*

$$(\vartheta^\circ - \hat{\vartheta}_t^{LS})^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\vartheta^\circ - \hat{\vartheta}_t^{LS}) = O\left(\log \lambda_{max} \left(\sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T \right) \right) \quad a.s. \tag{2.7}$$

In particular, this implies that under the conditions

(i) $\lambda_{min} \left(\sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T \right) \to \infty$ *a.s.,*

(ii) $\log \lambda_{max} \left(\sum_{s=0}^{t} \varphi_{s-1} \varphi_{s-1}^T \right) = o\left(\lambda_{min} \left(\sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T \right) \right)$ *a.s.,* *the LS estimate is consistent.*

In adaptive control, identification is performed in closed-loop. Therefore, one cannot ensure the satisfaction of conditions (i) and (ii) and the true parameter vector is generally not consistently estimated. Nevertheless, property (2.7) still provides a valuable bound on the discrepancy between the estimated parameter and the true parameter. We call this property "closed-loop identification property" to emphasize that it holds even in closed-loop. On the other hand, as discussed in section 1, the LS identification algorithm generally provides estimates with an optimal LQG cost larger than the optimal cost associated with the true system. This is the reason why optimality of an LS-based adaptive control scheme is not guaranteed.

Motivated by these considerations, we introduce a cost-biased identification algorithm with the twofold objective of preserving the LS property (2.7) and forcing the estimates to lie asymptotically in the parameter region with an optimal cost not larger than the optimal cost associated with the true system.

Consider the estimate $\hat{\vartheta}_t$ computed through the following algorithm:

$$\hat{\vartheta}_t = \begin{cases} \arg\min_{\vartheta \in \Theta} D_t(\vartheta) & \text{if } t = t_i,\ i = 0, 1, 2, \dots, \\ \hat{\vartheta}_{t-1} & \text{otherwise,} \end{cases} \tag{2.8}$$

where the time instants $\{t_i\}$ are obtained by the recursive equation $t_{i+1} = t_i + T_i$ initialized with $t_0 = 0$ and the cost-biased identification index $D_t(\vartheta)$ is given by

$$D_t(\vartheta) = V_t(\vartheta) + \alpha_t J^\star(\vartheta) + \gamma_t \|\vartheta - \hat{\vartheta}_{t-1}\|,\ \hat{\vartheta}_{-1} = 0, \tag{2.9}$$

where $V_t(\vartheta)$ is the LS cost (2.6) and $J^\star(\vartheta)$ is the optimal LQG cost for system (1.1) with parameter ϑ. The identification algorithm is completely defined by specifying the sequences of

- freezing time intervals $\{T_i\}$,
- cost-biasing weights $\{\alpha_t\}$,
- friction parameters $\{\gamma_t\}$.

We discuss hereafter the meaning of these parameters, while their actual choice is postponed to the following section.

The freezing parameter T_i is used to ensure stability of the closed-loop system. Since the parameter estimate changes with time and the control law is tuned to such an estimate, the adaptive control system is time-varying. On the other hand, it is well known that guaranteeing a stability property at each time instant for the "frozen dynamics" does not imply that the overall time-varying system has a stable dynamics. This problem can be solved by updating the estimate at a slower rate than the updating of the system variables, and this is achieved by a suitable choice of T_i. This same approach is exploited, for instance, in [17], [27], and [29].

The cost-biasing term $\alpha_t J^\star(\vartheta)$ is introduced with the objective of penalizing those parameterizations with high optimal LQG cost. The weight α_t has to be appropriately selected so as to balance the contrasting objectives of preserving the closed-loop identification property (2.7) and forcing the asymptotic estimate to correspond to a model with value of the optimal LQG performance index not larger than the optimal performance value for the true system.

Finally, the friction term $\gamma_t \|\vartheta - \hat{\vartheta}_{t-1}\|$ is introduced so as to avoid the estimate $\hat{\vartheta}_t$ being subject to undesired jumps in the time instants t_i when it is updated. This is necessary to prove optimality of the adaptive scheme.

3. Selection of $\{\mathbf{T_i}\}$, $\{\alpha_\mathbf{t}\}$, $\{\gamma_\mathbf{t}\}$ and properties of $\hat{\vartheta}_\mathbf{t}$. The adaptive control law is given by the optimal control law (2.2) with the estimate $\hat{\vartheta}_t$ in place of ϑ° (certainty equivalence principle):

$$u_t = \mathcal{S}(\hat{\vartheta}_t; q^{-1})\, y_t + \mathcal{R}(\hat{\vartheta}_t; q^{-1})\, u_t.$$

The system

$$\begin{cases} y_{t+1} = [1 - \mathcal{A}(\hat{\vartheta}_t; q^{-1})]\, y_{t+1} + \mathcal{B}(\hat{\vartheta}_t; q^{-1})\, u_t, \\ u_t = \mathcal{S}(\hat{\vartheta}_t; q^{-1})\, y_t + \mathcal{R}(\hat{\vartheta}_t; q^{-1})\, u_t \end{cases} \tag{3.1}$$

is then given the name of *time-varying estimated system.* We will select T_i so as to stabilize system (3.1) and later on in section 4 we shall see that this leads to the stability of the true closed-loop system. Letting $x_t := [y_t \dots y_{t-(n-1)} u_{t-1} \dots u_{t-(m-1)}]^T$, this system can be given the state space representation

$$x_{t+1} = F(\hat{\vartheta}_t)\, x_t \tag{3.2}$$

with

$$F(\vartheta) = A(\vartheta) + B(\vartheta)L(\vartheta), \tag{3.3}$$

where matrices $A(\vartheta)$, $B(\vartheta)$, and $L(\vartheta)$ have been introduced in section 2.1.

Choose now a constant $\mu < 1$ (*contraction* constant). The time interval T_i is then defined as

$$T_i := \inf\{\tau \in Z_+ : \|F(\hat{\vartheta}_{t_i})^\tau\| \le \mu\} \tag{3.4}$$

(note that such a T_i exists since $\hat{\vartheta}_{t_i}$ belongs to Θ and therefore corresponds to a stabilizable system). In this way, the time-varying system (3.1) is kept constant until its dynamics is contracted by a factor μ, whence guaranteeing its stability. The following proposition makes this precise.

PROPOSITION 3.1. *The autonomous estimated system $x_{t+1} = F(\hat{\vartheta}_t)\, x_t$ is a.s. exponentially stable, uniformly in time.*

Proof. The proof is given in the appendix. □

The choice of $\{\alpha_t\}$ and $\{\gamma_t\}$ is discussed in the next theorem.

THEOREM 3.2. *Suppose that u_t is $\mathcal{F}_t$-measurable. Given $\delta > 0$, select*

$$\alpha_t := \log^{1+\delta} \lambda_{max} \left(\sum_{s=1}^{t} \varphi_{s-1}\varphi_{s-1}^T \right) \tag{3.5}$$

and $\{\gamma_t\}$ to be a positive diverging sequence of real numbers satisfying $\gamma_t = o(\alpha_t)$. Then,

(i) $(\vartheta^\circ - \hat{\vartheta}_{t_i})^T \sum_{s=1}^{t_i} \varphi_{s-1}\varphi_{s-1}^T (\vartheta^\circ - \hat{\vartheta}_{t_i}) = O\left(\log^{1+\delta} \lambda_{max} \left(\sum_{s=1}^{t_i} \varphi_{s-1}\varphi_{s-1}^T \right) \right)$ *a.s.,*

(ii) $\limsup_{t\to\infty} J^\star(\hat{\vartheta}_t) \le J^\star(\vartheta^\circ)$ *a.s.,*

(iii) *if* $\sum_{t=1}^{N} \|\varphi_{t-1}\|^2 = O(N)$ *a.s., then* $\sum_{t=1}^{N} \|\hat{\vartheta}_t - \hat{\vartheta}_{t-1}\| = o(N)$ *a.s.*

Proof. The proof is given in the appendix. □

According to (3.5), $\{\alpha_t\}$ is chosen to be an increasing sequence of real numbers adaptively selected on the basis of the data generated by the controlled system. According to result (ii), this selection is effective in pushing the estimate towards the region where the optimal LQG cost is not larger than $J^\star(\vartheta^\circ)$. In turn, result (i) shows that the closed-loop identification property (2.7) is preserved with two slight differences: (1) the exponent $1+\delta$ appears in the right-hand side, (2) the rate of divergence in point (i) of Theorem 3.2 concerns only the time instants t_i when the estimate $\hat{\vartheta}_t$ is updated, while the original closed-loop identification property refers to all t's.

Before ending this section, we state a proposition regarding the estimation error

$$e_t := \varphi_t^T[\vartheta^\circ - \hat{\vartheta}_t]. \tag{3.6}$$

The technical proof of this proposition is given in the appendix and is obtained by a suitable manipulation of the sole result (i) in Theorem 3.2.

Proposition 3.3. *The estimation error $e_t = \varphi_t^T[\vartheta^\circ - \hat{\vartheta}_t]$ satisfies the following equation:*

$$\sum_{t=0,\ t\notin\mathcal{B}_N}^{N} |e_t|^p = o\left(\sum_{t=0}^{N} \|\varphi_t\|^p + N\right), \quad p \geq 2,\ a.s.,$$

where $\mathcal{B}_N$ is a set of instant points which depends on N, whose cardinality is bounded: $|\mathcal{B}_N| \leq C_{\mathcal{B}}\ \forall N$.

4. Stability and optimality.

The closed-loop system

$$\left\{\begin{array}{l} y_{t+1} = [1-\mathcal{A}(\vartheta^\circ;q^{-1})]\,y_{t+1} + \mathcal{B}(\vartheta^\circ;q^{-1})\,u_t + n_{t+1}, \\ u_t = \mathcal{S}(\hat{\vartheta}_t;q^{-1})\,y_t + \mathcal{R}(\hat{\vartheta}_t;q^{-1})\,u_t \end{array}\right. \tag{4.1}$$

can be represented as a variation system with respect to the so-called estimated system of (3.1) as follows:

$$\left\{\begin{array}{l} y_{t+1} = [1-\mathcal{A}(\hat{\vartheta}_t;q^{-1})]\,y_{t+1} + \mathcal{B}(\hat{\vartheta}_t;q^{-1})\,u_t + n_{t+1} + e_t, \\ u_t = \mathcal{S}(\hat{\vartheta}_t;q^{-1})\,y_t + \mathcal{R}(\hat{\vartheta}_t;q^{-1})\,u_t, \end{array}\right. \tag{4.2}$$

where e_t is defined in (3.6). The uniform stability property of the estimated system (3.1) (Proposition 3.1) and the property of e_t stated in Proposition 3.3 are exploited in the next theorem to prove stability of system (4.1).

Theorem 4.1 (L^p-stability). *The adaptive LQG control scheme*

$$\left\{\begin{array}{l} y_{t+1} = [1-\mathcal{A}(\vartheta^\circ;q^{-1})]\,y_{t+1} + \mathcal{B}(\vartheta^\circ;q^{-1})\,u_t + n_{t+1}, \\ u_t = \mathcal{S}(\hat{\vartheta}_t;q^{-1})\,y_t + \mathcal{R}(\hat{\vartheta}_t;q^{-1})\,u_t \end{array}\right.$$

is L^p-stable: $\limsup_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1}[|y_t|^p + |u_t|^p] < \infty$ *a.s.* $\forall p > 0$.

Proof. Fix a time point $N > 0$ and an integer $d \geq 1$.

From Proposition 3.3, there exists a set of instant points $\mathcal{B}_{N-1}$ such that

$$\sum_{t=0,\ t\notin\mathcal{B}_{N-1}}^{N-1} e_t^{2^d} = o\left(\sum_{t=0}^{N-1} \|\varphi_t\|^{2^d} + N\right) \quad \text{a.s.} \tag{4.3}$$

In view of representation (4.2) of system (4.1), it is easily seen that the state vector

$$x_t = [y_t \dots y_{t-(n-1)}\, u_{t-1} \dots u_{t-(m-1)}]^T$$

is governed by the equation

$$\begin{aligned} x_{t+1} &= F^\circ(\hat{\vartheta}_t)\, x_t + C n_{t+1} && (4.4)\\ &= F(\hat{\vartheta}_t)\, x_t + C[e_t + n_{t+1}], && (4.5)\end{aligned}$$

where $F^\circ(\vartheta) = A(\vartheta^\circ) + B(\vartheta^\circ)L(\vartheta)$, $A(\vartheta^\circ)$, $B(\vartheta^\circ)$, $L(\vartheta)$, and C are defined in section 2.1, and $F(\vartheta)$ is given in (3.3).

For the following derivations, it is convenient to use representation (4.4) in the time instants $t \in \mathcal{B}_{N-1}$ and representation (4.5) for $t \notin \mathcal{B}_{N-1}$, thus finally leading to

$$x_{t+1} = \begin{cases} F^\circ(\hat{\vartheta}_t)\, x_t + C n_{t+1}, & t \in \mathcal{B}_{N-1}, \\ F(\hat{\vartheta}_t)\, x_t + C[e_t + n_{t+1}], & t \notin \mathcal{B}_{N-1}. \end{cases} \tag{4.6}$$

Note now that since $\hat{\vartheta}_t$ belongs to the compact set Θ and $F^\circ(\vartheta)$ is a continuous function of ϑ, $\vartheta \in \Theta$, we then have that $\|F^\circ(\hat{\vartheta}_t)\|$ is uniformly bounded. From this fact and the uniform exponential stability of the autonomous system $x_{t+1} = F(\hat{\vartheta}_t)x_t$ (Proposition 3.1), and the fact that $|\mathcal{B}_{N-1}| \leq C_{\mathcal{B}}$ $\forall N$ (see Proposition 3.3), it is easy to show that the state vector x_t generated by system (4.6) can be bounded as follows:

$$\|x_t\| \leq k_1 \left\{ \sum_{i=1}^{t} \nu^{t-i}|n_i| + \sum_{i=0, i\notin\mathcal{B}_{N-1}}^{t-1} \nu^{t-i}|e_i| \right\}, \quad t \leq N,$$

where k_1 and $\nu \in (0,1)$ are suitable constants. We now have

$$\begin{aligned} &\left[k_1 \left\{ \sum_{i=1}^{t} \nu^{t-i}|n_i| + \sum_{i=0, i\notin\mathcal{B}_{N-1}}^{t-1} \nu^{t-i}|e_i| \right\} \right]^{2^d} \\ &\quad \leq k_1^{2^d} \left[\left\{ \sum_{i=1}^{t} \nu^{t-i}|n_i| + \sum_{i=0, i\notin\mathcal{B}_{N-1}}^{t-1} \nu^{t-i}|e_i| \right\}^2 \right]^{2^{d-1}} \\ &\quad \leq k_1^{2^d} \left[2 \left\{ \sum_{i=1}^{t} \nu^{\frac{t-i}{2}} (\nu^{\frac{t-i}{2}}|n_i|) \right\}^2 + 2 \left\{ \sum_{i=0, i\notin\mathcal{B}_{N-1}}^{t-1} \nu^{\frac{t-i}{2}} (\nu^{\frac{t-i}{2}}|e_i|) \right\}^2 \right]^{2^{d-1}} \\ &\quad \leq k_1^{2^d} \left[2 \sum_{i=1}^{t} \nu^{t-i} \sum_{i=1}^{t} \nu^{t-i} n_i^2 + 2 \sum_{i=0, i\notin\mathcal{B}_{N-1}}^{t-1} \nu^{t-i} \sum_{i=0, i\notin\mathcal{B}_{N-1}}^{t-1} \nu^{t-i} e_i^2 \right]^{2^{d-1}} \\ &\quad \leq k_1^{2^d} \left(\frac{2}{1-\nu} \right)^{2^{d-1}} \left[\sum_{i=1}^{t} \nu^{t-i} n_i^2 + \sum_{i=0, i\notin\mathcal{B}_{N-1}}^{t-1} \nu^{t-i} e_i^2 \right]^{2^{d-1}}. \end{aligned}$$

Iterating this same equation d times, we then obtain

$$\|x_t\|^{2^d} \leq k_2 \left\{ \sum_{i=1}^{t} \nu^{t-i} n_i^{2^d} + \sum_{i=0, i\notin\mathcal{B}_{N-1}}^{t-1} \nu^{t-i} e_i^{2^d} \right\}, \quad t \leq N, \tag{4.7}$$

k_2 being a suitable constant, from which we finally get

$$\frac{1}{N}\sum_{t=1}^{N}\|x_t\|^{2^d} \le k_3 \left\{ \frac{1}{N}\sum_{t=1}^{N} n_t^{2^d} + \frac{1}{N}\sum_{t=0,t\notin\mathcal{B}_{N-1}}^{N-1} e_t^{2^d} \right\}, \tag{4.8}$$

where k_3 is a suitable constant, independent of N.

We next bound the two terms in the right-hand side of (4.8).

The term $\frac{1}{N}\sum_{t=1}^{N} n_t^{2^d}$ is handled as follows. Define $v_t := n_t^{2^d} - E[n_t^{2^d}|\mathcal{F}_{t-1}]$. Then $\{v_t\}$ is a martingale difference satisfying

$$\sum_{t=1}^{\infty}\frac{1}{t^2}E[v_t^2|\mathcal{F}_{t-1}] \le \sum_{t=1}^{\infty}\frac{1}{t^2}E[n_t^{2^{d+1}}|\mathcal{F}_{t-1}] < \infty,$$

due to Assumption 2.1. By applying Theorem 2.18 in [16], we then conclude that $\frac{1}{N}\sum_{t=0}^{N-1}[n_t^{2^d} - E[n_t^{2^d}|\mathcal{F}_{t-1}]]$ tends to zero, a.s. Since $\frac{1}{N}\sum_{t=0}^{N-1} E[n_t^{2^d}|\mathcal{F}_{t-1}]$ is bounded by Assumption 2.1, we finally have $\limsup_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1} n_t^{2^d} < \infty$, a.s.

The term $\frac{1}{N}\sum_{t=0,t\notin\mathcal{B}_{N-1}}^{N-1} e_t^{2^d}$ is immediately bounded by means of (4.3) and the final bound for $\frac{1}{N}\sum_{t=1}^{N}\|x_t\|^{2^d}$ is obtained

$$\frac{1}{N}\sum_{t=1}^{N}\|x_t\|^{2^d} = O(1) + o\left(\frac{1}{N}\sum_{t=0}^{N-1}\|\varphi_t\|^{2^d}\right) \quad \text{a.s.}$$

Since $\frac{1}{N}\sum_{t=0}^{N-1}\|\varphi_t\|^{2^d} \le \frac{1}{N}\sum_{t=0}^{N}\|x_t\|^{2^d}$, this implies that $\frac{1}{N}\sum_{t=0}^{N-1}\|\varphi_t\|^{2^d}$ remains bounded. Then, the thesis immediately follows from the arbitrariness of d and the fact that $\frac{1}{N}\sum_{t=0}^{N-1}\|\varphi_t\|^{p} \le \frac{1}{N}\sum_{t=0}^{N-1}[\|\varphi_t\|^{2^d}+1]$, $p \le 2^d$. □

In the next theorem we show that the LQG adaptive control scheme is self-optimizing.

THEOREM 4.2 (optimality). *The adaptive LQG control scheme*

$$\begin{cases} y_{t+1} = [1-\mathcal{A}(\vartheta^\circ;q^{-1})]\,y_{t+1} + \mathcal{B}(\vartheta^\circ;q^{-1})\,u_t + n_{t+1}, \\ u_t = \mathcal{S}(\hat{\vartheta}_t;q^{-1})\,y_t + \mathcal{R}(\hat{\vartheta}_t;q^{-1})\,u_t \end{cases}$$

is self-optimizing: $\limsup_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1}\left[y_t^2 + \beta\,u_t^2\right] = J^\star(\vartheta^\circ)$ a.s.

Proof. We start by showing that $x_t := [y_t \dots y_{t-(n-1)}\ u_{t-1} \dots u_{t-(m-1)}]^T$ satisfies the following equation:

$$\|x_t\|^p = o(t), \quad \forall p > 0, \quad \text{a.s.} \tag{4.9}$$

This condition will be useful in the subsequent derivations. By contradiction, suppose that there exist $\{t_k\}_{k\ge 0}$ and a real number $\eta > 0$, such that $\|x_{t_k}\| > \eta t_k\ \forall k$. Then

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\|x_t\|^{1+p} \ge \limsup_{k\to\infty}\frac{1}{t_k}\|x_{t_k}\|^{1+p} \ge \limsup_{k\to\infty}\frac{1}{t_k}\eta^{1+p}t_k^{1+p} = \infty,$$

which contradicts Theorem 4.1.

Observe now that the dynamic programming equation for the estimated model $x_{t+1} = A(\hat{\vartheta}_t)x_t + B(\hat{\vartheta}_t)u_t + Cn_{t+1}$ writes

$$\begin{aligned} J^\star(\hat{\vartheta}_t) + x_t^T P(\hat{\vartheta}_t)x_t = x_t^T T x_t + \beta u_t^2 + E\big[&(A(\hat{\vartheta}_t)x_t + B(\hat{\vartheta}_t)u_t + Cn_{t+1})^T \\ &P(\hat{\vartheta}_t)(A(\hat{\vartheta}_t)x_t + B(\hat{\vartheta}_t)u_t + Cn_{t+1})\,|\,\mathcal{F}_t\big], \end{aligned} \tag{4.10}$$

where $P(\vartheta)$ is the solution to the Riccati equation (2.4). By (2.1) and the definition (3.6), x_t can be given the following expression:

$$x_{t+1} = A(\hat{\vartheta}_t)x_t + B(\hat{\vartheta}_t)u_t + Cn_{t+1} + Ce_t. \tag{4.11}$$

Substituting (4.11) in (4.10), we then get

$$J^\star(\hat{\vartheta}_t) + x_t^T P(\hat{\vartheta}_t)x_t = x_t^T T x_t + \beta u_t^2 + E\big[(x_{t+1} - Ce_t)^T P(\hat{\vartheta}_t)(x_{t+1} - Ce_t) \,|\, \mathcal{F}_t\big],$$

from which

$$\begin{aligned}\frac{1}{N}\sum_{t=0}^{N-1} J^\star(\hat{\vartheta}_t) - \frac{1}{N}\sum_{t=0}^{N-1}[x_t^T T x_t + \beta u_t^2]&\\ = -\frac{1}{N}\sum_{t=0}^{N-1}&\Big[x_t^T P(\hat{\vartheta}_t)x_t - E\big[x_{t+1}^T P(\hat{\vartheta}_{t+1})x_{t+1} \,|\, \mathcal{F}_t\big]\Big]\\ +\frac{1}{N}\sum_{t=0}^{N-1}& E\big[x_{t+1}^T(P(\hat{\vartheta}_t) - P(\hat{\vartheta}_{t+1}))x_{t+1} \,|\, \mathcal{F}_t\big]\\ +\frac{1}{N}\sum_{t=0}^{N-1}& E\big[e_t^T C^T P(\hat{\vartheta}_t)Ce_t \,|\, \mathcal{F}_t\big]\\ -2\frac{1}{N}\sum_{t=0}^{N-1}& E\big[x_{t+1}^T P(\hat{\vartheta}_t)Ce_t \,|\, \mathcal{F}_t\big].\end{aligned} \tag{4.12}$$

From property (ii) in Theorem 3.2, we get $\limsup_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} J^\star(\hat{\vartheta}_t) \le J^\star(\vartheta^\circ)$ a.s. Therefore, the thesis will be proved if we show that all the terms in the right-hand side of (4.12) tend to zeros as $N \to \infty$. We shall study each term separately.

First term:

$$\begin{aligned}\frac{1}{N}\sum_{t=0}^{N-1}\Big[x_t^T P(\hat{\vartheta}_t)x_t - E\big[x_{t+1}^T P(\hat{\vartheta}_{t+1})x_{t+1} \,|\, \mathcal{F}_t\big]\Big] &= -\frac{1}{N}x_N^T P(\hat{\vartheta}_N)x_N\\ +\frac{1}{N}x_0^T P(\hat{\vartheta}_0)x_0 + \frac{1}{N}\sum_{t=0}^{N-1}&\Big[x_{t+1}^T P(\hat{\vartheta}_{t+1})x_{t+1} - E\big[x_{t+1}^T P(\hat{\vartheta}_{t+1})x_{t+1} \,|\, \mathcal{F}_t\big]\Big].\end{aligned}$$

The term $\frac{1}{N}x_0^T P(\hat{\vartheta}_0)x_0$ equals zero. As for $\frac{1}{N}x_N^T P(\hat{\vartheta}_N)x_N$, observe that

$$\frac{1}{N}x_N^T P(\hat{\vartheta}_N)x_N \le k_1\frac{1}{N}\|x_N\|^2,$$

k_1 being a suitable constant, since $P(\vartheta)$ is uniformly bounded on the compact set Θ (see Remark 2.3). Therefore, from (4.9) we get

$$\lim_{N\to\infty}\frac{1}{N}x_N^T P(\hat{\vartheta}_N)x_N = 0.$$

Define $w_t := x_{t+1}^T P(\hat{\vartheta}_{t+1})x_{t+1} - E\big[x_{t+1}^T P(\hat{\vartheta}_{t+1})x_{t+1} \,|\, \mathcal{F}_t\big]$. Then $\{w_t\}$ is a martingale difference. Hence, $\frac{1}{N}\sum_{t=0}^{N-1} w_t$ asymptotically vanishes if $\sum_{t=0}^{\infty}\frac{1}{t^2}E[w_{t+1}^2|\mathcal{F}_t] < \infty$ (see Theorem 2.18 in [16]). We have

$$E[w_{t+1}^2|\mathcal{F}_t] \le E\big[(x_{t+1}^T P(\hat{\vartheta}_{t+1})x_{t+1})^2|\, \mathcal{F}_t\big] \le k_2 E\big[\|x_{t+1}\|^4|\, \mathcal{F}_t\big] \le k_3\Big[\|x_t\|^4 + \|u_t\|^4 + 1\Big],$$

k_2, k_3 being suitable constants, since $P(\vartheta)$ is uniformly bounded over Θ and $\{n_t\}$ satisfies point 1 in Assumption 2.1. We then need to prove that $\sum_{t=0}^{\infty}\frac{1}{t^2}[\|x_t\|^4 + \|u_t\|^4] < \infty$. This is easily shown through (4.9) with $p = 8$, which implies $\|x_t\|^4 = o(t^{1/2})$ and $u_t^4 = o(t^{1/2})$, since $\sum_{t=0}^{\infty}\frac{1}{t^2}[\|x_t\|^4 + \|u_t\|^4] = \sum_{t=0}^{\infty}\frac{1}{t^{3/2}}\frac{1}{t^{1/2}}[\|x_t\|^4 + \|u_t\|^4]$, where $\sum_{t=0}^{\infty}\frac{1}{t^{3/2}}$ converges.

Second term:

Observe that $\{v_t\} := \{x_{t+1}^T(P(\hat{\vartheta}_t)-P(\hat{\vartheta}_{t+1}))x_{t+1}-E\big[x_{t+1}^T(P(\hat{\vartheta}_t)-P(\hat{\vartheta}_{t+1}))x_{t+1}|\,\mathcal{F}_t\big]\}$ is a martingale difference. By derivations similar to those for the first term, we can prove that $\frac{1}{N}\sum_{t=0}^{N} v_t \to 0$. Then

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1} E\big[x_{t+1}^T(P(\hat{\vartheta}_t) - P(\hat{\vartheta}_{t+1}))x_{t+1}\,|\,\mathcal{F}_t\big] = 0$$

is proven by showing that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1} x_{t+1}^T(P(\hat{\vartheta}_t) - P(\hat{\vartheta}_{t+1}))x_{t+1} = 0. \tag{4.13}$$

To prove (4.13), apply the Schwarz inequality to obtain

$$\begin{aligned}\left|\frac{1}{N}\sum_{t=0}^{N-1} x_{t+1}^T(P(\hat{\vartheta}_t) - P(\hat{\vartheta}_{t+1}))x_{t+1}\right| &\le \frac{1}{N}\sum_{t=0}^{N-1}\|P(\hat{\vartheta}_t) - P(\hat{\vartheta}_{t+1})\|\,\|x_{t+1}\|^2\\ &\le \left(\frac{1}{N}\sum_{t=0}^{N-1}\|P(\hat{\vartheta}_t) - P(\hat{\vartheta}_{t+1})\|^2\right)^{\frac{1}{2}}\left(\frac{1}{N}\sum_{t=0}^{N-1}\|x_{t+1}\|^4\right)^{\frac{1}{2}}.\end{aligned}$$

By Theorem 4.1 $\frac{1}{N}\sum_{t=0}^{N-1}\|x_{t+1}\|^4$ is bounded. Moreover, $\lim_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1}\|P(\hat{\vartheta}_t) - P(\hat{\vartheta}_{t+1})\|^2 = 0$ because of property (iii) in Theorem 3.2 and the Lipschitz continuity of $P(\vartheta)$ over Θ ($P(\vartheta)$ is analytic on Θ and Θ is compact). This concludes the proof of (4.13).

Third term:

Since $\hat{\vartheta}_t \in \Theta$ and $P(\vartheta)$ is uniformly bounded on Θ, then

$$0 \le \frac{1}{N}\sum_{t=0}^{N-1} E\big[e_t^T C^T P(\hat{\vartheta}_t)Ce_t\,|\,\mathcal{F}_t\big] = \frac{1}{N}\sum_{t=0}^{N-1} e_t^T C^T P(\hat{\vartheta}_t)Ce_t \le h_1\frac{1}{N}\sum_{t=0}^{N-1} e_t^2,$$

h_1 being a suitable constant. We now show that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1} e_t^2 = 0 \quad \text{a.s.} \tag{4.14}$$

From Proposition 3.3 it follows that there exists a set of instant points $\mathcal{B}_{N-1}$ whose cardinality is upper bounded by a constant $C_{\mathcal{B}} < \infty$ such that $\frac{1}{N}\sum_{t=0,\ t\notin\mathcal{B}_{N-1}}^{N-1} e_t^2 = \frac{1}{N}o(\sum_{t=0}^{N-1}\|\varphi_t\|^2)$ a.s. Then, recalling the definition (3.6) of e_t, we have

$$\frac{1}{N}\sum_{t=0}^{N-1} e_t^2 = \frac{1}{N}o\left(\sum_{t=0}^{N-1}\|\varphi_t\|^2\right) + \frac{1}{N}\sum_{t\in\mathcal{B}_{N-1}}|\varphi_t^T(\vartheta^\circ - \hat{\vartheta}_t)|^2.$$

By Theorem 4.1, the first term tends to zero. As for the second term, we have that it can be upper bounded as follows:

$$\frac{1}{N}\sum_{t\in\mathcal{B}_{N-1}}|\varphi_t^T(\vartheta^\circ-\hat{\vartheta}_t)|^2\le h_2\,C_{\mathcal{B}}\frac{1}{N}\max_{0\le t\le N-1}\|\varphi_t\|^2,\quad h_2=\text{suitable constant,}$$

since $\hat{\vartheta}_t$ is bounded uniformly in time. Noting that $\|\varphi_t\|^2\le\|x_{t+1}\|^2+\|x_t\|^2$, from (4.9) we get

$$\|\varphi_t\|^2=o(t)\quad\text{a.s.,}\tag{4.15}$$

which implies $\frac{1}{N}\max_{0\le t\le N-1}\|\varphi_t\|^2\to 0$.

Fourth term:

$$\begin{aligned}\frac{1}{N}\sum_{t=0}^{N-1}&E\big[x_{t+1}^TP(\hat{\vartheta}_t)Ce_t\,|\,\mathcal{F}_t\big]\\&=\frac{1}{N}\sum_{t=0}^{N-1}E\big[(A(\vartheta^\circ)x_t+B(\vartheta^\circ)u_t+Cn_{t+1})^TP(\hat{\vartheta}_t)Ce_t\,|\,\mathcal{F}_t\big]\\&=\frac{1}{N}\sum_{t=0}^{N-1}x_t^TA(\vartheta^\circ)^TP(\hat{\vartheta}_t)Ce_t+\frac{1}{N}\sum_{t=0}^{N-1}u_t^TB(\vartheta^\circ)^TP(\hat{\vartheta}_t)Ce_t.\end{aligned}$$

We next show that each term on the right-hand side goes to zero as N tends to infinity.

Since $\hat{\vartheta}_t\in\Theta$, with Θ compact, and $P(\vartheta)$ is analytic on Θ, by using Schwarz inequality, we have

$$\left|\frac{1}{N}\sum_{t=0}^{N-1}x_t^TA(\vartheta^\circ)^TP(\hat{\vartheta}_t)Ce_t\right|\le k\left(\frac{1}{N}\sum_{t=0}^{N-1}\|x_t\|^2\right)^{\frac{1}{2}}\left(\frac{1}{N}\sum_{t=0}^{N-1}e_t^2\right)^{\frac{1}{2}}.$$

Then $\lim_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1}x_t^TA(\vartheta^\circ)^TP(\hat{\vartheta}_t)Ce_t=0$ a.s. follows from Theorem 4.1 and (4.14).

Similarly, one can prove that $\lim_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1}u_t^TB(\vartheta^\circ)^TP(\hat{\vartheta}_t)Ce_t=0$ a.s. □

5. Conclusions. The more commonly adopted strategy for the design of adaptive control laws is the certainty equivalence approach. Although the approach is conceptually simple, working out stability and optimality results for certainty equivalence adaptive control schemes is a difficult task even in the ideal case when the true system belongs to the model class. This is due to the intricate interaction between control and identification in closed-loop, which can cause identifiability problems.

We introduced a new LQG adaptive control scheme based on the certainty equivalence principle able to ensure both stability and optimality results irrespectively of the excitation characteristics of the involved signals by adopting a cost-biased approach.

This paper presents the following limitations:

- *the true system is described as an ARX system subject to white noise.* This hypothesis is necessary mainly for the applicability of the proposed cost-biased LS identification method, whose properties are in fact derived on the basis of the LS estimate properties. As a consequence of this fact, the extension to the ARMAX system case is not straightforward.

- *the proposed identification method is nonrecursive.* The cost-biased identification index has, in general, multiple local minima and its minimization is not straightforward. Therefore, it should be minimized by resorting to some global optimization algorithm (see, e.g., [30, 31, 4, 15]). This limitation must be removed by introducing some recursive way to minimize our performance index so as to retain all the properties relevant to control.

These problems constitute interesting research issues. In particular, inspired by the result obtained for the white noise case, one can conceive of introducing appropriate cost-biased identification algorithms for the colored noise case. In this regard, much work has to be done, but an encouraging starting point is represented by the fact that the extended LS algorithm satisfies closed-loop properties similar to those valid for the LS algorithm (see, e.g., [10]).

Appendix. Proofs of the results in section 3.

Proof of Proposition 3.1. Recall that $\hat{\vartheta}_t \in \Theta$, $t \geq 0$, where Θ is compact and is such that all the parametrizations in Θ correspond to stabilizable models. We start by proving that $T(\vartheta) := \inf\{\tau \in Z_+ : \|F(\vartheta)^\tau\| \leq \mu\}$ is uniformly bounded in the compact set Θ, i.e., $\sup_{\vartheta \in \Theta} T(\vartheta) < \infty$.

Condition $\vartheta \in \Theta$ implies that the system $\mathcal{A}(\vartheta; q^{-1}) y_{t+1} = \mathcal{B}(\vartheta; q^{-1}) u_t$ associated with parameter ϑ is stabilizable and therefore stabilized by the control law $u_t = \mathcal{S}(\vartheta; q^{-1}) y_t + \mathcal{R}(\vartheta; q^{-1}) u_t$. From this it follows that the dynamic matrix $F(\vartheta)$ of the time-invariant system

$$x_{t+1} = F(\vartheta) x_t \tag{A.1}$$

is exponentially stable.

Denote by $\{\lambda_i(\vartheta)\}_{i=1,\ldots,n+m-1}$ the eigenvalues of $F(\vartheta)$.

By the observation that $F(\vartheta)$ is a continuous function of ϑ, $\mathcal{C} = \{\vartheta \in \Re^{n+m} : q^s \mathcal{A}(\vartheta; q^{-1})$ and $q^{s-1} \mathcal{B}(\vartheta; q^{-1})$ have no unstable pole-zero cancellations$\}$, we have that $\bar{\lambda}(\vartheta) := \max_{i \in \{1,\ldots,n+m-1\}} |\lambda_i(\vartheta)|$ is also a continuous function of ϑ, $\vartheta \in \mathcal{C}$. Being Θ compact and included in $\mathcal{C}$, the conclusion is finally drawn that

$$\bar{\lambda} := \max_{\vartheta \in \Theta} \bar{\lambda}(\vartheta) < 1.$$

Fix now a real number $\bar{\nu} \in (\bar{\lambda}, 1)$ and introduce the system

$$w_{t+1} = \frac{1}{\bar{\nu}} F(\vartheta)\, w_t. \tag{A.2}$$

System (A.2) is exponentially stable $\forall \vartheta \in \Theta$, since $|\frac{\lambda_i(\vartheta)}{\bar{\nu}}| \leq \frac{\bar{\lambda}}{\bar{\nu}} < 1$ $\forall i$, $\forall \vartheta \in \Theta$. Hence, the solution $S(\vartheta)$ to the Lyapunov equation associated with matrix $\frac{1}{\bar{\nu}} F(\vartheta)$

$$\frac{1}{\bar{\nu}} F(\vartheta)^T\, S(\vartheta)\, \frac{1}{\bar{\nu}} F(\vartheta) - S(\vartheta) = -I$$

is positive definite. Moreover, it is a standard fact that the state vector w_t of system (A.2) can be bounded as follows in terms of $S(\vartheta)$:

$$\|w_t\| \leq \sqrt{\frac{\lambda_{max}(S(\vartheta))}{\lambda_{min}(S(\vartheta))}}\, \|\, w_{t^\star}\|, \;\; t \geq t^\star \geq 0, \tag{A.3}$$

where $\lambda_{max}(S(\vartheta))$ and $\lambda_{min}(S(\vartheta))$ are, respectively, the maximum and minimum eigenvalues of $S(\vartheta)$. Since $S(\vartheta)$ is continuous in the closed set Θ (see [12]), we can define $c := \max_{\vartheta\in\Theta} \sqrt{\frac{\lambda_{max}(S(\vartheta))}{\lambda_{min}(S(\vartheta))}}$ and rewrite inequality (A.3) as $\|w_t\| \le c\,\|w_{t^\star}\|,\ \ t \ge\ t^\star$ $\forall\,\vartheta \in \Theta$. Setting $w_{t^\star} = x_{t^\star}$, we finally get a bound on the state vector x_t of the time-invariant system (A.1)

$$\|x_t\| \le c\,\bar\nu^{t-t^\star}\|x_{t^\star}\|, \quad t \ge t^\star, \forall\,\vartheta \in \Theta. \tag{A.4}$$

Set $\bar T = \inf\{\tau \in Z_+ : c\,\bar\nu^\tau \le \mu\} < \infty$. Since $\|x_{\bar T+t^\star}\| = \|F(\vartheta)^{\bar T} x_{t^\star}\| \le \mu\|x_{t^\star}\|$ $\forall\,\vartheta \in \Theta$, $\forall\, x_{t^\star}$, then $\|F(\vartheta)^{\bar T}\| = \sup_{\|x\|\neq 0} \frac{\|F(\vartheta)^{\bar T} x\|}{\|x\|} \le \mu\ \forall\,\vartheta \in \Theta$, and therefore $T(\vartheta) = \inf\{\tau \in Z_+ : \|F(\vartheta)^\tau\| \le \mu\}$ satisfies $T(\vartheta) \le \bar T\ \forall\,\vartheta \in \Theta$. This finally implies that

$$\sup_{\vartheta\in\Theta}\{T(\vartheta)\} \le \bar T < \infty. \tag{A.5}$$

Let us turn now to considering the time-varying system $x_{t+1} = F(\hat\vartheta_t)\,x_t$.

Being $\hat\vartheta_t \in \Theta$, $t \ge 0$, from (A.5) it follows that the updating time interval T_i in (3.4) is uniformly bounded:

$$T := \sup_{i\ge 0} T_i < \bar T. \tag{A.6}$$

We are now in a position to establish the uniform exponential stability. We apply (A.4) to the state vector x_t on each finite time interval $[t_i, t_{i+1}]$, thus getting

$$\|x_t\| \le c\,\bar\nu^{t-t^\star}\|x_{t^\star}\|, \ \ t_i \le t^\star \le t \le t_{i+1}. \tag{A.7}$$

If we choose $t^\star = t_i$, we have $\|x_t\| \le c\,\bar\nu^{t-t_i}\|x_{t_i}\|,\ \ t \in [t_i, t_{i+1}]$. From the definition (3.4) of $\{T_k\}$, it follows that $\|x_{t_i}\| \le \mu^{i-j}\|x_{t_j}\|,\ \ j \le i$. By applying (A.7) in the time interval $[t_{j-1}, t_j]$ with $t = t_j$, we get $\|x_{t_j}\| \le c\,\bar\nu^{t_j-t^\star}\|x_{t^\star}\|,\ \ t^\star \in [t_{j-1}, t_j]$. These last three inequalities lead to

$$\|x_t\| \le c\,\bar\nu^{t-t_i}\mu^{i-j}\,c\,\bar\nu^{t_j-t^\star}\|x_{t^\star}\|,\ \ t_{j-1} \le t^\star \le t_j \le t_i \le t \le t_{i+1},\ \ j \le i.$$

By setting $\nu = \max\{\bar\nu, \mu^{\frac{1}{T}}\} < 1$, we have that $\mu \le \nu^{T_k}\ \forall k$ and therefore

$$\begin{aligned}\|x_t\| &\le c^2\,\nu^{t-t_i}\nu^{t_i-t_{i-1}}\dots\nu^{t_{j+1}-t_j}\nu^{t_j-t^\star}\|x_{t^\star}\| \\ &= c^2\,\nu^{t-t^\star}\|x_{t^\star}\|,\ \ t_{j-1} \le t^\star \le t_j \le t_i \le t \le t_{i+1}.\end{aligned}$$

Finally, from this last inequality and inequality (A.7), we get $\|x_t\| \le c^2\,\nu^{t-t^\star}\|x_{t^\star}\|$, $t^\star \le t$, i.e., the thesis.

Proof of Theorem 3.2. Point (i): $D_t(\vartheta) - V_t(\hat\vartheta_t^{LS})$ can be written as follows:

$$\begin{aligned}D_t(\vartheta) - V_t(\hat\vartheta_t^{LS}) &= \sum_{s=1}^t (y_s - \varphi_{s-1}^T\vartheta)^2 + \alpha_t J^\star(\vartheta) + \gamma_t\|\vartheta - \hat\vartheta_{t-1}\| - \sum_{s=1}^t (y_s - \varphi_{s-1}^T\hat\vartheta_t^{LS})^2 \\ &= \vartheta^T\sum_{s=1}^t \varphi_{s-1}\varphi_{s-1}^T\vartheta - 2\vartheta^T\sum_{s=1}^t \varphi_{s-1}y_s + \alpha_t J^\star(\vartheta) + \gamma_t\|\vartheta - \hat\vartheta_{t-1}\| \\ &\quad -(\hat\vartheta_t^{LS})^T\sum_{s=1}^t \varphi_{s-1}\varphi_{s-1}^T\hat\vartheta_t^{LS} + 2(\hat\vartheta_t^{LS})^T\sum_{s=1}^t \varphi_{s-1}y_s.\end{aligned} \tag{A.8}$$

The LS estimate $\hat{\vartheta}_t^{LS}$ minimizing $V_t(\vartheta)$ satisfies the following equality:

$$\sum_{s=1}^{t} \varphi_{s-1} y_s = \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T \hat{\vartheta}_t^{LS}.$$

Substituting this last expression in (A.8), we obtain

$$D_t(\vartheta) - V_t(\hat{\vartheta}_t^{LS}) = (\vartheta - \hat{\vartheta}_t^{LS})^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\vartheta - \hat{\vartheta}_t^{LS}) + \alpha_t J^\star(\vartheta) + \gamma_t \|\vartheta - \hat{\vartheta}_{t-1}\|. \tag{A.9}$$

Set $\vartheta_t := \arg\min_{\vartheta \in \Theta} D_t(\vartheta)$. By definition of ϑ_t we have

$$D_t(\vartheta_t) - V_t(\hat{\vartheta}_t^{LS}) \leq D_t(\vartheta) - V_t(\hat{\vartheta}_t^{LS}), \quad \vartheta \in \Theta.$$

By choosing $\vartheta = \vartheta^\circ$ and using expression (A.9), we then get

$$\begin{aligned} &(\vartheta_t - \hat{\vartheta}_t^{LS})^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\vartheta_t - \hat{\vartheta}_t^{LS}) + \alpha_t J^\star(\vartheta_t) + \gamma_t \|\vartheta_t - \hat{\vartheta}_{t-1}\| \\ &\leq (\vartheta^\circ - \hat{\vartheta}_t^{LS})^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\vartheta^\circ - \hat{\vartheta}_t^{LS}) + \alpha_t J^\star(\vartheta^\circ) + \gamma_t \|\vartheta^\circ - \hat{\vartheta}_{t-1}\| \\ &= O(\alpha_t) \quad \text{a.s.,} \end{aligned} \tag{A.10}$$

where the last equality follows from Theorem 2.4, the definition (3.5) of α_t, the fact that $\|\vartheta^\circ - \hat{\vartheta}_{t-1}\|$ is bounded, and the relation $\gamma_t = o(\alpha_t)$. Since $\alpha_t J^\star(\vartheta_t) + \gamma_t \|\vartheta_t - \hat{\vartheta}_{t-1}\| \geq 0$, we have $(\vartheta_t - \hat{\vartheta}_t^{LS})^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\vartheta_t - \hat{\vartheta}_t^{LS}) = O(\alpha_t)$ a.s. From definition (3.5) of α_t and Theorem 2.4, we then have

$$\begin{aligned} (\vartheta_t - \vartheta^\circ)^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\vartheta_t - \vartheta^\circ) &\leq 2\Big[(\vartheta_t - \hat{\vartheta}_t^{LS})^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\vartheta_t - \hat{\vartheta}_t^{LS}) \\ &\quad + (\hat{\vartheta}_t^{LS} - \vartheta^\circ)^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\hat{\vartheta}_t^{LS} - \vartheta^\circ)\Big] \\ &= O(\alpha_t) \text{ a.s.,} \end{aligned}$$

thus concluding the proof of point (i), since $\hat{\vartheta}_t = \vartheta_t$, for $t = t_i$, $i = 0, 1, \ldots$.

Point (ii): A simple elaboration of (A.10) shows that

$$\begin{aligned} J^\star(\vartheta_t) &\leq \frac{(\vartheta^\circ - \hat{\vartheta}_t^{LS})^T \sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T (\vartheta^\circ - \hat{\vartheta}_t^{LS})}{\alpha_t} + J^\star(\vartheta^\circ) + \frac{\gamma_t}{\alpha_t} \|\vartheta^\circ - \hat{\vartheta}_{t-1}\| \\ &= \frac{O(\log \lambda_{max}(\sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T))}{\log^{1+\delta} \lambda_{max}(\sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T)} + \frac{o(\alpha_t)}{\alpha_t} + J^\star(\vartheta^\circ) \text{ a.s.,} \end{aligned}$$

where in the second equation we have used the definition of α_t given in equation (3.5) and the fact that $\gamma_t = o(\alpha_t)$. To conclude the proof, it suffices to show that $\lim_{t \to \infty} \log \lambda_{max}(\sum_{s=1}^{t} \varphi_{s-1} \varphi_{s-1}^T) = \infty$. The easy proof of this fact is omitted.

Point (iii): By the definition (2.8) of $\hat{\vartheta}_t$, we have

$$V_t(\hat{\vartheta}_t) + \alpha_t J^\star(\hat{\vartheta}_t) + \gamma_t \|\hat{\vartheta}_t - \hat{\vartheta}_{t-1}\| \leq V_t(\hat{\vartheta}_{t-1}) + \alpha_t J^\star(\hat{\vartheta}_{t-1}),$$

which implies

$$\text{(A.11)} \quad \sum_{t=1}^{N} \gamma_t \|\hat{\vartheta}_t - \hat{\vartheta}_{t-1}\| \leq \sum_{t=1}^{N} [V_t(\hat{\vartheta}_{t-1}) - V_t(\hat{\vartheta}_t)] + \sum_{t=1}^{N} \alpha_t [J^\star(\hat{\vartheta}_{t-1}) - J^\star(\hat{\vartheta}_t)].$$

The first term in the right-hand side of (A.11) can be bounded as follows:

$$\begin{aligned} \sum_{t=1}^{N} [V_t(\hat{\vartheta}_{t-1}) - V_t(\hat{\vartheta}_t)] &\leq V_1(\hat{\vartheta}_0) - V_N(\hat{\vartheta}_N) + \sum_{t=1}^{N-1} [V_{t+1}(\hat{\vartheta}_t) - V_t(\hat{\vartheta}_t)] \\ &\leq V_1(\hat{\vartheta}_0) + \sum_{t=1}^{N-1} [\varphi_t^T(\vartheta^\circ - \hat{\vartheta}_t) + n_{t+1}]^2 \\ &\leq V_1(\hat{\vartheta}_0) + 2\sum_{t=1}^{N-1} [\varphi_t^T(\vartheta^\circ - \hat{\vartheta}_t)]^2 + 2\sum_{t=1}^{N-1} n_{t+1}^2 \\ &\leq k_1 \left[1 + \sum_{t=1}^{N} \|\varphi_{t-1}\|^2 + \sum_{t=1}^{N-1} n_{t+1}^2 \right], \end{aligned}$$

k_1 being a suitable constant, where we used the boundedness of $\hat{\vartheta}_t$.

By Remark 2.3, the second term in the right-hand side of (A.11) can be bounded as follows:

$$\begin{aligned} \sum_{t=1}^{N} \alpha_t [J^\star(\hat{\vartheta}_{t-1}) - J^\star(\hat{\vartheta}_t)] &= \alpha_1 J^\star(\hat{\vartheta}_0) - \alpha_N J^\star(\hat{\vartheta}_N) + \sum_{t=1}^{N-1} (\alpha_{t+1} - \alpha_t) J^\star(\hat{\vartheta}_t) \\ &\leq \alpha_1 J^\star(\hat{\vartheta}_0) + \max_{\vartheta \in \Theta} J^\star(\vartheta) \sum_{t=1}^{N-1} (\alpha_{t+1} - \alpha_t) \\ &= k_2[1 + \alpha_N], \end{aligned}$$

where k_2 is a suitable constant.

Substituting these bounds in (A.11), we then have

$$\text{(A.12)} \quad \frac{1}{N} \sum_{t=1}^{N} \gamma_t \|\hat{\vartheta}_t - \hat{\vartheta}_{t-1}\| \leq \bar{k} \left[\frac{1}{N} + \frac{\alpha_N}{N} + \frac{1}{N} \sum_{t=1}^{N} \|\varphi_{t-1}\|^2 + \frac{1}{N} \sum_{t=1}^{N-1} n_{t+1}^2 \right],$$

with $\bar{k}$ = suitable constant. Observe now that all the terms in the right-hand side of (A.12) are $O(1)$. This, in particular, follows from the assumption of point (iii) in Theorem 3.2 that $\sum_{t=1}^{N} \|\varphi_{t-1}\|^2 = O(N)$ and Assumption 2.1, point 2. Then $\frac{1}{N} \sum_{t=1}^{N} \gamma_t \|\hat{\vartheta}_t - \hat{\vartheta}_{t-1}\| = O(1)$. Since γ_t tends to infinity, this last equation implies $\frac{1}{N} \sum_{t=1}^{N} \|\hat{\vartheta}_t - \hat{\vartheta}_{t-1}\| = o(1)$, that is, the thesis.

Proof of Proposition 3.3. Fix a real number $\epsilon > 0$ and a time instant N. Consider the set of instant points in the interval $[0, N]$ where $\tilde{\vartheta}_t := \vartheta^\circ - \hat{\vartheta}_t$ changes: $t_0, t_1, \ldots, t_{i(N)}$, where $i(N) := \max\{i : t_i \leq N\}$. In these instant points we define a set of subspaces $\{S_{t_i}\}_{i=0}^{i(N)}$ through the following backward recursive procedure:

for $i = i(N) + 1$, set $S_i = \emptyset$,

for $i = i(N), i(N) - 1, \ldots, 0$, set (here and throughout the symbol $\tilde{\vartheta}_{t,S}$ stands for the projection of vector $\tilde{\vartheta}_t$ onto the subspace S)

$$\text{(A.13)} \qquad S_{t_i} = \begin{cases} S_{t_{i+1}} & \text{if } \|\tilde{\vartheta}_{t_i, S_{t_{i+1}}^\perp}\| \leq \epsilon, \\ S_{t_{i+1}} \oplus span\{\tilde{\vartheta}_{t_i}\} & \text{otherwise.} \end{cases}$$

For each $t \in [0, N]$, with the notation $i(t) := \max\{i : t_i \leq t\}$, we have

$$|\varphi_t^T \tilde{\vartheta}_t|^p \leq c_p \, |\varphi^T_{t,S^\perp_{t_{i(t)}}} \tilde{\vartheta}_{t,S^\perp_{t_{i(t)}}}|^p + c_1 \, |\varphi^T_{t,S_{t_{i(t)}}} \tilde{\vartheta}_{t,S_{t_{i(t)}}}|^p, \tag{A.14}$$

where c_p is a suitable constant depending on p. By definition (A.13), the first term in the right-hand side can be upper bounded as follows:

$$|\varphi^T_{t,S^\perp_{t_{i(t)}}} \tilde{\vartheta}_{t,S^\perp_{t_{i(t)}}}|^p \leq \epsilon^p \|\varphi_t\|^p. \tag{A.15}$$

To handle the second term, we first work out a basis in $S_{t_{i(t)}}$. For this purpose, consider the subset $\{\tau_j\}_{j=1}^{dim(S_{t_0})}$ of instant points $\{t_i\}_{i=0}^{i(N)}$ such that subspace S_{t_i} enlarges: $S_{\tau_j} \supset S_{t_i}$, $t_i > \tau_j$. The searched basis is $\{\tilde{\vartheta}_{\tau_j}\}_{j=dim(S_{t_0})-dim(S_{t_{i(t)}})+1}^{dim(S_{t_0})}$.

In view of the uniform boundedness of $\tilde{\vartheta}_t$ and also considering the very definition of subspaces S_{t_i} (equation (A.13)), it is easy to see that vectors $\{\tilde{\vartheta}_{\tau_j}\}$ are spread in subspace $S_{t_{i(t)}}$ in such a way that the angle between each pair of vectors tends to zero only when $\epsilon \to 0$. Consequently, there exists a constant $c(\epsilon)$, depending on ϵ, but independent of N, such that term $|\varphi^T_{t,S_{t_{i(t)}}} \tilde{\vartheta}_{t,S_{t_{i(t)}}}|^p$ in the right-hand side of inequality (A.14) can be bounded as follows:

$$\begin{aligned} |\varphi^T_{t,S_{t_{i(t)}}} \tilde{\vartheta}_{t,S_{t_{i(t)}}}|^p &\leq \Delta^p \|\varphi_{t,S_{t_{i(t)}}}\|^p \\ &\leq \Delta^p c(\epsilon) \sum_{j=dim(S_{t_0})-dim(S_{t_{i(t)}})+1}^{dim(S_{t_0})} \|\varphi_{t,span\{\tilde{\vartheta}_{\tau_j}\}}\|^p, \end{aligned} \tag{A.16}$$

where $\Delta = \max_{\vartheta_1,\vartheta_2 \in \Theta} \|\vartheta_1 - \vartheta_2\|$.

By plugging estimates (A.15) and (A.16) in (A.14), we obtain

$$|\varphi_t^T \tilde{\vartheta}_t|^p \leq c_p \epsilon^p \|\varphi_t\|^p + c_p \, \Delta^p \, c(\epsilon) \sum_{j=dim(S_{t_0})-dim(S_{t_{i(t)}})+1}^{dim(S_{t_0})} \|\varphi_{t,span\{\tilde{\vartheta}_{\tau_j}\}}\|^p.$$

Summing up these relations from time $t = 0$ to $t = N$, we finally have

$$\sum_{t=0}^N |\varphi_t^T \tilde{\vartheta}_t|^p \leq c_p \epsilon^p \sum_{t=0}^N \|\varphi_t\|^p + c_p \, \Delta^p \, c(\epsilon) \sum_{t=0}^N \sum_{j=dim(S_{t_0})-dim(S_{t_{i(t)}})+1}^{dim(S_{t_0})} \|\varphi_{t,span\{\tilde{\vartheta}_{\tau_j}\}}\|^p. \tag{A.17}$$

Introduce now the time-varying set of instant points

$$\mathcal{B}_N := \cup_{j=1}^{dim(S_{t_0})} \{\tau_j, \tau_j + 1, \ldots, \tau_j + T - 1\},$$

where $T := \sup_{i \geq 0} T_i < \infty$ (see (A.6) in the proof of Proposition 3.1). Since $\dim(S_{t_0}) \leq n + m$, we obviously have $|\mathcal{B}_N| \leq T(n + m)$.

Then

$$\sum_{t=0,\, t \notin \mathcal{B}_N}^N \sum_{j=dim(S_{t_0})-dim(S_{t_{i(t)}})+1}^{dim(S_{t_0})} \|\varphi_{t,span\{\tilde{\vartheta}_{\tau_j}\}}\|^p \leq \sum_{j=1}^{dim(S_{t_0})} \sum_{t=0}^{\tau_j - 1} \|\varphi_{t,span\{\tilde{\vartheta}_{\tau_j}\}}\|^p.$$

We now show that

$$\text{(A.18)} \qquad \sum_{t=0}^{t_i-1} |\varphi_t^T \tilde{\vartheta}_{t_i}|^p = o\left(\sum_{t=0}^{t_i-1} \|\varphi_t\|^p\right) \quad \text{a.s.},$$

from which it follows that

$$\sum_{t=0,\ t\not\in\mathcal{B}_N}^{N} \ \sum_{j=dim(S_{t_0})-dim(S_{t_{i(t)}})+1}^{dim(S_{t_0})} \|\varphi_{t,span\{\tilde{\vartheta}_{\tau_j}\}}\|^p \leq \frac{n+m}{\epsilon^p}\left[o\left(\sum_{t=0}^{N}\|\varphi_t\|^p\right) + O(1)\right],$$
(A.19)

where we used the fact that $\dim(S_{t_0}) \leq n+m\ \forall N$.

Observe first that

$$\text{(A.20)} \qquad \sum_{t=0}^{t_i-1} \|\varphi_t\|^2 = O\left(\sum_{t=0}^{t_i-1} \|\varphi_t\|^p\right) \quad \text{a.s.}$$

Indeed, using Jensen's inequality [11, Corollary 1 in section 4.3])

$$\begin{aligned}\sum_{t=0}^{t_i-1} \|\varphi_t\|^2 &= t_i\left[\left(\frac{1}{t_i}\sum_{t=0}^{t_i-1}\|\varphi_t\|^2\right)^{p/2}\right]^{2/p} \\ &\leq t_i\left[\frac{1}{t_i}\sum_{t=0}^{t_i-1}\|\varphi_t\|^p\right]^{2/p} = \sum_{t=0}^{t_i-1}\|\varphi_t\|^p\left[\frac{t_i}{\sum_{t=0}^{t_i-1}\|\varphi_t\|^p}\right]^{1-2/p},\end{aligned}$$

where

$$\text{(A.21)} \qquad \limsup_{i\to\infty} \frac{t_i}{\sum_{t=0}^{t_i-1}\|\varphi_t\|^p} < \infty \quad \text{a.s.}$$

This last equation is easily derived as follows. From the regression-like form $y_t = \varphi_{t-1}^T\vartheta^\circ + n_t$, it follows that $|n_t|^p \leq 2^{p-1}\max\{\|\vartheta^\circ\|,1\}[\,|y_t|^p + \|\varphi_{t-1}\|^p\,]$. Taking into account that the autoregressive part of system is not trivial ($n>0$), this in turn implies that $|n_t|^p \leq h_1[\,\|\varphi_t\|^p + \|\varphi_{t-1}\|^p\,]$, from which it is easily shown that $\sum_{t=1}^{N-1}|n_t|^p \leq h_1\sum_{t=0}^{N-1}\|\varphi_t\|^p$, where h_1 is a suitable constant. Since $\frac{1}{N}\sum_{t=1}^{N-1}|n_t|^p \geq [\frac{1}{N}\sum_{t=0}^{N-1}n_t^2]^{p/2}$ (using Jensen's inequality), from Assumption 2.1, we then get

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N-1}\|\varphi_t\|^p > 0 \quad \text{a.s.},$$

from which (A.21) follows.

By means of (A.20), we now show that $\sum_{t=0}^{t_i-1}|\varphi_t^T\tilde{\vartheta}_{t_i}|^p = o(\sum_{t=0}^{t_i-1}\|\varphi_t\|^p)$ a.s., which implies (A.18). This equation is easily derived from property (i) in Theorem 3.2

as follows:

$$\begin{aligned}\sum_{t=0}^{t_i-1} |\varphi_t^T \tilde{\vartheta}_{t_i}|^p &\leq \left| \sum_{t=0}^{t_i-1} |\varphi_t^T(\vartheta^\circ - \vartheta_{t_i})|^2 \right|^{p/2} \\ &= o\left(\left(\operatorname{Log} \sum_{t=0}^{t_i-1} \|\varphi_t\| \right)^{p(1+\delta)/2} \right) \qquad \text{(by property (i))} \\ &= o\left(\sum_{t=0}^{t_i-1} \|\varphi_t\|^2 \right) \\ &= o\left(\sum_{t=0}^{t_i-1} \|\varphi_t\|^p \right) \qquad \text{(by (A.20)).}\end{aligned}$$

By using inequality (A.17) and inequality (A.19), we obtain

$$\begin{aligned}\sum_{t=0,\ t\notin \mathcal{B}_N}^{N} |\varphi_t^T \tilde{\vartheta}_t|^p &\leq c_p \epsilon^p \sum_{t=0}^{N} \|\varphi_t\|^p + c_p \Delta^p c(\epsilon) \frac{n+m}{\epsilon^p} \left[o\left(\sum_{t=0}^{N} \|\varphi_t\|^p \right) + O(1) \right] \\ &\leq c_p \epsilon^p O\left(\sum_{t=0}^{N} \|\varphi_t\|^p + N \right) + c_p \Delta^p c(\epsilon) \frac{n+m}{\epsilon^p} o\left(\sum_{t=0}^{N} \|\varphi_t\|^p + N \right),\end{aligned}$$

which finally implies that

$$\limsup_{N\to\infty} \frac{\displaystyle\sum_{t=0,\ t\notin \mathcal{B}_N}^{N} |\varphi_t^T \tilde{\vartheta}_t|^p}{\displaystyle\sum_{t=0}^{N} \|\varphi_t\|^p + N} \leq c_p \epsilon^p.$$

Since ϵ is arbitrarily chosen, the thesis follows.

REFERENCES

[1] K. Åström and B. Wittenmark, *On self-tuning regulators*, Automatica, 9 (1973), pp. 185–189.

[2] A. Becker, P. R. Kumar, and C. Z. Wei, *Adaptive control with the stochastic approximation algorithm: Geometry and convergence*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 330–338.

[3] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, Vols. I and II, Athena Scientific, Belmont, MA, 1995.

[4] B. Betrò and F. Schoen, *Sequential stopping rules for the multistart algorithm in global optimisation*, Math. Programming, 38 (1987), pp. 271–286.

[5] V. Borkar and P. P. Varaiya, *Adaptive control of Markov chains,* I: *Finite parameter set*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 953–958.

[6] M. C. Campi, *The problem of pole-zero cancellation in transfer function identification and application to adaptive stabilization*, Automatica, 32 (1996), pp. 849–857.

[7] M. C. Campi and P. R. Kumar, *Adaptive linear quadratic Gaussian control: The cost-biased approach revisited*, SIAM J. Control Optim., 36 (1998), pp. 1890–1907.

[8] H. F. Chen and L. Guo, *Convergence rate of least-squares identification and adaptive control for stochastic systems*, Internat. J. Control, 44 (1986), pp. 1459–1476.

[9] H. F. Chen and L. Guo, *Optimal adaptive control and consistent parameter estimates for ARMAX model with quadratic cost*, SIAM J. Control Optim., 25 (1987), pp. 845–867.

[10] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.

[11] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, 3rd ed., Springer Texts Statist., Springer-Verlag, New York, 1997.

[12] D. F. Delchamps, *Analytic feedback control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1031–1033.

[13] G. C. Goodwin, P. J. Ramadge, and P. E. Caines, *Discrete-time stochastic adaptive control*, SIAM J. Control Optim., 19 (1981), pp. 829–853.

[14] L. Guo, *Self-convergence of weighted least-squares with applications to stochastic adaptive control*, IEEE Trans. Automat. Control, 41 (1996), pp. 79–89.

[15] B. Hajek, *A tutorial survey of theory and applications of simulated annealing*, in Proceedings of the 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, December 1985, IEEE, Piscataway, NJ, 1985, pp. 755–760.

[16] P. Hall and C. Heyde, *Martingale limit theory and its application*, Probability and Mathematical Statistics, Z.W. Birnbaum and E. Lukacs, eds., Academic Press, New York, 1980.

[17] J. Kanniah, O. P. Malik, and G. S. Hope, *Self-tuning regulator based on dual-rate sampling*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 755–759.

[18] P. R. Kumar, *Optimal adaptive control of linear-quadratic-Gaussian systems*, SIAM J. Control Optim., 21 (1983), pp. 163–178.

[19] P. R. Kumar, *Simultaneous identification and adaptive control of unknown systems over finite parameter sets*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 68–76.

[20] P. R. Kumar, *Convergence of adaptive control schemes using least-squares parameter estimates*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 416–424.

[21] P. R. Kumar and A. Becker, *A new family of optimal adaptive controllers for Markov chains*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 137–146.

[22] P. R. Kumar and W. Lin, *Optimal adaptive controllers for unknown Markov chains*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 765–774.

[23] T. L. Lai and C. Z. Wei, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, Ann. Statist., 10 (1982), pp. 154–166.

[24] W. Lin and P. Kumar, *Stochastic control of a queue with two servers of different rates*, in Analysis and Optimization of Systems, A. Bensoussan and J.-L. Lions, eds., Lecture Notes in Control and Inform. Sci. 44, Springer-Verlag, New York, 1982, pp. 719–728.

[25] W. Lin, P. R. Kumar, and T. I. Seidman, *Will the self-tuning approach work for general cost criteria?*, System Control Lett., 6 (1985), pp. 77–85.

[26] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1999.

[27] R. Ortega, R. Kelly, and R. Lozano-Leal, *On global stability of adaptive systems using an estimator with parameter freezing*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 343–346.

[28] M. Prandini, *Adaptive LQG Control: Optimality Analysis and Robust Controller Design*, Ph.D. Thesis, University of Brescia, Brescia, Italy, 1998.

[29] M. Prandini, S. Bittanti, and M. C. Campi, *A penalized identification criterion for securing controllability in adaptive control*, J. Math. Systems Estim. Control, 8 (1998), pp. 491–494. (Retrieval code for the electronic version: 29460.)

[30] A. H. G. Rinnooy Kan and G. T. Timmer, *Stochastic global optimization methods.* I. *Clustering methods*, Math. Programming, 39 (1987), pp. 27–56.

[31] A. H. G. Rinnooy Kan and G. T. Timmer, *Stochastic global optimization methods.* II. *Multilevel methods*, Math. Programming, 39 (1987), pp. 57–78.

[32] K. S. Sin and G. C. Goodwin, *Stochastic adaptive control using a modified least squares algorithm*, Automatica, 18 (1982), pp. 815–321.

[33] J. H. van Schuppen, *Tuning of Gaussian stochastic control systems*, IEEE Trans. Automat. Control, AC-39 (1994), pp. 2178–2190.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1520–1539

ZERO-SUM STOCHASTIC GAMES IN BOREL SPACES: AVERAGE PAYOFF CRITERIA*

ONÉSIMO HERNÁNDEZ-LERMA[†] AND JEAN B. LASSERRE[‡]

Abstract. This paper is concerned with two-person zero-sum dynamic stochastic games in Borel spaces, with possibly unbounded payoff function, and several average (or ergodic) payoff criteria. We give conditions under which the long-run *expected average* payoff criterion, the *sample-path average* criterion, the existence of solutions to the average payoff *Shapley equations*, and a certain *"martingale condition"* are all equivalent.

Key words. zero-sum stochastic games, Borel spaces, expected average payoff, sample-path average payoff, Shapley equations

AMS subject classifications. 90D15, 93E05, 90D10

PII. S0363012999361962

1. Introduction.

In this paper we study noncooperative, discrete-time, two-person zero-sum stochastic games in Borel spaces with possibly unbounded payoff function. We consider three average (or ergodic) payoff criteria: the long-run *expected average payoff* (EAP), the *sample-path average payoff* (SPAP), and the existence of *canonical pairs* of stationary strategies, by which we mean a pair of strategies that satisfy the *Shapley equations* for the average payoff criterion (see Definition 4.3). We give conditions under which these criteria are all equivalent and also equivalent to a certain "martingale condition" (Theorem 5.9(c)).

This equivalence is not obvious at all. For instance, there are well-known counterexamples (e.g., [4]) showing that EAP optimality (Definition 4.1) does *not* imply the existence of solutions to the Shapley equation. Similarly, the comparison—not to mention the *equivalence*—between EAP optimality and SPAP optimality (Definition 4.2) is not quite straightforward, not even in the case of a bounded payoff function or for *Markov control processes* (MCPs), which correspond to the one-player case (see Remark 5.11).

The average payoff criteria have been widely studied for finite or countable state spaces (see, for instance, [1, 3, 6, 24, 26] and their references), but for *uncountable* spaces they are restricted to just a few publications, for instance, [7, 16, 17, 18, 21, 22, 25]. In the latter case, which is the one we are interested in, one can distinguish two main approaches: the "contraction" and the "vanishing discount" approaches. In the former, the idea is to impose conditions to ensure that Fan's minimax operator (see (4.15)) is a contraction map in some norm. This has been done in [7, 22, 25] using the "span" seminorm (see also section 3.3 in [9] for the one-player case), which requires the immediate payoff function to be *bounded*, and in [18] (see also [1, 24]) using the weighted w-norm in (5.2) below. The second, "vanishing discount," approach is used

*Received by the editors September 29, 1999; accepted for publication (in revised form) August 2, 2000; published electronically January 25, 2001. This research was partially supported by the CNRS (France)-CONACYT (México) Scientific Cooperation Program and by the ECOS (France)-ANUIES (México) Educational and Scientific Cooperation Program.

http://www.siam.org/journals/sicon/39-5/36196.html

[†]Departamento de Matemáticas, CINVESTAV-IPN, A. Postal 14-740, México D.F. 07000, México (ohernand@math.cinvestav.mx). The research of this author was also supported by CONACYT grant 32299-E.

[‡]LAAS-CNRS, 7 Avenue du Colonel Roche, 31077 Toulouse Cédex, France (lasserre@laas.fr).

in [21]. Here we use a *third approach* to obtain the Shapley equation, which is roughly related to the "policy iteration" (or Howard's) algorithm for MCPs.

At any rate, a common feature to all these works and ours is that they all require strong ergodicity conditions. For instance, the hypotheses in [7, 22, 25] guarantee geometric ergodicity in the *total variation* norm, whereas in [1, 18, 21, 24] this kind of ergodicity is with respect to a *weighted* w-*norm* (see (5.3) and (5.4)). This w-norm is of common use to analyze MCPs—see, for example, [8, 11, 12, 13, 14, 19] and their references.

The remainder of the paper is organized as follows. Sections 2 and 3 introduce standard material on stochastic games and strategies, respectively. The optimality criteria we are concerned with are presented in section 4. The core of the paper is contained in section 5: after introducing some assumptions, we present our main results on the *equivalence* of EAP optimality, the existence of "canonical pairs" of strategies (Theorem 5.8), a certain martingale condition (Theorem 5.9), and SPAP optimality (Theorem 5.10). These results are specialized in the obvious manner to MCPs (Corollary 5.12). Finally, after some technical preliminaries in section 6, the proofs of Theorems 5.8, 5.9, and 5.10 are presented in sections 7, 8, and 9, respectively.

2. The game model. In this section we introduce the (discrete-time, time-homogeneous) two-person zero-sum stochastic game model we are interested in. We begin by introducing the following terminology and notation—for further details the reader may refer to Bertsekas and Shreve [2], for instance.

DEFINITION 2.1. (a) *A Borel subset* X *of a complete and separable metric space is called a* Borel space*, and its Borel σ-algebra is denoted by* $\mathcal{B}(\mathrm{X})$*. We deal only with Borel spaces, and so "measurable" (for either sets or functions) always means "Borel-measurable." Given a Borel space* X*, we denote by* $\mathbb{P}(\mathrm{X})$ *the family of probability measures on* X*, endowed with the weak topology* $\sigma(\mathbb{P}(\mathrm{X}), C_b(\mathrm{X}))$*, where* $C_b(\mathrm{X})$ *stands for the space of continuous bounded functions on* X*. In this case,* $\mathbb{P}(\mathrm{X})$ *is a Borel space. Moreover, if* X *is compact, then so is* $\mathbb{P}(\mathrm{X})$*.*

(b) *Let* X *and* Y *be Borel spaces. A measurable function* $\varphi : \mathrm{Y} \to \mathbb{P}(\mathrm{X})$ *is called a* transition probability from Y *to* X *(also known as a* stochastic kernel *on* X *given* Y*), and we denote by* $\mathbb{P}(\mathrm{X}|\mathrm{Y})$ *the family of all those transition probabilities. If* φ *is in* $\mathbb{P}(\mathrm{X}|\mathrm{Y})$*, then we write its values either as* $\varphi(y)(B)$ *or as* $\varphi(B|y)$ *for all* $y \in \mathrm{Y}$ *and* $B \in \mathcal{B}(\mathrm{X})$*. Finally, if* $\mathrm{X} = \mathrm{Y}$*, then* φ *is called a* Markov *transition probability on* X*.*

The stochastic game model. We shall consider the two-person zero-sum game model

$$GM := (\mathrm{X}, A, B, \mathbb{K}_A, \mathbb{K}_B, Q, r), \tag{2.1}$$

where X is the *state space*, and A and B are the *action spaces* for players 1 and 2, respectively. These spaces are all assumed to be Borel spaces. The sets $\mathbb{K}_A \in \mathcal{B}(\mathrm{X}\times A)$ and $\mathbb{K}_B \in \mathcal{B}(\mathrm{X}\times B)$ are the *constraint sets.* That is, for each state $x \in \mathrm{X}$, the x-section in $\mathbb{K}_A$, namely,

$$A(x) := \{a \in A | (x, a) \in \mathbb{K}_A\},$$

represents the set of admissible actions for player 1 in the state x. Similarly, the x-section in $\mathbb{K}_B$,

$$B(x) := \{b \in B | (x, b) \in \mathbb{K}_B\},$$

stands for the family of admissible actions for player 2 in the state x. Let

$$\mathbb{K} := \{(x, a, b) | x \in \mathrm{X}, a \in A(x), b \in B(x)\},$$

which is a Borel subset of $\mathrm{X} \times A \times B$ (see Lemma 1.1 in [20], for instance). Then $Q \in \mathbb{P}(\mathrm{X}|\mathbb{K})$ is the game's *transition law*, and, finally, $r : \mathbb{K} \to \mathbb{R}$ is a measurable function representing the *reward function* for player 1 (or the *cost function* for player 2).

The game is played as follows. At each stage (or time) $t = 0, 1, \dots,$ players 1 and 2 observe the current state $x \in \mathrm{X}$ of the system and then independently choose actions $a \in A(x)$ and $b \in B(x)$, respectively. As a consequence of this, the following happens: (1) player 1 receives an immediate reward $r(x, a, b)$; (2) player 2 incurs a cost $r(x, a, b)$; and (3) the system moves to a new state x' with distribution $Q(\cdot\, |x, a, b)$. Thus, the goal of player 1 is to maximize his/her reward, whereas that of player 2 is to minimize his/her cost.

3. Strategies. Let $H_0 := \mathrm{X}$ and $H_t := \mathbb{K} \times H_{t-1}$ for $t = 1, 2, \dots$. For each t, an element $h_t = (x_0, a_0, b_0, \dots, x_{t-1}, a_{t-1}, b_{t-1}, x_t)$ of H_t represents a "history" of the game up to time t. A *strategy* for player 1 is then defined as a sequence $\pi^1 = \{\pi^1_t, t = 0, 1, \dots\}$ of transition probabilities π^1_t in $\mathbb{P}(A|H_t)$ such that

$$\pi^1_t(A(x_t))|h_t) = 1 \quad \forall h_t \in H_t,\ t = 0, 1, \dots .$$

We denote by Π_1 the family of all strategies for player 1.

Now define $\mathbb{A}(x) := \mathbb{P}(A(x))$ for each state $x \in \mathrm{X}$, and let Φ_1 be the class of all transition probabilities $\varphi \in \mathbb{P}(A|\mathrm{X})$ such that $\varphi(x)$ is in $\mathbb{A}(x)$ for all $x \in \mathrm{X}$. Then a strategy $\pi^1 = \{\pi^1_t\} \in \Pi_1$ is called *stationary* if there exists $\varphi \in \Phi_1$ such that

$$\pi^1_t(\cdot\, |h_t) = \varphi(x_t)(\cdot\,) \quad \forall h_t \in H_t,\ t = 0, 1, \dots .$$

We will identify Φ_1 with the family of stationary strategies for player 1.

The sets of strategies Π_2 and Φ_2 for player 2 are defined similarly, writing $B(x)$ and $\mathbb{B}(x) := \mathbb{P}(B(x))$ in lieu of $A(x)$ and $\mathbb{A}(x)$, respectively.

Let $(\Omega, \mathcal{F})$ be the (canonical) measurable space that consists of the sample space $\Omega := (\mathrm{X} \times A \times B)^\infty$ and its product σ-algebra $\mathcal{F}$. Then for each pair of strategies $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$ and each "initial state" $x \in \mathrm{X}$, there exists a probability measure P^{π^1,π^2}_x and an stochastic process $\{(x_t, a_t, b_t), t = 0, 1, \dots\}$ defined on $(\Omega, \mathcal{F})$ in a canonical way, where $x_t, a_t,$ and b_t represent the state and the actions of players 1 and 2, respectively, at each stage $t = 0, 1, \dots$. The expectation operator with respect to P^{π^1,π^2}_x is denoted by E^{π^1,π^2}_x.

Remark 3.1. As was already mentioned at the end of section 2, the players choose their actions *independently.* This means, more precisely, that for any pair of strategies $\pi^i = \{\pi^i_t\} \in \Pi_i$ $(i = 1, 2)$ and any initial state $x \in \mathrm{X}$, the corresponding action processes $\{a_t\}$ and $\{b_t\}$ are *conditionally independent* in the sense that

$$P^{\pi^1,\pi^2}_x(a_t \in C, b_t \in D|h_t) = \pi^1_t(C|h_t)\pi^2_t(D|h_t)$$

for all $C \in \mathcal{B}(A), D \in \mathcal{B}(B), h_t \in H_t,$ and $t = 0, 1, \dots .$

4. Average payoff criteria. For each $n = 1, 2, \dots$ and each history $h_\infty := (x_0, a_0, b_0, x_1, a_1, b_1, \dots)$, let

$$J^0_n(h_\infty) := \sum_{t=0}^{n-1} r(x_t, a_t, b_t) \tag{4.1}$$

be the *n-stage sample-path payoff* when player $i\,(i = 1, 2)$ uses the strategy $\pi^i \in \Pi_i$, given the initial state $x_0 = x$. The corresponding *n-stage expected payoff* is

$$J_n(\pi^1, \pi^2, x) := E_x^{\pi^1,\pi^2}\left[\sum_{t=0}^{n-1} r(x_t, a_t, b_t)\right]. \tag{4.2}$$

We then define the long-run SPAP

$$J^0(h_\infty) := \liminf_{n\to\infty} J_n^0(h_\infty)/n, \tag{4.3}$$

and, similarly, the long-run EAP

$$J(\pi^1, \pi^2, x) := \liminf_{n\to\infty} J_n(\pi^1, \pi^2, x)/n. \tag{4.4}$$

To introduce the optimality criteria we are concerned with we use the following concepts. The functions on X defined as

$$L(x) := \sup_{\pi^1\in\Pi_1} \inf_{\pi^2\in\Pi_2} J(\pi^1, \pi^2, x) \quad \text{and} \quad U(x) := \inf_{\pi^2\in\Pi_2} \sup_{\pi^1\in\Pi_1} J(\pi^1, \pi^2, x) \tag{4.5}$$

are called the *lower value* and the *upper value*, respectively, of the (expected) average payoff game. It is clear that $L(\cdot) \le U(\cdot)$ in general, but if it holds that $L(x) = U(x)$ for all $x \in$ X, then the common function is called the *value* of the game and is denoted by $V(\cdot)$.

DEFINITION 4.1. *Suppose that the game has a value $V(\cdot)$. Then a strategy π^{*1} in Π_1 is said to be* expected average payoff optimal *(briefly, EAP optimal)* for player 1 *if*

$$\inf_{\pi^2\in\Pi_2} J(\pi^{*1}, \pi^2, x) = V(x) \quad \forall x \in \text{X}. \tag{4.6}$$

Similarly, $\pi^{*2} \in \Pi_2$ is EAP optimal for player 2 *if*

$$\sup_{\pi^1\in\Pi_1} J(\pi^1, \pi^{*2}, x) = V(x) \quad \forall x \in \text{X}. \tag{4.7}$$

If $\pi^{*i} \in \Pi_i$ is EAP optimal for player $i\,(i = 1, 2)$, then (π^{*1}, π^{*2}) is called an EAP optimal pair *of strategies (also known as a* saddle point *or as a* noncooperative equilibrium*).*

For the SPAP we introduce a similar optimality criterion. (We use below the usual abbreviation "a.s." for "almost surely.")

DEFINITION 4.2. *Suppose that the game has a value $V(\cdot)$. Then a pair of strategies $(\pi^{*1}, \pi^{*2}) \in \Pi_1 \times \Pi_2$ is said to be* SPAP optimal *if it satisfies that for all $x \in$ X and $\pi^i \in \Pi_i\,(i = 1, 2)$,*

$$J^0(h_\infty) = V(x), \quad P_x^{\pi^{*1},\pi^{*2}} a.s., \tag{4.8}$$

$$J^0(h_\infty) \ge V(x), \quad P_x^{\pi^{*1},\pi^{2}} a.s., \tag{4.9}$$

$$J^0(h_\infty) \le V(x), \quad P_x^{\pi^{1},\pi^{*2}} a.s. \tag{4.10}$$

SPAP optimality is also studied in [3, 7, 24], for instance.

To introduce our last optimality criterion we use the following notation. (Recall that $\mathbb{A}(x) := \mathbb{P}(A(x))$ and $\mathbb{B}(x) := \mathbb{P}(B(x))$; see section 3.) For any given function $f : \mathbb{K} \to \mathbb{R}$ and probability measures $\varphi \in \mathbb{A}(x)$ and $\psi \in \mathbb{B}(x)$, we write

$$f(x, \varphi, \psi) := \int_{A(x)} \int_{B(x)} f(x, a, b)\psi(db)\varphi(da) \tag{4.11}$$

whenever the integrals are well defined. In particular, for r and Q as in (2.1),

$$r(x, \varphi, \psi) := \int_{A(x)} \int_{B(x)} r(x, a, b)\psi(db)\varphi(da)$$

and

$$Q(\cdot\,|x, \varphi, \psi) := \int_{A(x)} \int_{B(x)} Q(\cdot\,|x, a, b)\psi(db)\varphi(da).$$

DEFINITION 4.3. *A four-tuple $(\xi_*, u_*, \varphi_*, \psi_*)$ that consists of a constant $\xi_* \in \mathbb{R}$, a measurable function $u_* : \mathrm{X} \to \mathbb{R}$, and a pair $(\varphi_*, \psi_*) \in \Phi_1 \times \Phi_2$ of stationary strategies is said to be a* canonical four-tuple *if it holds that, for all $x \in \mathrm{X}$,*

$$\xi_* + u_*(x) = r(x, \varphi_*(x), \psi_*(x)) + \int_{\mathrm{X}} u_*(y)Q(dy|x, \varphi_*(x), \psi_*(x)) \tag{4.12}$$

$$= \max_{\varphi\in\mathbb{A}(x)} \left[r(x, \varphi, \psi_*(x)) + \int_{\mathrm{X}} u_*(y)Q(dy|x, \varphi, \psi_*(x)) \right] \tag{4.13}$$

$$= \min_{\psi\in\mathbb{B}(\mathrm{X})} \left[r(x, \varphi_*(x), \psi) + \int_{\mathrm{X}} u_*(y)Q(dy|x, \varphi_*(x), \psi) \right]. \tag{4.14}$$

In this case, it is also said that (φ_*, ψ_*) is a canonical pair of stationary strategies. (Concerning the name "canonical four-tuple," see the second paragraph in section 6.)

Definition 4.3 is of course related to the so-called *Shapley* (or *dynamic programming*) *equation*

$$\xi_* + u_*(x) = Tu_*(x) \quad \forall x \in \mathrm{X},$$

where T is the minimax operator defined by

$$\begin{aligned} Tu_*(x) &:= \max_{\varphi\in\mathbb{A}(x)} \min_{\psi\in\mathbb{B}(\mathrm{X})} \left[r(x, \varphi, \psi) + \int_{\mathrm{X}} u_*(y)Q(dy|x, \varphi, \psi) \right] \\ &= \min_{\psi\in\mathbb{B}(x)} \max_{\varphi\in\mathbb{A}(x)} \left[r(x, \varphi, \psi) + \int_{\mathrm{X}} u_*(y)Q(dy|x, \varphi, \psi) \right]. \end{aligned} \tag{4.15}$$

Our assumptions in section 5 will ensure that the maximum and the minimum in (4.15) are indeed attained and also that the second equality holds.

On the other hand, the relation between Definitions 4.3 and 4.1 is that if u_* satisfies that

$$\lim_{n\to\infty} n^{-1}E_x^{\pi^1,\pi^2} u_*(x_n) = 0 \tag{4.16}$$

for all $x \in \mathrm{X}$ and $\pi^i \in \Pi_i\,(i = 1, 2)$, then one can easily show that ξ_* *is the value* $V(\cdot\,)$ of the game and that *the canonical pair (φ_*, ψ_*) is EAP optimal.* (See the proof of Theorem 5.8 in section 7.) Thus, denoting by $(\Phi_1 \times \Phi_2)_{ca}$ the family of canonical pairs and by $(\Phi_1 \times \Phi_2)_{eap}$ the family of EAP optimal strategies, we have

$$(\Phi_1 \times \Phi_2)_{ca} \subset (\Phi_1 \times \Phi_2)_{eap} \tag{4.17}$$

if (4.16) holds. Moreover, the relation (4.17) is, in general, *strict*; in other words, as shown by well-known counterexamples (for instance, [4]), without suitable hypotheses, an EAP optimal pair is not necessarily canonical. The key fact to note here is that iteration of (4.12) yields

$$n\xi_* + u_*(x) = J_n(\varphi_*, \psi_*, x) + E_x^{\varphi_*,\psi_*} u_*(x_n) \tag{4.18}$$

for all $x \in \mathrm{X}$ and $n = 1, 2, \ldots$. This establishes an explicit relation between the n-stage averages $J_n(\varphi_*, \psi_*, x)/n$ and ξ_*, whereas for an arbitrary EAP optimal pair an expression such as (4.18) is virtually impossible to obtain. However, we show below (Theorem 5.8) that under appropriate assumptions we have *equality* in (4.17), i.e.,

$$(\Phi_1 \times \Phi_2)_{ca} = (\Phi_1 \times \Phi_2)_{eap}, \tag{4.19}$$

and in fact these sets also coincide with the family $(\Phi_1 \times \Phi_2)_{spap}$ of SPAP optimal pairs of stationary strategies (Theorem 5.10).

5. Main results.

As we already mentioned in section 1, our assumptions are an obvious variant of hypotheses previously used to study MCPs and stochastic games [1, 8, 11, 12, 13, 14, 18, 19, 21, 24]. In particular, the following Assumption 5.1 consists of standard continuity-compactness hypotheses, together with a growth condition (5.1) on the reward/cost function r.

Assumption 5.1. (a) For each state $x \in \mathrm{X}$, the (nonempty) sets $A(x)$ and $B(x)$ of admissible actions are compact.

(b) For each (x, a, b) in $\mathbb{K}$, $r(x, \cdot\,, b)$ is upper semicontinuous (u.s.c.) on $A(x)$, and $r(x, a, \cdot\,)$ is lower semicontinuous (l.s.c.) on $B(x)$.

(c) For each (x, a, b) in $\mathbb{K}$ and each bounded measurable function v on X, the functions

$$\int_{\mathrm{X}} v(y)Q(dy|x, \cdot\,, b) \quad \text{and} \quad \int_{\mathrm{X}} v(y)Q(dy|x, a, \cdot\,)$$

are continuous on $A(x)$ and $B(x)$, respectively.

(d) There exists a constant r_1 and a measurable function $w(\cdot\,) \geq 1$ on X such that

$$|r(x, a, b)| \leq r_1 w(x) \quad \forall (x, a, b) \in \mathbb{K}, \tag{5.1}$$

and, in addition, part (c) holds when v is replaced with w.

The next two assumptions are used to guarantee that the state process $\{x_t\}$ is "ergodic" in a suitable sense—see Remark 5.5(b).

Assumption 5.2. There exists a probability measure $\nu \in \mathbb{P}(\mathrm{X})$, a positive number $\alpha < 1$, and a measurable function $\beta : \mathbb{K} \to [0, 1]$ for which the following holds for all (x, a, b) in $\mathbb{K}$ and D in $\mathcal{B}(\mathrm{X})$:

(a) $Q(D|x, a, b) \geq \beta(x, a, b)\nu(D)$.

(b) $\int_{\mathrm{X}} w(y)Q(dy|x, a, b) \leq \alpha w(x) + \beta(x, a, b) \parallel \nu \parallel_w$, where $w(\cdot\,) \geq 1$ is the function in Assumption 5.1(d), and $\parallel \nu \parallel_w := \int w d\nu$.

(c) $\inf \int_{\mathrm{X}} \beta(x, \varphi(x), \psi(x))\nu(dx) > 0$, where the infimum is over all pairs (φ, ψ) in $\Phi_1 \times \Phi_2$.

Assumption 5.3. There exists a σ-finite measure λ on X with respect to which, for each pair (φ, ψ) in $\Phi_1 \times \Phi_2$, the Markov transition probability $Q(\cdot\,|x, \varphi(x), \psi(x))$ is λ-irreducible.

We next introduce some notation and then we mention some important consequences of the above assumptions.

DEFINITION 5.4. $\mathbb{B}_w(\mathrm{X})$ *denotes the linear space of real-valued measurable functions* u *on* X *with a finite* w*-norm, which is defined as*

$$\| u \|_w := \sup_{x \in \mathrm{X}} |u(x)|/w(x), \tag{5.2}$$

and $\mathbb{M}_w(\mathrm{X})$ *stands for the normed linear space of finite signed measures* μ *on* X *such that*

$$\| \mu \|_w := \int_{\mathrm{X}} w d|\mu| < \infty, \tag{5.3}$$

where $|\mu| := \mu^+ + \mu^-$ *denotes the total variation of* μ.

Note that the integral $\int u d\mu$ is finite for each u in $\mathbb{B}_w(\mathrm{X})$ and μ in $\mathbb{M}_w(\mathrm{X})$ because, by (5.2) and (5.3),

$$\left| \int u d\mu \right| \leq \| u \|_w \int w d|\mu| = \| u \|_w \| \mu \|_w < \infty.$$

Remark 5.5. Suppose that Assumptions 5.2 and 5.3 are satisfied. Then we have the following:

(a) For each pair (φ, ψ) in $\Phi_1 \times \Phi_2$ the state (Markov) process $\{x_t\}$ is *positive Harris recurrent*; hence, in particular, the Markov transition probability

$$Q(\cdot\,|x, \varphi(x), \psi(x))$$

admits a unique *invariant probability measure* in $\mathbb{M}_w(\mathrm{X})$, which will be denoted by $q(\varphi, \psi)$; thus

$$q(\varphi, \psi)(D) = \int_{\mathrm{X}} Q(D|x, \varphi(x), \psi(x)) q(\varphi, \psi)(dx) \quad \forall D \in \mathcal{B}(\mathrm{X}).$$

(b) $\{x_t\}$ is w*-geometrically ergodic*, that is, there exist positive constants $\theta < 1$ and M such that

$$\left| \int_{\mathrm{X}} u(y) Q^n(dy|x, \varphi(x), \psi(x)) - \int_{\mathrm{X}} u(y) q(\varphi, \psi)(dy) \right| \leq w(x) \| u \|_w M \theta^n \tag{5.4}$$

for every $u \in \mathbb{B}_w(\mathrm{X}), x \in \mathrm{X}$, and $n = 0, 1, \ldots$, where Q^n denotes the n-step Markov transition probability. This result follows from Lemmas 3.3 and 3.4 in [8], where it was *assumed* the positive Harris recurrence in part (a). However, as noted in Lemma 4.1 of [19], (a) follows from our current Assumptions 5.2 and 5.3.

The w-geometric ergodicity (5.4) was obtained in [8] following ideas from Kartashov [15]. It turns out, however, that (5.4) can be obtained in several different ways; see, for instance, Hordijk and Yushkevich [14], Küenle [18], Nowak [21], or section 7.3.D and section 10.2.C in [11]. For the *countable* case, see [1] or [24], for instance. For geometric ergodicity in the *total variation norm*, which is obtained by taking $w(\cdot) \equiv 1$ in (5.3), see [7], section 3.3 in [9], or the notes to section 7.3 in [11].

It should be noted that all of the examples in [1, 7, 11, section 10.9, 14, 19, 24] satisfy our Assumptions 5.1, 5.2, and 5.3, as well as Assumption 5.7 below (taking γ as the "counting measure" if X is a *countable* set).

Another important consequence of Assumptions 5.2 and 5.3, together with Assumption 5.1 is that, as proved in Theorem 3 of Nowak [21], the set $(\Phi_1 \times \Phi_2)_{eap}$ in (4.17) is nonempty. More precisely, we have the following.

PROPOSITION 5.6 (Nowak [21]). *If Assumptions* 5.1*,* 5.2*, and* 5.3 *are satisfied, then the (expected) average payoff game has a constant value, say, $V(x) = V^*$ for all $x \in \mathrm{X}$, and there exists an EAP optimal pair of stationary strategies.*

The hypotheses C5 and C6 used in [21] to prove Proposition 5.6 are somewhat different from our Assumptions 5.2 and 5.3, but the fact is that they also give the w-geometric ergodicity (5.4), which combined with standard dynamic programming arguments is a key tool to prove Proposition 5.6.

Here we use Proposition 5.6 as our point of departure, and, together with the following assumption, use it to prove our first main result, Theorem 5.8. In fact, Nowak [21] also obtains the Shapley equations, but our proof, in section 7, is quite different from his—see Remark 7.1. Moreover, he does not obtain the *equivalence* stated in Theorem 5.8.

Assumption 5.7. There exists a σ-finite measure γ on X and a strictly positive density function $g(x, a, b, \cdot\,)$ such that

$$Q(D|x, a, b) = \int_D g(x, a, b, y)\gamma(dy)$$

for all $D \in \mathcal{B}(\mathrm{X})$ and $(x, a, b) \in \mathbb{K}$.

Note that *Assumption* 5.7 *implies Assumption* 5.3 *with* $\lambda = \gamma$.

THEOREM 5.8. *If Assumptions* 5.1*,* 5.2*, and* 5.7 *are satisfied, then* (4.19) *holds, that is,*

$$(\Phi_1 \times \Phi_2)_{ca} = (\Phi_1 \times \Phi_2)_{eap}.$$

In fact, there exists a canonical four-tuple $(\xi_*, u_*, \varphi_*, \psi_*)$ with u_* in $\mathbb{B}_w(\mathrm{X})$ and (by Proposition 5.6*) $\xi_* = V^*$.*

In addition to the equivalence between EAP optimality and the existence of canonical four-tuples, the hypotheses of Theorem 5.8 give other characterizations of EAP optimality, as in Theorem 5.9 below, where we use the following notation.

Let $\mathcal{F}_n$ be the σ-algebra generated by (x_t, a_t, b_t) for $t = 0, \dots, n$, that is,

$$\mathcal{F}_n := \sigma\{x_0, a_0, b_0, \dots, x_n, a_n, b_n\}. \tag{5.5}$$

Moreover, let J_n^0 be as in (4.1) and ξ_*, u_* as in Theorem 5.8, and then define the stochastic process

$$M_n(h_\infty) := J_n^0(h_\infty) + u_*(x_n) - n\xi_* \quad \text{for} \quad n = 1, 2, \dots, \tag{5.6}$$

with $M_0(h_\infty) := u_*(x_0)$. Finally, let $\Delta : \mathbb{K} \to \mathbb{R}$ be the so-called *discrepancy function* given by

$$\Delta(x, a, b) := r(x, a, b) + \int_{\mathrm{X}} u_*(y)Q(dy|x, a, b) - u_*(x) - \xi_*. \tag{5.7}$$

From this definition of Δ it is immediate that (5.8), below, is just another way of expressing (4.12)–(4.14). Thus the equivalence of (a) and (b) in the following theorem is a direct consequence of Theorem 5.8.

THEOREM 5.9. *Under the hypotheses of Theorem* 5.8*, the following statements are equivalent:*

(a) *The pair* $(\varphi_*, \psi_*) \in \Phi_1 \times \Phi_2$ *is EAP optimal.*

(b) *For each* $x \in \mathrm{X}$

$$\Delta(x, \varphi_*(x), \psi_*(x)) = \max_{\varphi \in \mathbb{A}(x)} \Delta(x, \varphi, \psi_*(x)) = \min_{\psi \in \mathbb{B}(x)} \Delta(x, \varphi_*(x), \psi) = 0. \tag{5.8}$$

(c) *For each* $x \in \mathrm{X}, \pi^1 \in \Pi_1$*, and* $\pi^2 \in \Pi_2$

(c$_1$) $\{M_n(h_\infty), \mathcal{F}_n\}$ *is a* $P_x^{\varphi_*, \psi_*}$*-martingale,*

(c$_2$) $\{M_n(h_\infty), \mathcal{F}_n\}$ *is a* $P_x^{\varphi_*, \pi^2}$*-submartingale, and*

(c$_3$) $\{M_n(h_\infty), \mathcal{F}_n\}$ *is a* $P_x^{\pi^1, \psi_*}$*-supermartingale.*

On the other hand, if we add the "second order" condition (5.9) to (5.1), it turns out that the sets in (4.19) coincide with $(\Phi_1 \times \Phi_2)_{spap}$. That is, we have the following.

THEOREM 5.10. *Suppose that the hypotheses of Theorem* 5.8 *are satisfied and, in addition, there is a constant* $r_2 \geq 0$ *such that*

$$r^2(x, a, b) \leq r_2 w(x) \quad \forall (x, a, b) \in \mathbb{K}. \tag{5.9}$$

Then a pair of strategies in $\Phi_1 \times \Phi_2$ *is EAP optimal if and only if it is SPAP optimal; hence, by Theorem* 5.8*,*

$$(\Phi_1 \times \Phi_2)_{eap} = (\Phi_1 \times \Phi_2)_{ca} = (\Phi_1 \times \Phi_2)_{spap}. \tag{5.10}$$

To conclude this section and proceed to prove Theorems 5.8, 5.9, and 5.10, we shall specialize these theorems to MCPs or one-player stochastic games. To fix ideas we begin with the following obvious remark.

Remark 5.11. (a) If there is only one player, say player 1, then the game model GM in (2.1) reduces to the *Markov control model*

$$MCM = (\mathrm{X}, A, \mathbb{K}_A, \widehat{Q}, \widehat{r}), \tag{5.11}$$

where X, A, and $\mathbb{K}_A$ are exactly as in (2.1), but the transition law $\widehat{Q}$ and the reward function $\widehat{r}$ are defined on $\mathbb{K}_A$, that is, $\widehat{Q}$ is in $\mathbb{P}(\mathrm{X}|\mathbb{K}_A)$ and $\widehat{r} : \mathbb{K}_A \to \mathbb{R}$.

(b) Another way in which GM reduces to a Markov control model is to assume that one of the players, say player 2, selects a *fixed* stationary strategy ψ in Φ_2. In this case, the corresponding Markov control model is given by (5.11) with

$$\widehat{r}(x, a) := r(x, a, \psi(x)) \quad \text{and} \quad \widehat{Q}(\cdot\,|x, a) := Q(\cdot\,|x, a, \psi(x)) \;\; \forall x \in \mathrm{X}.$$

In either case, the optimality criteria in section 4 reduce in an obvious manner. In particular, the *value* of the MCP becomes

$$\widehat{V}(x) := \sup_{\pi \in \Pi_c} \widehat{J}(\pi, x) \,\text{for}\, x \in \mathrm{X}, \tag{5.12}$$

where Π_c is the set of all control strategies obtained from Π_1 (see section 3).

With this notation, Proposition 5.6 and Theorems 5.8, 5.9, and 5.10 yield the following.

COROLLARY 5.12. *Consider the MCP associated to* (5.11) *and suppose that Assumptions* 5.1, 5.2, *and* 5.3 *are satisfied. Then we have the following:*

(a) *There exists an* expected average reward *(EAR) optimal strategy* φ^* *in* Φ_1*, the set of stationary control strategies, and the value* $\widehat{V}(\cdot)$ *in* (5.12) *satisfies*

$$\widehat{J}(\varphi^*, x) = \widehat{V}(x) = \sup_{\varphi \in \Phi_1} \widehat{J}(\varphi, x) =: \xi^* \quad \forall x \in \mathrm{X}, \quad \textit{and}$$

(b) *there exists a* canonical triplet (ξ^*, h^*, φ^*) *with* $\xi^* \in \mathbb{R}, h^* \in \mathbb{B}_w(\mathrm{X})$*, and* $\varphi^* \in \Phi_1$*, that is (cf. Definition* 4.3*) for all* $x \in \mathrm{X}$*,*

$$\xi^* + h^*(x) = \widehat{r}(x, \varphi^*(x)) + \int_{\mathrm{X}} h^*(y)\widehat{Q}(dy|x, \varphi^*(x)) \tag{5.13}$$
$$= \max_{\varphi \in \mathbb{A}(x)} \left[\widehat{r}(x, \varphi) + \int_{\mathrm{X}} h^*(y)\widehat{Q}(dy|x, \varphi) \right].$$

In this case, φ^* *is said to be a* canonical strategy*.*

Furthermore, if we replace Assumption 5.3 *with Assumption* 5.7*, then the following statements* (c) *to* (f) *are equivalent:*

(c) $\varphi^* \in \Phi_1$ *is EAR optimal.*

(d) $\varphi^* \in \Phi_1$ *is a canonical strategy.*

(e) *For each* $x \in \mathrm{X}$*,*

$$\widehat{\Delta}(x, \varphi^*(x)) = \max_{\varphi \in \mathbb{A}(x)} \widehat{\Delta}(x, \varphi) = 0,$$

where $\widehat{\Delta} : \mathbb{K}_A \to \mathbb{R}$ *is the (average reward) discrepancy function:*

$$\widehat{\Delta}(x, a) := \widehat{r}(x, a) + \int_{\mathrm{X}} h^*(y)\widehat{Q}(dy|x, a) - h^*(x) - \xi^*.$$

(f) *For each* $\pi \in \Pi_c$ *and* $x \in \mathrm{X}$*, the stochastic process (cf.* (5.6)*)*

$$\widehat{M}_n(\widehat{h}_\infty) := \sum_{t=0}^{n-1} \widehat{r}(x_t, a_t) + h^*(x_n) - n\xi^* \quad \textit{for} \quad n = 1, 2, \dots,$$

with $\widehat{M}_0(\widehat{h}_\infty) := h^*(x_0)$ *and* $\widehat{h}_\infty = (x_0, a_0, x_1, a_1, \dots)$*, is a* P_x^π*-supermartingale (with respect to the* σ*-algebra generated by* $\{x_0, a_0, \dots, x_n, a_n\}$*—see* (5.5)*), whereas* $\{\widehat{M}_n(\widehat{h}_\infty), n = 0, 1, \dots\}$ *is a* $P_x^{\varphi^*}$*-martingale. Finally, if in addition* $\widehat{r}(x, a)$ *satisfies* (5.9)*, then each of the statements* (c) *to* (f) *is equivalent to the following:*

(g) $\varphi^* \in \Phi$ *is* sample-path average reward *(SPAR)* optimal*, that is (with* ξ^* *as in* (a) *and the obvious changes in* (4.3)*),*

$$J^0(\widehat{h}_\infty) = \xi^* \quad P_x^{\varphi^*} \, a.s. \ \ \forall x \in \mathrm{X}$$

and

$$J^0(\widehat{h}_\infty) \le \xi^* \quad P_x^\pi \ a.s. \ \ \forall \pi \in \Pi_c, x \in \mathrm{X}.$$

Proof. Parts (a) and (b) are well known—see, for instance, part (a) of Theorem 10.3.6 in [11]. Furthermore, part (b) of the same theorem shows that if $\varphi^* \in \Phi_1$ is EAR optimal, then φ^* satisfies (5.13) for *almost every* $x \in \mathrm{X}$ (with respect to some probability measure on X). However, under Assumption 5.7, it follows from Theorem 5.8 that (5.13) *holds for all* $x \in \mathrm{X}$, which gives the equivalence of (c) and (d). The remaining parts follow from Theorems 5.9 and 5.10. □

The rest of the paper is devoted to proving Theorems 5.8, 5.9, and 5.10 in sections 7, 8, and 9, respectively.

6. Preliminaries. In this section we present some concepts and preliminary results needed to prove the theorems in section 5. Some of these results are well known, but we state them here for completeness and ease of reference.

First of all, let us recall that if $P \in \mathbb{P}(\mathrm{X}|\mathrm{X})$ is a Markov transition probability on X and $c : \mathrm{X} \to \mathbb{R}$ is a given measurable function, then the equation

$$\xi + u(x) = c(x) + \int_{\mathrm{X}} u(y) P(dy|x) \quad \forall x \in \mathrm{X} \tag{6.1}$$

is called the (strictly unichain) *Poisson equation* (P.E.) for P with "charge" c, and the pair $(\xi, u(\cdot))$, with $\xi \in \mathbb{R}$ and $u(\cdot) : \mathrm{X} \to \mathbb{R}$, is called a solution to the P.E. This solution is also known as a *canonical pair*, which partly explains the name of "canonical triplet" for (ξ^*, h^*, φ^*) in (5.13), and of "canonical four-tuple" for $(\xi_*, u_*, \varphi_*, \psi_*)$ in (4.12). Note, in particular, that (4.12) is the P.E. for the Markov transition probability and the charge given by

$$P(\cdot\,|x) := Q(\cdot\,|x, \varphi_*(x), \psi_*(x)) \quad \text{and} \quad c(x) := r(x, \varphi_*(x), \psi_*(x)), \tag{6.2}$$

respectively. The following lemma shows, in particular, how to obtain the P.E. associated to an arbitrary pair of strategies in $\Phi_1 \times \Phi_2$.

LEMMA 6.1. *Suppose that Assumptions* 5.1, 5.2, *and* 5.3 *hold. Then for each pair of stationary strategies* $(\varphi, \psi) \in \Phi_1 \times \Phi_2$ *we have the following:*

(a) *The (finite) constant*

$$j(\varphi, \psi) := \int_{\mathrm{X}} r(x, \varphi(x), \psi(x)) q(\varphi, \psi)(dx),$$

with $q(\varphi, \psi)$ *as in Remark* 5.5(a), *is such that* (4.3) *and* (4.4) *can be written, for all* $x \in \mathrm{X}$, *as*

$$J^0(h_\infty) = \lim_{n\to\infty} J_n^0(h_\infty)/n = j(\varphi, \psi) \quad P_x^{\varphi,\psi} a.s. \tag{6.3}$$

and

$$J(\varphi, \psi, x) = \lim_{n\to\infty} J_n(\varphi, \psi, x)/n = j(\varphi, \psi), \tag{6.4}$$

respectively.

(b) *The function* $h_{\varphi,\psi}$ *defined on* X *as*

$$\begin{aligned} h_{\varphi,\psi}(x) &:= \lim_{n\to\infty} [J_n(\varphi, \psi, x) - nj(\varphi, \psi)] \\ &= \sum_{t=0}^{\infty} E_x^{\varphi,\psi} [r(x_t, \varphi(x_t), \psi(x_t)) - j(\varphi, \psi)] \end{aligned}$$

belongs to $\mathbb{B}_w(\mathrm{X})$ *(the set in Definition* 5.4*), and, moreover, its* w*-norm is independent of* (φ, ψ)*; in fact,*

$$\| h_{\varphi,\psi} \|_w \leq r_1 M/(1-\theta) \quad \forall (\varphi, \psi) \in \Phi_1 \times \Phi_2,$$

with r_1 *as in* (5.1), *and* M *and* θ *as in* (5.4).

(c) *The pair* $(j(\varphi, \psi), h_{\varphi,\psi})$ *in* $\mathbb{R} \times \mathbb{B}_w(\mathrm{X})$ *is the unique solution of the P.E.*

$$j(\varphi, \psi) + h_{\varphi,\psi}(x) = r(x, \varphi(x), \psi(x)) + \int_{\mathrm{X}} h_{\varphi,\psi}(y) Q(dy|x, \varphi(x), \psi(x)) \tag{6.5}$$

(*compare with* (6.1)) *that satisfies the condition*

$$\int_{\mathrm{X}} h_{\varphi,\psi}(x) q(\varphi,\psi)(dx) = 0.$$

Proof. The convergence in (6.3), which is a special case of the strong law of large numbers for Markov chains, follows from the positive Harris recurrence in Remark 5.5(a). (If necessary, for further details see Theorem 11.2.1 and Corollary 11.2.2 in [11], for instance.) The remaining statements in the lemma are part of Proposition 10.2.3 in [11]. □

Following Küenle [18], one can also get the P.E. (6.5) using a "contraction argument."

LEMMA 6.2. *Suppose that Assumptions* 5.2 *and* 5.3 *are satisfied. Let* (φ,ψ) *be an arbitrary pair in* $\Phi_1 \times \Phi_2$*, and define*

$$P(\cdot\,|x) := Q(\cdot\,|x,\varphi(x),\psi(x)) \quad and \quad \mu(\cdot\,) := q(\varphi,\psi)(\cdot\,). \tag{6.6}$$

If a function u *in* $\mathbb{B}_w(\mathrm{X})$ *is* P-subharmonic *(or subinvariant), that is,*

$$u(x) \le \int_{\mathrm{X}} u(y) P(dy|x) \quad \forall x \in \mathrm{X},$$

then $u(\cdot\,)$ *is constant* μ*-a.e.; in fact,*

$$u(\cdot\,) = \inf_{x\in\mathrm{X}} u(x) = \int_{\mathrm{X}} u d\mu \quad \mu\text{-}a.e.$$

(with P *and* μ *as in* (6.6)*). Similarly, if* u *is* P-superharmonic *(or superinvariant), that is,* $u(x) \ge \int u(y)P(dy|x)$ *for all* $x \in \mathrm{X}$*, then*

$$u(\cdot\,) = \sup_{x\in\mathrm{X}} u(x) = \int_{\mathrm{X}} u d\mu \quad \mu\text{-}a.e.$$

Proof. See, for instance, Lemma 7.5.12 in [11]. □

Let μ_1 and μ_2 be two measures on X and recall that, by definition, μ_1 is *absolutely continuous* with respect to μ_2 (in symbols: $\mu_1 \ll \mu_2$) if $\mu_2(D) = 0$, with D in $\mathcal{B}(\mathrm{X})$, implies $\mu_1(D) = 0$. In addition, μ_1 is *equivalent* to μ_2 if $\mu_1 \ll \mu_2$ and $\mu_2 \ll \mu_1$.

LEMMA 6.3. *Suppose that Assumptions* 5.2 *and* 5.7 *hold, and let* γ *be as in the latter assumption. Then, for any pair* (φ,ψ) *in* $\Phi_1 \times \Phi_2$,

(a) γ *is equivalent to the invariant probability measure* $q(\varphi,\psi)$*; hence,*

(b) *if* $q(\varphi,\psi)(D) = 0$*, then* $Q(D|x,a,b) = 0$ *for all* (x,a,b) *in* $\mathbb{K}$.

Proof. (a) If $\gamma(D) = 0$, then (by Assumption 5.7)

$$Q(D|x,a,b) = \int_D g(x,a,b,y)\gamma(dy) = 0 \quad \forall (x,a,b) \in \mathbb{K}. \tag{6.7}$$

Therefore, by the invariance of $q(\varphi,\psi)$ (see Remark 5.5(a)),

$$q(\varphi,\psi)(D) = \int_{\mathrm{X}} Q(D|x,\varphi(x),\psi(x)) q(\varphi,\psi)(dx) = 0;$$

that is, $q(\varphi,\psi) \ll \gamma$.

To prove the converse, we already noted that Assumption 5.7 implies Assumption 5.3 with the irreducibility measure $\lambda = \gamma$. On the other hand, if a Markov chain is λ-irreducible and has an invariant probability measure, say, μ, then $\lambda \ll \mu$. (For a proof of this fact see, for instance, Theorem 7.2 in [23].) Therefore, taking μ as $q(\varphi, \psi)$, we conclude that $\lambda = \gamma$ is absolutely continuous with respect to $q(\varphi, \psi)$.

(b) By part (a), $q(\varphi, \psi)(D) = 0$ implies $\gamma(D) = 0$, which in turn yields (6.7). □

The following result, which in particular gives (4.16), follows from straightforward calculations using (5.2) and the inequality in Assumption 5.2(b); see, for example, [8, 12] or Lemma 10.4.1 in [11].

Lemma 6.4. *Suppose that* (5.1) *and Assumptions* 5.1(a) *and* 5.2(b) *are satisfied, and let* $k := 1 + \| \nu \|_w /(1-\alpha)$. *Then for any* $\pi^i \in \Pi_i$ $(i = 1, 2)$, $x \in \mathrm{X}$, *and* $n = 0, 1, \ldots$

(a) $E_x^{\pi^1,\pi^2} w(x_n) \le k w(x)$, *and*

(b) $E_x^{\pi^1,\pi^2} |r(x_n, a_n, b_n)| \le r_1 k w(x)$.

In particular, by (a) *and* (5.2), *for any function* u *in* $\mathbb{B}_w(\mathrm{X})$ *we have*

(c) $E_x^{\pi^1,\pi^2} |u(x_n)| \le \| u \|_w \, k w(x)$, *and so*

(d) $\lim_{n\to\infty} n^{-1} E_x^{\pi^1,\pi^2} |u(x_n)| = 0$, *and the convergence is uniform on* $\Pi_1 \times \Pi_2$.

Finally, let us consider the minimax operator T in (4.15). Choose an arbitrary function u in $\mathbb{B}_w(\mathrm{X})$ and let

$$H(u; x, a, b) := r(x, a, b) + \int_{\mathrm{X}} u(y) Q(dy|x, a, b) \quad \text{for} \quad (x, a, b) \in \mathbb{K}. \tag{6.8}$$

By Assumptions 5.1(c) and the second part of 5.1(d), the integral in (6.8) is continuous in both $a \in A(x)$ and $b \in B(x)$ (see Lemma 8.3.7(a) in [11], for instance). This fact and Assumption 5.1(b) yield that $H(u; x, \cdot, b)$ is u.s.c. on $A(x)$ and that $H(u; x, a, \cdot)$ is l.s.c. on $B(x)$. Therefore (using the notation in (4.11)), the function $H(u; x, \varphi, \psi)$ is u.s.c. in $\varphi \in \mathbb{A}(x) := \mathbb{P}(A(x))$ and l.s.c. in $\psi \in \mathbb{B}(\mathrm{X}) := \mathbb{P}(B(x))$; see, for example, the "extended Fatou Lemma" 8.3.7(b) and the statement (12.3.37) in [11, p. 225]. Moreover, $H(u; x, \varphi, \psi)$ is concave (as it is linear) in φ and convex in ψ. Thus, by Fan's minimax theorem [5] and well-known minimax measurable selection theorems (see, for instance, [20]), we get the following.

Lemma 6.5. *Suppose that Assumptions* 5.1 *and* 5.2(b) *hold. Then* Tu *is in* $\mathbb{B}_w(\mathrm{X})$ *for each* u *in* $\mathbb{B}_w(\mathrm{X})$, *and there exist stationary strategies* $\varphi^* \in \Phi_1$ *and* $\psi^* \in \Phi_2$ *such that, for all* $x \in \mathrm{X}$,

$$\begin{aligned} Tu(x) &= H(u; x, \varphi^*(x), \psi^*(x)) \\ &= \max_{\varphi \in \mathbb{A}(x)} H(u; x, \varphi, \psi^*(x)) \\ &= \min_{\psi \in \mathbb{B}(\mathrm{X})} H(u; x, \varphi^*(x), \psi). \end{aligned} \tag{6.9}$$

Proof. By the remarks in the previous paragraph, we need only to show that if u is in $\mathbb{B}_w(\mathrm{X})$, then so is Tu. This follows from (5.1), (5.2), and Assumption 5.2(b), which give, for any $(x, a, b) \in \mathbb{K}$,

$$\begin{aligned} |H(u; x, a, b)| &\le r_1 w(x) + \| u \|_w \int_{\mathrm{X}} w(y) Q(dy|x, a, b) \\ &\le (r_1 + \|u\|_w (\alpha + \|\nu\|_w)) w(x) \end{aligned} \tag{6.10}$$

because $0 \le \beta(x, a, b) \le 1$ and $w(\cdot) \ge 1$. Thus, by (6.9) and (6.10), Tu is indeed in $\mathbb{B}_w(\mathrm{X})$. □

7. Proof of Theorem 5.8. We begin with the easiest (in fact, standard) part (4.17). That is, we wish to show that if $(\xi_*, u_*, \varphi_*, \psi_*)$ is a canonical four-tuple with u_* in $\mathbb{B}_w(\mathrm{X})$, then (φ_*, ψ_*) is EAP optimal.

Dividing by n both sides of (4.18) and letting $n \to \infty$, Lemma 6.4(d) and (6.4) give

$$J(\varphi_*, \psi_*, x) = j(\varphi_*, \psi_*) = \xi_* \quad \forall x \in \mathrm{X}. \tag{7.1}$$

Thus, to complete the proof that (φ_*, ψ_*) is EAP optimal it remains only to prove that ξ_* equals the value V^* in Proposition 5.6, i.e.,

$$\xi_* = V^*. \tag{7.2}$$

To prove this let us first note that (4.13) gives

$$\xi_* + u_*(x) \geq r(x, \varphi, \psi_*(x)) + \int_{\mathrm{X}} u_*(y) Q(dy|x, \varphi, \psi_*(x)) \tag{7.3}$$

for all $x \in \mathrm{X}$ and $\varphi \in \mathbb{A}(x)$. Then, by standard dynamic programming arguments (see [1, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19]),

$$n\xi_* + u_*(x) \geq J_n(\pi^1, \psi_*, x) + E_x^{\pi^1, \psi_*} u_*(x_n) \tag{7.4}$$

for any $\pi^1 \in \Pi_1, x \in \mathrm{X}$, and $n = 1, 2, \ldots$. Multiplying by $1/n$ both sides of (7.4) and taking the lim inf as $n \to \infty$, it follows from (4.4) and Lemma 6.4(d) that

$$\xi_* \geq J(\pi^1, \psi_*, x) \geq \inf_{\pi^2 \in \Pi_2} J(\pi^1, \pi^2, x) \quad \forall \pi^1 \in \Pi_1, x \in \mathrm{X}. \tag{7.5}$$

This implies that $\xi^* \geq L(x) = V^*$ for all $x \in \mathrm{X}$.

A similar argument, but using now (4.14), gives

$$\xi_* \leq J(\varphi_*, \pi^2, x) \leq \sup_{\pi^1 \in \Pi_1} J(\pi^1, \pi^2, x) \quad \forall \pi^2 \in \Pi_2, x \in \mathrm{X}, \tag{7.6}$$

so that $\xi_* \leq U(x) = V^*$ for all $x \in \mathrm{X}$, and so (7.2) follows. This completes the proof of (4.17).

To prove (4.19), let us now take an EAP optimal pair (φ_*, ψ_*). We wish to prove that (φ_*, ψ_*) is a canonical pair, so that it satisfies (4.12)–(4.14) with $\xi_* = V^*$ and some function u_* in $\mathbb{B}_w(\mathrm{X})$. With this in mind, first note that, by (6.4) and Proposition 5.6, we have

$$J(\varphi_*, \psi_*, x) = j(\varphi_*, \psi_*) = V^* \quad \forall x \in \mathrm{X}. \tag{7.7}$$

Thus, by Lemma 6.1(b), (c), there is a function u_0 in $\mathbb{B}_w(\mathrm{X})$ such that the P.E. (6.5) for (φ_*, ψ_*) can be written as

$$V^* + u_0(x) = r(x, \varphi_*(x), \psi_*(x)) + \int_{\mathrm{X}} u_0(y) Q(dy|x, \varphi_*(x), \psi_*(x)) \tag{7.8}$$

for all $x \in \mathrm{X}$. It follows that

$$V^* + u_0(x) \geq \min_{\psi \in \mathbb{B}(\mathrm{X})} \left[r(x, \varphi_*(x), \psi) + \int_{\mathrm{X}} u_0(y) Q(dy|x, \varphi_*(x), \psi) \right], \tag{7.9}$$

which (by Lemma 6.5) yields

$$V^* + u_0(x) \geq r(x, \varphi_*(x), \psi_0(x)) + \int_X u_0(y) Q(dy|x, \varphi_*(x), \psi_0(x)) \tag{7.10}$$

for all $x \in X$ and some $\psi_0 \in \Phi_2$. On the other hand, iteration of (7.10) gives that (as in (7.3)–(7.5)), for all $x \in X$,

$$V^* \geq J(\varphi_*, \psi_0, x) \geq \inf_{\pi^2 \in \Pi_2} J(\varphi_*, \pi^2, x) = V^*,$$

where the latter equality is due to (4.6). In other words,

$$J(\varphi_*, \psi_0, x) = j(\varphi_*, \psi_0) = V^* \quad \forall x \in X, \tag{7.11}$$

and so ψ_0 is a best reply of player 2 to the EAP optimal strategy φ_* of player 1. Now, write the P.E. for (φ_*, ψ_0) using (7.11). Then, repeating the argument used to obtain (7.8), there is a function u_1 in $\mathbb{B}_w(X)$ such that, for all $x \in X$,

$$V^* + u_1(x) = r(x, \varphi_*(x), \psi_0(x)) + \int_X u_1(x) Q(dy|x, \varphi_*(x), \psi_0(x)). \tag{7.12}$$

If we now subtract (7.12) from (7.10), we see that

$$u_0(x) - u_1(x) \geq \int_X [u_0(y) - u_1(y)] \, Q(dy|x, \varphi_*(x), \psi_0(x)) \quad \forall x \in X,$$

so that $u_0 - u_1$ is superharmonic with respect to the Markov transition probability $Q(\cdot\,|x, \varphi_*(x), \psi_0(x))$. Consequently, by Lemma 6.2, there is a constant k_1 and a Borel set $D_1 \subset X$, with $q(\varphi_*, \psi_0)(D_1) = 1$, such that

$$u_0(x) = u_1(x) + k_1 \quad \forall x \in D_1.$$

Finally, define

$$v_0(\cdot\,) := u_0(\cdot\,) = u_1(\cdot\,) + k_1 \,\text{on}\, D_1$$

and

$$v_0(x) := \min_{\psi \in \mathbb{B}(X)} \left[r(x, \varphi_*(x), \psi) + \int_X u_0(y) Q(dy|x, \varphi_*(x), \psi) \right] - V^* \tag{7.13}$$

for $x \in D_1^c$, where $D_1^c := X \backslash D_1$ denotes the complement of D_1. Then, as

$$q(\varphi_*, \psi_0)(D_1^c) = 0,$$

Lemma 6.3(b) gives that $Q(D_1^c|x, a, b) = 0$ for all (x, a, b) in $\mathbb{K}$, which in turn implies that in the integral in (7.13) we may replace u_0 with v_0. Thus,

$$V^* + v_0(x) = \min_{\psi \in \mathbb{B}(x)} \left[r(x, \varphi_*(x), \psi) + \int_X v_0(y) Q(dy|x, \varphi_*(x), \psi) \right] \tag{7.14}$$

for all $x \in X$.

We now go back to the initial P.E. (7.8), but instead of taking the "minimum" as in (7.9), we now take the "maximum," i.e.,

$$\begin{aligned} V^* + u_0(x) &\le \max_{\varphi \in \mathbb{A}(x)} \left[r(x, \varphi, \psi_*(x)) + \int_{\mathrm{X}} u_0(y) Q(dy|x, \varphi, \psi_*(x)) \right] \\ &= r(x, \varphi_0(x), \psi_*(x)) + \int_{\mathrm{X}} u_0(y) Q(dy|x, \varphi_0(x), \psi_*(x)) \end{aligned}$$

for some $\varphi_0 \in \Phi_1$. Then, with obvious changes, the arguments used in (7.11)–(7.14) give the existence of a Borel set D_2 with $q(\varphi_0, \psi_*)(D_2) = 1$ and a function v_1 in $\mathbb{B}_w(\mathrm{X})$ such that

$$v_1(\cdot\,) = u_0(\cdot\,) \,\text{on}\, D_2, \tag{7.15}$$

and

$$V^* + v_1(x) = \max_{\varphi \in \mathbb{A}(x)} \left[r(x, \varphi, \psi_*(x)) + \int_{\mathrm{X}} v_1(y) Q(dy|x, \varphi, \psi_*(x)) \right] \tag{7.16}$$

for all $x \in \mathrm{X}$.

To conclude, let $D := D_1 \cap D_2$ and define $u_*(\cdot\,) := u_0(\cdot\,)$ on D and

$$u_*(x) := Tu_0(x) - V^* \quad \text{for} \quad x \in D^c. \tag{7.17}$$

By Lemma 6.3(a), we have $\gamma(D_1^c) = \gamma(D_2^c) = 0$, and so $\gamma(D^c) = 0$, which by Assumption 5.7 gives $Q(D^c|x, a, b) = 0$ for all (x, a, b) in $\mathbb{K}$. Hence, instead of (7.17) we may write

$$V^* + u_*(x) = Tu_*(x) \quad \forall x \in \mathrm{X}, \tag{7.18}$$

which together with (7.8), (7.12)–(7.14), and (7.15)–(7.16) gives that $(V^*, u_*, \varphi_*, \psi_*)$ is a canonical four-tuple. □

Remark 7.1. It is interesting to note that the previous proof is quite different from Nowak's [21] "vanishing discount" proof of a solution to the Shapley equations, but still at the last step of his proof he uses an argument such as that in (7.17)–(7.18). Moreover, while we assume a strictly positive density $g(x, a, b, \cdot\,)$ (see Assumption 5.7), he uses a condition on g that implies the continuity of $(a, b) \mapsto Q(\cdot\,|x, a, b)$ in the *total variation norm* for each $x \in \mathrm{X}$; see hypothesis C7 in [21]. Thus, it is unclear (to us, at least) if his hypotheses would allow a proof such as ours, and, conversely, if our assumptions would allow a proof such as his.

8. Proof of Theorem 5.9. It is evident that (5.8) is simply another way of writing (4.12)–(4.14). Therefore, the equivalence of (a) and (b) follows from Theorem 5.8. (For a different way of getting this equivalence see Remark 8.1 below.)

Now choose an arbitrary initial state $x \in \mathrm{X}$ and arbitrary strategies π^i in Π_i ($i = 1, 2$), and let $\mathcal{F}_n$ be as in (5.5). To prove that (b) *implies* (c), first note that

$$E_x^{\pi^1,\pi^2}\left[u_*(x_{n+1})|\mathcal{F}_n\right] = \int_{\mathrm{X}} u_*(y) Q(dy|x_n, a_n, b_n), \tag{8.1}$$

and, therefore, from (5.7),

(8.2)

$$E_x^{\pi^1,\pi^2}\left[\Delta(x_n, a_n, b_n)|\mathcal{F}_n\right] = r(x_n, a_n, b_n) + E_x^{\pi^1,\pi^2}\left[u_*(x_{n+1})|\mathcal{F}_n\right] - u_*(x_n) - \xi_*.$$

Let us now consider the stochastic process in (5.6). From Lemma 6.4(b), (c), it follows that this process is integrable with respect to $P_x^{\pi^1,\pi^2}$, i.e.,

$$E_x^{\pi^1,\pi^2}|M_n(h_\infty)| < \infty \quad \text{for each} \quad n = 0, 1, \ldots .$$

Moreover,

$$M_{n+1}(h_\infty) = M_n(h_\infty) + r(x_n, a_n, b_n) + u_*(x_{n+1}) - u_*(x_n) - \xi_*,$$

so that, by (8.2),

$$E_x^{\pi^1,\pi^2}\left[M_{n+1}(h_\infty)|\mathcal{F}_n\right] = M_n(h_\infty) + E_x^{\pi^1,\pi^2}\left[\Delta(x_n, a_n, b_n)|\mathcal{F}_n\right]. \tag{8.3}$$

This expression immediately yields that (5.8) implies (c). Indeed, if

$$\Delta(x, \varphi_*(x), \psi_*(x)) = 0 \;\; \forall x \in \mathrm{X}, \tag{8.4}$$

then, by (8.3),

$$E_x^{\varphi_*,\psi_*}\left[M_{n+1}(h_\infty)|\mathcal{F}_n\right] = M_n(h_\infty) \;\; \forall n = 0, 1, \ldots, \tag{8.5}$$

and (c$_1$) follows. Similarly, if

$$\min_{\psi \in \mathbb{B}(x)} \Delta(x, \varphi_*(x), \psi) = 0 \;\; \forall x \in \mathrm{X}, \tag{8.6}$$

so that

$$\Delta(x, \varphi_*(x), \psi) \geq 0 \;\; \forall x \in \mathrm{X}, \psi \in \mathbb{B}(x), \tag{8.7}$$

then (8.3) yields

$$E_x^{\varphi_*,\pi^2}\left[M_{n+1}(h_\infty)|\mathcal{F}_n\right] \geq M_n(h_\infty) \;\; \forall n = 0, 1, \ldots, \tag{8.8}$$

and we get (c$_2$). Finally, if

$$\max_{\varphi \in \mathbb{A}(x)} \Delta(x, \varphi, \psi_*(x)) = 0, \tag{8.9}$$

then

$$\Delta(x, \varphi, \psi_*(x)) \leq 0 \;\; \forall x \in \mathrm{X}, \varphi \in \mathbb{A}(x). \tag{8.10}$$

Thus, from (8.3),

$$E_x^{\pi^1,\psi_*}\left[M_{n+1}(h_\infty)|\mathcal{F}_n\right] \leq M_n(h_\infty) \;\; \forall n = 0, 1, \ldots,$$

and (c$_3$) follows.

The converse, (c) *implies* (b), is just as easy; first note that taking expectations in both sides of (8.3) we get

$$E_x^{\pi^1,\pi^2} M_{n+1}(h_\infty) = E_x^{\pi^1,\pi^2} M_n(h_\infty) + E_x^{\pi^1,\pi^2} \Delta(x_n, a_n, b_n). \tag{8.11}$$

Therefore, if (c$_1$) holds, then

$$E_x^{\varphi_*,\psi_*} \Delta(x_n, a_n, b_n) = 0,$$

which for $n = 0$ gives (8.4). Similarly, under (c$_2$), (8.11) gives

$$E_x^{\varphi_*,\pi^2}\Delta(x_n, \varphi_*(x_n), \pi_n^2(h_n)) \geq 0$$

for all $x \in \mathrm{X}, h_n \in H_n$ (see section 3), and $n = 0, 1, \ldots$. In particular, for $n = 0$ we obtain

$$\Delta(x, \varphi_*(x), \pi_0^2(x)) \geq 0 \quad \forall x \in \mathrm{X},$$

and so, as $\pi^2 \in \Pi_2$ was arbitrary, we get (8.7) and (8.6). Finally, from (8.11) and (c$_3$) we deduce (8.10) and (8.9). □

Remark 8.1. Let $J_n(\Delta; \pi^1, \pi^2, x)$ and $J(\Delta; \pi^1, \pi^2, x)$ be the functions in (4.2) and (4.4), respectively, obtained by replacing $r(x, a, b)$ with the discrepancy function $\Delta(x, a, b)$; that is,

$$J_n(\Delta; \pi^1, \pi^2, x) := E_x^{\pi^1,\pi^2}\left[\sum_{t=0}^{n-1}\Delta(x_t, a_t, b_t)\right] \tag{8.12}$$

and

$$J(\Delta; \pi^1, \pi^2, x) := \liminf_{n\to\infty} J_n(\Delta; \pi^1, \pi^2, x)/n. \tag{8.13}$$

Now take expectations on both sides of (8.2) and then sum over n to get

$$J_n(\Delta; \pi^1, \pi^2, x) = J_n(\pi^1, \pi^2, x) + E_x^{\pi^1,\pi^2}u_*(x_n) - u_*(x) - n\xi_*.$$

Thus multiplying by $1/n$ and taking lim inf as $n \to \infty$, Lemma 6.4(d) and (8.13) yield

$$J(\pi^1, \pi^2, x) = \xi_* + J(\Delta; \pi^1, \pi^2, x). \tag{8.14}$$

From this expression we can immediately deduce the equivalence of (a) and (b) in Theorem 5.9. (A similar expression for the SPAP criterion is obtained in (9.5).)

9. Proof of Theorem 5.10. We shall use the notation in (8.12) and (8.13) but for the SPAP criterion, that is,

$$J_n^0(\Delta; h_\infty) := \sum_{t=0}^{n-1}\Delta(x_t, a_t, b_t) \tag{9.1}$$

and

$$J^0(\Delta; h_\infty) := \liminf_{n\to\infty} J_n^0(\Delta; h_\infty)/n,$$

where $x_0 = x \in \mathrm{X}$ and $\pi^i \in \Pi_i$ $(i = 1, 2)$ are arbitrary. Moreover, with $\mathcal{F}_n$ as in (5.5), we consider the stochastic processes

$$\begin{aligned} Y_t(\pi^1, \pi^2) : &= u_*(x_t) - E_x^{\pi^1,\pi^2}\left[u_*(x_t)|\mathcal{F}_{t-1}\right] \\ &= u_*(x_t) - \int_{\mathrm{X}} u_*(y)Q(dy|x_{t-1}, a_{t-1}, b_{t-1}) \end{aligned} \tag{9.2}$$

for $t = 1, 2, \ldots,$ and

$$S_n(\pi^1, \pi^2) := \sum_{t=1}^{n} Y_t(\pi^1, \pi^2). \tag{9.3}$$

From (8.1), (8.2), and (9.1), we can also write the latter process as

$$S_n(\pi^1,\pi^2) = u_*(x_t) - u_*(x_0) - J_n^0(\Delta; h_\infty) + J_n^0(h_\infty) - n\xi_*. \tag{9.4}$$

On the other hand, (9.2), (9.3), and straightforward calculations show that $S_n(\pi^1,\pi^2)$ is a $P_x^{\pi^1,\pi^2}$-martingale with respect to $\mathcal{F}_n$, and, in addition, Lemma 11.3.11 in [11], which is the same as Lemma 4.4 in [13], gives that $P_x^{\pi^1,\pi^2}$ a.s.

$$\lim_{n\to\infty} n^{-1} S_n(\pi^1,\pi^2) = 0 \quad \text{and} \quad \lim_{n\to\infty} n^{-1} u_*(x_n) = 0.$$

Therefore, multiplying both sides of (9.4) by n^{-1} and letting $n \to \infty$ we get

$$J^0(h_\infty) = \xi_* + J^0(\Delta; h_\infty) \quad P_x^{\pi^1,\pi^2}\text{a.s.} \tag{9.5}$$

We next use (9.5) to prove the “only if” part of Theorem 5.10.

Suppose that $(\varphi_*,\psi_*) \in \Phi_1 \times \Phi_2$ is EAP optimal, that is, (φ_*,ψ_*) is in $(\Phi_1 \times \Phi_2)_{eap}$. Then write (5.8) more explicitly, as in (8.4), (8.6)–(8.7), and (8.9)–(8.10), to get the following: (8.4), (9.5), and (6.3) yield (4.8) with $V(\cdot) = \xi_*$, that is,

$$J^0(h_\infty) = \xi_* \quad P_x^{\varphi_*,\psi_*}\text{a.s.} \tag{9.6}$$

Similarly, (9.5) and (8.7) give

$$J^0(h_\infty) \geq \xi_* \quad P_x^{\varphi_*,\pi^2}\text{a.s.}, \tag{9.7}$$

whereas (9.5) and (8.10) give

$$J^0(h_\infty) \leq \xi_* \quad P_x^{\pi^1,\psi_*}\text{a.s.} \tag{9.8}$$

Hence, as $x \in \mathrm{X}, \pi^1 \in \Pi_1$ and $\pi^2 \in \Pi_2$ were arbitrary, we conclude (from Definition 4.2) that (φ_*,ψ_*) is SPAP optimal with the SPAP optimal payoff $V(\cdot) = \xi_*$.

Conversely, suppose that (φ_*,ψ_*) satisfies (9.6)–(9.8) for all $x \in \mathrm{X}$ and $\pi^i \in \Pi_i$ $(i = 1, 2)$. Then, by (9.6), (6.3), and (6.4), we see that (φ_*,ψ_*) satisfies

$$J(\varphi_*,\psi_*,x) = \xi_* \quad \forall x \in \mathrm{X}. \tag{9.9}$$

On the other hand, as (9.8) holds for all $\pi^1 \in \Pi_1 \supset \Phi_1$, using again (6.3) and (6.4) we see that

$$J(\varphi,\psi_*,x) = j(\varphi,\psi_*) \leq \xi_* \quad \forall \varphi \in \Phi_1, x \in \mathrm{X}; \tag{9.10}$$

thus, if we *fix* $\psi_* \in \Phi_2$ as in Remark 5.11(b), then (9.9), (9.10), and Corollary 5.12(a) yield

$$\sup_{\pi^1 \in \Pi_1} J(\pi^1,\psi_*,x) = \xi_* \quad \forall x \in \mathrm{X}. \tag{9.11}$$

In like manner, with the obvious (notational) changes in Remark 5.11(b) and Corollary 5.12(a), we can see that (9.7) implies that

$$\inf_{\pi^2 \in \Pi_2} J(\varphi_*,\pi^2,x) = \xi_* \quad \forall x \in \mathrm{X}. \tag{9.12}$$

Thus, from (9.11) and (9.12), it follows that the pair (φ_*,ψ_*) is EAP optimal. □

REFERENCES

[1] E. Altman, A. Hordijk and F.M. Spieksma, *Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards*, Math. Oper. Res., 22 (1997), pp. 588–618.

[2] D.P. Bertsekas and S.E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[3] V.S. Borkar and M.K. Ghosh, *Denumerable state stochastic games with limiting average payoff*, J. Optim. Theory Appl., 76 (1993), pp. 539–560.

[4] R. Cavazos-Cadena, *A counterexample on the optimality equation in Markov decision chains with the average cost criterion*, Systems Control Lett., 16 (1991), pp. 387–392.

[5] K. Fan, *Minimax theorems*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 42–47.

[6] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer-Verlag, New York, 1997.

[7] M.K. Ghosh and A. Bagchi, *Stochastic games with average payoff criterion*, Appl. Math. Optim., 38 (1998), pp. 283–301.

[8] E. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: Existence of canonical policies*, Appl. Math. (Warsaw), 23 (1995), pp. 199–218.

[9] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.

[10] O. Hernández-Lerma and J.B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[11] O. Hernández-Lerma and J.B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[12] O. Hernández-Lerma and O. Vega-Amaya, *Infinite-horizon Markov control processes with undiscounted cost criteria: From average to overtaking optimality*, Appl. Math. (Warsaw), 25 (1998), pp. 153–178.

[13] O. Hernández-Lerma, O. Vega-Amaya, and G. Carrasco, *Sample-path optimality and variance-minimization of average cost Markov control processes*, SIAM J. Control Optim., 38 (1999), pp. 79–93.

[14] A. Hordijk and A.A. Yushkevich, *Blackwell optimality in the class of all policies in Markov decision chains with a Borel state space and unbounded rewards*, Math. Methods Oper. Res., 49 (1999), pp. 421–448.

[15] N.V. Kartashov, *Strong Stable Markov Chains*, VSP, Utrecht, The Netherlands, 1996.

[16] H.-U. Küenle, *Stochastische Spiele und Entscheidungsmodelle*, Teubner-Texte Math. 89, Teubner-Verlag, Leipzig, 1986.

[17] H.-U. Küenle, *Stochastic games with complete information and average cost criteria*, in Advances in Dynamic Games and Applications, J.A. Filar, V. Gaitsgory, and K. Mizukami, eds., Ann. Internat. Soc. Dynam. Games 5, Birkhäuser, Boston, 2000, pp. 325–338.

[18] H.-U. Küenle, *On multichain Markov games*, Ann. Internat. Soc. Dynam. Games, to appear.

[19] F. Luque and O. Hernández-Lerma, *Semi-Markov control models with average costs*, Appl. Math. (Warsaw), 26 (1999), pp. 315–331.

[20] A.S. Nowak, *Measurable selection theorems for minimax stochastic optimization problems*, SIAM J. Control Optim., 23 (1985), pp. 466–476.

[21] A.S. Nowak, *Optimal strategies in a class of zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 50 (1999), pp. 399–419.

[22] A.S. Nowak, *Zero-sum average payoff stochastic games with general state space*, Games Econom. Behav., 7 (1994), pp. 221–232.

[23] S. Orey, *Lecture Notes on Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand Reinhold, London, 1971.

[24] O. Passchier, *The Theory of Markov Games and Queueing Control*, Ph.D. thesis, Department of Mathematics and Computer Science, Leiden University, The Netherlands, 1996.

[25] U. Rieder, *Average optimality in Markov game with general state space*, in Proceedings of the 3rd International Conference on Approximation Theory and Optimization, Puebla, México, 1995. Also available online at www.emis.de/proceedings/.

[26] L.I. Sennott, *Zero-sum stochastic games with unbounded cost: Discounted and average cost cases*, Z. Oper. Res., 39 (1994), pp. 209–225.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1540–1557

PATHWISE OPTIMALITY IN STOCHASTIC CONTROL*

PAOLO DAI PRA†, GIOVANNI B. DI MASI†, AND BARBARA TRIVELLATO†

Abstract. We introduce a notion of pathwise optimality for stochastic control problems over an infinite time horizon, and give sufficient conditions for the existence of pathwise optimal controls. We analyze both diffusion processes and processes with discrete state space.

Key words. stochastic optimal control, Hamilton–Jacobi–Bellman equations, pathwise optimality

AMS subject classification. 93E20

PII. S0363012998334778

1. Introduction.

Suppose we are given a probability space $(\Omega, \mathcal{F}, P)$ and a family of stochastic processes $x^u = (x_t^u)_{t\in[0,T]}$, indexed by a control parameter u. A stochastic control problem is defined by assigning real valued cost function $J(x^u, u)$ and consists in minimizing its expectation $E(J(x^u, u))$ over the controls.

This approach to optimization is quite appropriate when one deals with a controlled phenomenon that replicates many times, but may lack of reliability for a single trial. Indeed, even if a control u^* that minimizes the expected cost is known, the random variable $J(x^{u^*}, u^*)$ may have large fluctuations about its mean, making the choice of the control u^* a risky one. Various modifications of the performance index $E(J(x^u, u))$ have been proposed, in order to take these fluctuations into account. For example, one may penalize fluctuations around the mean by including in the index higher moments of the cost $J(x^u, u)$. An approach in this direction consists in minimizing over u the index $E\{e^{\mu J(x^u,u)}\}$, where $\mu > 0$ is a *risk* parameter (*risk sensitive control*).

Another point of view consists in looking for controls that are pathwise optimal, in some suitable sense. The most naive aim would be to find a control u^* such that $J(x^{u^*}, u^*) \leq J(x^u, u)$ almost surely (a.s.) for every control u. This exceedingly strong notion of optimality can be weakened by saying that a control u^* is a.s. optimal if for any other control u one can realize on *some* probability space $(\Omega', \mathcal{F}', P')$ the processes X^{u^*}, X^u having the same law of x^{u^*} and x^u, respectively, and such that $J(X^{u^*}, u^*) \leq J(X^u, u)$ P'-a.s. It is clear that any control that is a.s. optimal also minimizes the mean cost. In most interesting cases, however, as shown in [13], there is no optimal control for this notion of optimality.

The picture changes when one considers stochastic control problems over an infinite time horizon. Suppose the processes x^u to be defined in the whole time interval $[0, +\infty)$ and that the controls are themselves stochastic processes $u = (u_t)_{t\geq 0}$. We assign a real valued function $c(x_t^u, u_t)$ representing the running cost, and we let

$$J_T(u) = \int_0^T c(x_t^u, u_t)dt$$

*Received by the editors February 27, 1998; accepted for publication (in revised form) September 5, 2000; published electronically January 31, 2001.

http://www.siam.org/journals/sicon/39-5/33477.html

†Dipartimento di Matematica Pura e Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy (daipra@math.unipd.it, dimasi@math.unipd.it, trivellato@math.unipd.it).

(here and in what follows we simply write $J_T(u)$ rather than $J_T(x^u, u)$). For a given nonincreasing function $g : [0, +\infty) \to (0, +\infty)$ with $\lim_{T\to+\infty} g(T) = 0$, we say that a control u^* is g-optimal a.s. (or in probability) if for any other control u the random variables

$$g(T)[J_T(u^*) - J_T(u)]^+$$

go to zero a.s. (or in probability) as $T \to +\infty$. This notion of optimality, which we believe has been first introduced by Rotar, has been studied in special cases in [14, 18, 11, 12, 2, 19, 16, 7, 15, 3].

The purpose of this paper is to give results on almost sure optimality and optimality in probability for rather general nonlinear systems. In particular, we give conditions for the existence of an optimal control. These conditions can be verified in many interesting examples, including the LQG models treated, e.g., in [7], for which we can slightly weaken the assumptions, as well as several nonlinear models.

We begin in section 2 by giving our main definitions. Sections 3 and 5 are devoted to almost sure optimality and optimality in probability for controlled diffusion processes. Several examples are given in section 4. Sections 6 and 7 contain analogous results for point processes. In section 8 we consider controlled Markov chains with finite state space. In this case we can establish the existence of pathwise optimal controls in great generality.

2. Basic definitions.

In this paper we consider stochastic processes with values in $\mathbb{R}^d$ and whose trajectories belong to path space D. Here D is either $\mathcal{C}([0, +\infty), \mathbb{R}^d)$, provided with the topology of uniform convergence on the compact subsets of $[0, +\infty)$ or the set of *cadlag* functions $D = D([0, +\infty), \mathbb{R}^d)$, provided with the Skorohod topology (see, e.g., [5, Chapter 3, section 12]). Whenever the notion of measurability in D will occur, it will be meant with respect to the Borel σ-field associated to the above-mentioned topologies.

Let U be a measurable space. For any $v \in U$ we are given a Markov operator L^v, acting on a suitable domain $\mathcal{D}^v$ of functions from $\mathbb{R}^d$ to $\mathbb{R}$. We assume that there exists a subspace $\mathcal{C}$ of $\mathcal{D}^v$, which does not depend on v, such that the operator L^v is the closure of its restriction $\mathcal{C}$. The set $\mathcal{C}$ is called a *core* for the Markov operators L^v. Progressively measurable processes $u : [0, +\infty) \times D \to U$ are called *controls* (we shall write $u_t(x)$ for the value of u at time t on the path x). Suppose that a nonnegative measurable function $c : \mathbb{R}^d \times U \to [0, +\infty)$ and a probability measure μ on $\mathbb{R}^d$ are given. A control u is said to be *admissible* if the following conditions are satisfied:

1. For every $t \in [0, +\infty)$, the function $u_t : D \to U$ is measurable with respect to the σ-field $\mathcal{F}_t$ generated by the projections $\{\Pi_s : s \leq t\}$, where, for $x \in D$, $\Pi_s(x) = x_s$.
2. There exists a probability measure P^u on D such that for every $f \in \mathcal{C}$ the process

$$z_t = f(x_t) - \int_0^t (L^{u_s} f)(x_s) ds$$

is a P^u-local martingale, and $P^u \circ \Pi_0^{-1} = \mu$.
3. Let

$$J_t(u) = \int_0^t c(x_s, u_s) ds.$$

Then $J_t(u) < +\infty$ P^u-a.s. for all $t > 0$.

The set of admissible controls will be denoted by $\mathcal{U}$. Consider a nonincreasing function $g : [0, +\infty) \to (0, +\infty)$ such that $\lim_{t\to+\infty} g(t) = 0$.

DEFINITION 2.1. *We say that a control $u^* \in \mathcal{U}$ is g-optimal a.s. (respectively, in probability) if for all $u \in \mathcal{U}$ and for all probability measures P^{u,u^*} on $D \times D$ having marginals P^u and P^{u^*} and such that*

$$P^{u,u^*}\{(x, y) \in D \times D : x_0 \neq y_0\} = 0 \tag{2.1}$$

we have

$$\lim_{T\to+\infty} g(T)[J_T(u^*) - J_T(u)]^+ = 0 \qquad P^{u,u^*}\text{-a.s.} \tag{2.2}$$

(respectively, in probability with respect to P^{u,u^*}).

Remark 1. A special and interesting choice for g is $g(t) = 1/t$. Under rather mild assumptions (e.g., uniform integrability of $J_T(u)$) one can show that $1/t$-optimality implies optimality for the average cost per unit time

$$\limsup_{T\to+\infty} T^{-1}E^u(J_T(u)),$$

where E^u denotes expectation with respect to P^u. The results in this paper concern g-optimality for $g(t) = t^{-\alpha}$ for all $\alpha > 1/2$.

Remark 2. We have chosen to introduce our stochastic processes in a weak sense, i.e., as probability measures on the path space. In some cases there is a natural probability space over which all processes corresponding to admissible controls can be realized. This is the case, for instance, when considering diffusion processes, if one chooses to deal with strong solutions. In other cases, however, e.g., for counting processes, it is not so, the reason being that there is no natural "noise" that accounts for the randomness in the system. Thus, in the definition of controlled Markov process, we preferred not to make reference to any given probability space. The notion of g-optimality has been adapted accordingly. Note that our definition is stronger than the one given in previous works on this subject, since we require (2.2) to hold for any "coupling" P^{u,u^*} of P^u and P^{u^*}.

Remark 3. In the examples that we give in sections 4 and 8, where pathwise optimality is actually shown, condition (2.1) on the measures P^{u,u^*} would be irrelevant, so that we obtain an even stronger notion of optimality. In general, however, condition (2.1) is relevant. To see this, let X_1, X_2 be two disjoint subsets of $\mathbb{R}^d$. For $i = 1, 2$ let L_i^v, $v \in U$, be the generator of a controlled Markov process, with initial law μ_i and law P_i^u, u being an admissible control. Assume $P_i^u(x_t \in X_i) = 1$ for any admissible control u, i.e., the process is X_i-valued. Finally, let u_i^* be a g-optimal control, associated to a given running cost function $c_i(x, v)$.

Now consider the $X_1 \cup X_2$-valued process with generator $L^v f(x) = L_i^v f(x)$ if $x \in X_i$, initial condition $(\mu_1 + \mu_2)/2$, and running cost $c(x, v) = c_i(x, v)$ for $x \in X_i$. Clearly the control $u^* = u_i^*$ if $x_0 \in X_i$ is g-optimal for this combined problem.

Suppose now to remove condition (2.1). Then we can construct a coupling P^{u,u^*} satisfying $P^{u,u^*}(x_0 \in X_1, y_0 \in X_2 \text{ or } x_0 \in X_2, y_0 \in X_1) = 1$. In other words, under this P^{u,u^*}, the two process evolve with different generator and different cost function. There is no reason, therefore, for (2.2) to hold.

3. Pathwise optimality for controlled diffusions. In this section we assume the Markov operator L^v to be of the form

$$L^v = \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}(x,v)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{d} f_i(x,v)\frac{\partial}{\partial x_i},$$

where $a_{ij}, f_i : \mathbb{R}^d \times U \to \mathbb{R}$ are measurable functions and, for each $(x,u) \in \mathbb{R}^d \times U$, the matrix $a(x,u) = (a_{ij}(x,u))_{ij}$ is positive semidefinite.

Suppose $u \in \mathcal{U}$, and let P^u be the probability measure on $D = \mathcal{C}([0,+\infty), \mathbb{R}^d)$ introduced in the previous section. It is well known that P^u is the law of a weak solution of the stochastic differential equation

$$\begin{array}{rcl} dx_t & = & f(x_t,u_t)dt + \sigma(x_t,u_t)dw_t, \\ x_0 & \sim & \mu, \end{array}$$

where σ is a (possibly degenerate) square root of a and w_t is a d-dimensional Brownian motion.

The basic tool for solving the optimal control problem for diffusions is provided by the following partial differential equations, which will be referred to as the Hamilton–Jacobi–Bellman (HJB) equation and the stationary Hamilton–Jacobi–Bellman (SHJB) equation:

$$\text{(HJB)} : \begin{cases} \dfrac{\partial V_T}{\partial t}(t,x) + \inf\limits_{v \in U}[L^v V_T(t,x) + c(x,v)] = 0, \\ V_T(T,x) = 0. \end{cases}$$

$$\text{(SHJB)} : \inf_{v \in U}[L^v \phi(x) + c(x,v)] = \lambda.$$

Note that the (SHJB) equation is an equation for the pair (ϕ, λ), where λ is a real number. For generalities on these equations and their connections to stochastic control we refer the reader to [4].

We now give two different sets of assumptions on the model. We will show in Section 4 that these assumptions are satisfied for various classes of models.

The function g below is a positive, nonincreasing function vanishing at infinity.

Assumption A.

A1. For each $T > 0$ there exists a solution $V_T \in \mathcal{C}^{1,2}([0,T] \times \mathbb{R}^d)$ of (HJB).

A2. There exists a solution (ϕ, λ) of (SHJB), with $\phi \in \mathcal{C}^2(\mathbb{R}^d)$. The "inf" in (SHJB) is attained at $v = k(x)$, and the feedback $u_t^* = k(x_t)$ is an admissible control. Moreover, there exist constants $C, \epsilon > 0$ for which

$$E^{P^{u^*}}\left[\left(\phi^-(x_t)\right)^{2+\epsilon} \right] < C$$

for all $t \geq 0$, where ϕ^- denotes the negative part of ϕ.

A3. The following inequality holds:

$$\limsup_{T \to +\infty} g(T)\Big[\phi(x_0) + \lambda T - V_T(0,x_0)\Big] \leq 0 \quad \mu\text{-a.s.}$$

A4. There is a constant $B > 0$ such that

$$\left\| \frac{\partial V_T}{\partial x}(t,x)\sigma(x,v) \right\|^2 \leq B(c(x,v)+1),$$

uniformly in t, T.

A5. As A4, with ϕ replacing V_T.

Moreover, there exists an integer $m > 0$ such that the following further assumptions hold.

A6. There exists a constant $C > 0$ such that

$$E^{P^{u^*}}\left\{\left[\int_0^T c(x_t, u_t)\, dt\right]^m\right\} \leq CT^m$$

for all $T > 0$.

A7. Define $q(t) = t^{1/2}g(t)$. Then

$$\int_1^{+\infty} q^m(t)dt < +\infty.$$

Assumption B. There exists a solution (ϕ, λ) of (SHJB), with ϕ bounded. Moreover, Assumptions A2, A5, A6, and A7 hold.

We can now state the main results, whose proof is deferred to section 5.

THEOREM 3.1. *Suppose that either Assumption* A *or Assumption* B *holds. Then the feedback u^* is g-optimal a.s.*

THEOREM 3.2. *Suppose that either Assumption* A *or Assumption* B *holds, but in both cases condition* A7 *is weakened to $q(t) = o(1)$. Then the feedback u^* is g-optimal in probability.*

Remark 4. Note that if one can show that A6 holds for *all* m large enough, then one can choose $g(t) = t^{-(1/2+\delta)}$ for any $\delta > 0$.

4. Examples.

4.1. The linear quadratic regulator.

As an application, consider the completely observed stationary linear regulator with quadratic cost. For simplicity we let x_0 be a deterministic vector, i.e., $\mu = \delta_{x_0}$. Such assumption can be easily weakened.

$$dx_t = (Ax_t + Bu_t)\, dt + G\, dw_t, \tag{4.1}$$

$$J_T(u) = \int_0^T (x_t^{'}Qx_t + u_t^{'}Ru_t)\, dt, \tag{4.2}$$

where $x_t \in \mathbf{R}^n$, $u_t \in \mathbf{R}^m$, w is the standard p-dimensional Wiener process, A, B, G, Q, R are constant matrices of appropriate dimensions, Q, R are symmetric and positive definite, and the pair (A, B) is stabilizable. We recall that under the above assumptions, there exists the limit

$$\Pi = \lim_{T\to\infty} \Pi_T(t) \;\text{ for all } t \geq 0, \tag{4.3}$$

where $\Pi_T(t)$ is the solution to the Riccati differential equation on $[0, T]$

$$\frac{d}{dt}\Pi_T(t) + \Pi_T(t)A + A^{'}\Pi_T(t) - \Pi_T(t)BR^{-1}B^{'}\Pi_T(t) + Q = 0,\; \Pi_T(T) = 0, \tag{4.4}$$

and Π is the positive definite solution to the algebraic Riccati equation

$$\Pi A + A^{'}\Pi - \Pi BR^{-1}B^{'}\Pi + Q = 0. \tag{4.5}$$

Moreover, for some positive constants C, K

$$\|\Pi_T(t) - \Pi\| \leq Ce^{-K(T-t)}, \quad 0 \leq t \leq T, \tag{4.6}$$

and the matrix $A - BR^{-1}B^{'}\Pi$ is stable (see [1]).

We now show that Theorems 3.1 and 3.2 hold for this model, Assumption A being satisfied. The solutions to (HJB) and (SHJB) are given by, respectively,

$$V_T(t,x) = \phi_T(t,x) + \int_t^T \lambda_T(s)\,ds \tag{4.7}$$

with

$$\phi_T(t,x) = x^{'}\Pi_T(t)x, \quad \lambda_T(t) = \operatorname{tr}(G^{'}\Pi_T(t)G) \tag{4.8}$$

and

$$\phi(x) = x^{'}\Pi x, \quad \lambda = \operatorname{tr}(G^{'}\Pi G). \tag{4.9}$$

Moreover, the optimal feedback is

$$u_t^* = -R^{-1}B^{'}\Pi x_t, \tag{4.10}$$

which is easily seen to be admissible. Thus Assumptions A1 and A2 are satisfied. Assumption A3 is an immediate consequence of (4.6). Assumptions A4 and A5 are easily checked using the positivity of Q and the fact that

$$\frac{\partial V_T}{\partial x}(t,x)\sigma(x,u) = 2x^{'}\Pi_T(t), \qquad \frac{\partial \phi}{\partial x}(x)\sigma(x,u) = 2x^{'}\Pi.$$

Finally, by (4.10), denoting by x^* the process corresponding to the optimal control u^*, we have

$$c(x_t^*, u_t^*) = x_t^{*'}Qx_t^* + u_t^{*'}Ru_t^* = x_t^{*'}(Q + \Pi BR^{-1}B^{'}\Pi)x_t^* \leq C|x_t^*|^2. \tag{4.11}$$

We know that the matrix $A - BR^{-1}B^{'}\Pi$ is stable, so that the function $t \mapsto E|x_t^*|^2$ is bounded on $[0,\infty)$. As x_t^* is a Gaussian process it turns out that there exists a positive constant C_k such that

$$E|x_t^*|^k \leq C_k \ \text{ for all } t \geq 0, k \geq 2. \tag{4.12}$$

From (4.11), (4.12), and the inequality

$$\|\int_0^T \xi_t\,dt\|_m \leq \int_0^T \|\xi_t\|_m\,dt,$$

where $\|.\|_m$ is the $L^m(\Omega)$ norm, Assumption A6 follows for all $m \geq 1$.

4.2. Uniformly ergodic diffusions. In this section we consider a class of nonlinear controlled diffusions for which Assumption A is satisfied.

Consider diffusions satisfying the following stochastic differential equation:

$$\begin{array}{rcl} dx_t & = & (g(x_t) + u_t)dt + dw_t, \\ x_0 & \sim & \mu, \end{array}$$

where $x_t, u_t \in \mathbb{R}^d$. The cost functional is given by

$$J_T(u) = \int_0^T \left[l(x_t) + \frac{1}{2}\|u_t\|^2 \right] dt.$$

In other words, we are considering models as in section 3 with $U = \mathbb{R}^d$, $f(x,u) = g(x) + u$, $\sigma(x,u) = I$, and $c(x,u) = l(x) + \frac{1}{2}\|u\|^2$. We also make the following further assumptions on the model.

(i) The function $l(\cdot)$ is nonnegative, continuously differentiable, and its first partial derivatives are bounded.

(ii) The function $g(\cdot)$ is twice continuously differentiable, and its first partial derivatives are bounded.

(iii) There exists a constant $c > 0$ such that, for every $x, y \in \mathbb{R}^d$

$$(x-y)\cdot(g(x) - g(y)) \leq -c\|x-y\|^2,$$

where "$\cdot$" is the scalar product in $\mathbb{R}^d$.

PROPOSITION 4.1. *Under* (i), (ii), *and* (iii) *there exists a solution* $V_T \in \mathcal{C}^{1,2}$ *of* (HJB). *The first partial derivatives of* V_T *with respect to* x *are uniformly bounded in all variables. In particular, Assumptions* A1 *and* A4 *are satisfied. Moreover, the "inf" in* (HJB) *determines an admissible feedback control* u^T*; the corresponding process* x^T*, defined for times* $t \in [0,T]$*, is such that all its moments are uniformily bounded in both* t *and* T.

The existence of a regular solution V_T is proved in [9, Theorem 6.2]. The remaining part of Proposition 4.1 is proved in [8, Lemma 4.1 and Appendix A]. Note that in [9, Theorem 6.2], the control is supposed to take values a compact set. In this example, however, this is not a restriction since the "inf" in (HJB) is attained at the feedback $-\nabla_x V_T(t,x)$ that, as stated in Proposition 4.1, is bounded in all variables.

PROPOSITION 4.2. *There exists a solution* (ϕ, λ) *of* (SHJB) *with* $\phi \in \mathcal{C}^2$. *The first partial derivatives of* ϕ *are bounded. Moreover, the "inf" in* (SHJB) *determines an admissible feedback control* u^**; the corresponding process* x_t^* *is such that all its moments are bounded in time. Thus, in particular, Assumptions* A2 *and* A5 *hold, and* A6 *is satisfied for every* $m \geq 1$.

For the proof of Proposition 4.2 see [8, sections 3 and 4].

Thus, we have only to check Assumption A3. Define $W_T(t,x) = \phi(x) + \lambda(T-t)$. By the standard verification theorem of stochastic control (see, e.g., [9, Chapter 4, Theorem 4.1], we have, for all $x \in \mathbb{R}^d$,

$$\begin{aligned} V_T(0,x) &= \inf_{u \in \mathcal{U}} E \int_0^T c(x_t, u_t) dt, \\ W_T(0,x) = \phi(x) + \lambda T &= \inf_{u \in \mathcal{U}} E \left\{ \int_0^T c(x_t, u_t) dt + \phi(x_T) \right\} \\ &= E \left\{ \int_0^T c(x_t^*, u_t^*) dt + \phi(x_T^*) \right\}, \end{aligned}$$

where we have used the fact that the function $W_T(t,x)$ is a solution of (HJB) with final condition $W_T(T,x) = \phi(x)$. It follows that

$$W_T(0,x) \leq V_T(0,x) + E\phi(x_T^T), \tag{4.13}$$

where x^T is the process defined in Proposition 4.1. Note that, since ϕ has bounded gradient, it grows at most linearly. Therefore, by Proposition 4.1, the expectation $E\phi(x_T^T)$ is bounded in T, and Assumption A3 easily follows.

4.3. Diffusions with periodic coefficients. In this section we consider a class of nonlinear controlled diffusions for which Assumption B is satisfied.

We consider stochastic differential equations in $\mathbb{R}^d$ of the type

$$\begin{aligned} dx_t &= f(x_t, u_t)dt + dw_t, \\ x_0 &\sim \mu, \end{aligned}$$

where we assume (i) U is a compact metric space; (ii) f is jointly continuous in (x, u), and it is α-Holder continuous in x uniformly with respect to u, for some $\alpha > 0$; (iii) f is periodic in x. Moreover, the running cost function $c(x, u)$ is also assumed to be periodic in x, jointly continuous in (x, u), and α-Holder continuous in x uniformly with respect to u.

PROPOSITION 4.3. *There exists a solution of* (SHJB) *for the above models, such that $\phi \in \mathcal{C}^2(\mathbb{R}^d)$ and it is periodic.*

Proof. The existence of a periodic solution of (SHJB) in $W^{2,p}((0, \tau)^d)$ (τ is the period) for all $p \geq 2$ comes from [4, Theorem 6.1]. It follows that $\frac{\partial \phi}{\partial x} \in W^{1,p}$ for all large p, and therefore it is Holder continuous (see [6, Theorem IX.14]). Plugging this information into (SHJB), we get that $\Delta\phi$ is Holder continuous. Thus (see, e.g., [20, Theorem 4.1]) $\phi \in \mathcal{C}^2$. □

Note now that boundedness of ϕ follows from Proposition 4.3. Moreover, conditions A5 and A6 are trivially satisfied. As far as A2 is concerned, we have to show that the "inf" in (SHJB) is attained at an admissible feedback control. By continuity of f, c and compactness of U, the "inf" is attained at a feedback $k(x)$. Since the function $b(x) = f(x, k(x))$ is bounded, a weak solution of

$$\begin{aligned} dx_t &= b(x_t)dt + dw_t, \\ x_0 &\sim \mu \end{aligned}$$

can be constructed by a Girsanov transformation [10]. So $u_t^* = k(x_t)$ is admissible. In conclusion Assumption B is satisfied, which guarantees a.s. g-optimality of u^* for all g satisfying A7.

5. Proofs for diffusions.

Proof of Theorem 3.1. In this proof we let $(\Omega, \mathcal{F}, P)$ be any probability space in which stochastic processes $x_\cdot, x^*$, having law P^u and P^{u^*}, respectively, are defined. Without loss of generality (e.g., by using Theorem 4.2 in [10]), we may assume there are Brownian motions $w_\cdot, w_\cdot^*$ such that the equalities

$$\begin{aligned} dx_t &= f(x_t, u_t)dt + \sigma(x_t, u_t)dw_t, \\ dx_t^* &= f(x_t^*, u_t^*)dt + \sigma(x_t^*, u_t^*)dw_t^* \end{aligned}$$

hold in a strong sense.

Let $n = [T]$ be the integer part of $T > 0$. Then

$$\begin{aligned} g(T)[J_T(u^*) - J_T(u)] &= g(T)\left[\int_0^T c(x_t^*, u_t^*)\,dt - \int_0^T c(x_t, u_t)\,dt\right] \\ &\leq g(n)\left[\int_0^{n+1} c(x_t^*, u_t^*)\,dt - \int_0^n c(x_t, u_t)\,dt\right] \\ &= g(n)[J_{n+1}(u^*) - J_n(u)]. \end{aligned}$$

Thus it is sufficient to prove

$$\limsup_{n\to+\infty} g(n)[J_{n+1}(u^*) - J_n(u)] \leq 0 \quad \text{a.s.} \tag{5.1}$$

We first assume that Assumption A holds.

For $\varepsilon > 0$, we define $\Gamma_n = \{g(n)[J_{n+1}(u^*) - J_n(u)] \geq \varepsilon\}$, $\tau_n = \inf\{s : J_s(u) \geq h_n\} \wedge n$, and $h_n = nq(n)^{-1}$. We have

$$\begin{aligned} P(\Gamma_n, \tau_n < n) &= P(g(n)J_{n+1}(u^*) \geq g(n)J_n(u) + \varepsilon, \tau_n < n) \\ &\leq P(g(n)J_{n+1}(u^*) \geq g(n)h_n + \varepsilon) \\ &\leq P(J_{n+1}(u^*) \geq h_n). \end{aligned} \tag{5.2}$$

From the Chebyshev inequality, A6, and the definition of h_n

$$\begin{aligned} P(J_{n+1}(u^*) \geq h_n) &\leq h_n^{-m} E[J_{n+1}(u^*)]^m \\ &\leq C h_n^{-m}(n+1)^m = Cq(n)^m(1 + \tfrac{1}{n})^m, \end{aligned} \tag{5.3}$$

where we denoted by C different constants whose specific values are irrelevant. Condition A7 is equivalent to convergence of $\sum_{n=1}^{+\infty} q(n)^m$ and $\sum_{n=1}^{+\infty} q(n)^m(1 + 1/n)^m$. Thus from (5.2) and (5.3)

$$\sum_{n=1}^{+\infty} P(\Gamma_n, \tau_n < n) < +\infty, \tag{5.4}$$

which implies (5.1) on the set $\{\tau_n < n\}$.

We now deal with the set $\{\tau_n = n\}$. By applying Itô's rule to the solution of (HJB) we get

$$dV_T(t, x_t) = \left[\frac{\partial V_T}{\partial t}(t, x_t) + L^{u_t} V_T(t, x_t)\right] dt + \frac{\partial V_T}{\partial x}(t, x_t)\sigma(x_t, u_t) dw_t.$$

Using (HJB), adding and subtracting $c(x_t, u_t)dt$ to the previous equation, we get

$$J_T(u) = V_T(0, x_0) + \int_0^T \Delta^{V_T}(t, x_t, u_t) dt + M_T^{V_T}(u), \tag{5.5}$$

where

$$\Delta^{V_T}(t, x_t, u_t) = L^{u_t} V_T(t, x_t) + c(x_t, u_t) - \inf_{v \in U}[L^v V_T(t, x_t) + c(x_t, v)] \geq 0$$

and

$$M_T^{V_T}(u) = \int_0^T \frac{\partial V_T}{\partial x}(t, x_t)\sigma(x_t, u_t) dw_t$$

is a continuous local martingale. Similarly, by using the solution of (SHJB) rather than (HJB) we have

$$J_T(u^*) = \phi(x_0) + \lambda T - \phi(x_T^*) + M_T^\phi(u^*), \tag{5.6}$$

where

$$M_T^\phi(u^*) = \int_0^T \frac{\partial \phi}{\partial x}(t, x_t^*)\sigma(x_t^*, u_t^*) dw_t^*.$$

Thus, letting $W_T(t, x) = \phi(x) + \lambda(T - t)$, we obtain

$$g(n)[J_{n+1}(u^*) - J_n(u)] \leq g(n)[W_{n+1}(0,x_0) - V_n(0,x_0)] - g(n)\phi^-(x^*_{n+1})$$

$$+g(n)\int_0^{n+1} \frac{\partial W_{n+1}}{\partial x}(t,x^*_t)\sigma(x^*_t,u^*_t)\,dw^*_t - g(n)\int_0^n \frac{\partial V_n}{\partial x}(t,x_t)\sigma(x_t,u_t)\,dw_t. \tag{5.7}$$

It follows by Assumption A2 and the Borel–Cantelli lemma that

$$\lim_{n\to\infty} g(n)\phi^-(x^*_{n+1}) = 0$$

a.s., and by Assumption A3 that

$$\limsup_{n\to+\infty} g(n)[W_{n+1}(0,x_0) - V_n(0,x_0)] \leq 0 \quad \text{a.s.} \tag{5.8}$$

Now let Γ'_n denote the event where the expression in the second line of (5.7) takes a value greater that ε. Using first Chebyshev and then Burkholder inequality [17], we have

$$\begin{aligned} P(\Gamma'_n, \tau_n = n) &\leq P\left(g(n)\int_0^{n+1} \frac{\partial W_{n+1}}{\partial x}(t,x^*_t)\sigma(x^*_t,u^*_t)\,dw^*_t \geq \frac{\varepsilon}{3}\right) \\ &+P\left(-g(n)\int_0^n \frac{\partial V_n}{\partial x}(t,x_t)\sigma(x_t,u_t)\,dw_t \geq \frac{\varepsilon}{3}, \tau_n = n\right) \\ &\leq Cg(n)^{2m}E\left[\int_0^{n+1}\left|\frac{\partial W_{n+1}}{\partial x}(t,x^*_t)\sigma(x^*_t,u^*_t)\right|^2 dt\right]^m \\ &+Cg(n)^{2m}E\left[\int_0^{\tau_n}\left|\frac{\partial V_n}{\partial x}(t,x_t)\sigma(x_t,u_t)\right|^2 dt\right]^m. \end{aligned} \tag{5.9}$$

Now using Assumptions A5, A6, and the inequality $(a+b)^m \leq 2^m(a^m+b^m)$ we have

$$\begin{aligned} &E\left[\int_0^{n+1}\left|\frac{\partial W_{n+1}}{\partial x}(t,x^*_t)\sigma(x^*_t,u^*_t)\right|^2 dt\right]^m \\ &\leq B2^m\left[(n+1)^m + E\left[\int_0^{n+1} c(x^*_t,u^*_t)\,dt\right]^m\right] \leq B'(n+1)^m. \end{aligned} \tag{5.10}$$

Moreover, by Assumption A4 and the definition of τ_n,

$$\int_0^{\tau_n}\left|\frac{\partial V_n}{\partial x}(t,x_t)\sigma(x_t,u_t)\right|^2 dt \leq B[J_{\tau_n}(u)+n] \leq B[h_n+n]. \tag{5.11}$$

Thus, by (5.9)–(5.11), we obtain, for some constant A

$$P(\Gamma'_n, \tau_n = n) \leq Ag(n)^{2m}[(n+1)^m + (h_n+n)^m] = A\left[q_n^{2m}\left(1+\frac{1}{n}\right)^m + (q_n+q_n^2)^m\right], \tag{5.12}$$

which implies

$$\sum_{n=1}^{+\infty} P(\Gamma'_n, \tau_n = n) < +\infty.$$

This inequality, together with the Borel–Cantelli lemma, (5.4), and (5.8) completes the proof under Assumption A.

Under Assumption B the proof needs a slight modification. In particular, the proof of (5.1) on $\{\tau_n < n\}$ is unchanged. For the case $\{\tau_n = n\}$ we use Itô's rule for both $\phi(x_t)$ and $\phi(x_t^*)$, obtaining

$$g(n)[J_{n+1}(u^*) - J_n(u)] \leq g(n)[\lambda + \phi(x_n) - \phi(x_{n+1}^*)]$$

$$+g(n)\int_0^{n+1} \frac{\partial W_{n+1}}{\partial x}(t, x_t^*)\sigma(x_t^*, u_t^*)\, dw_t^* - g(n)\int_0^n \frac{\partial W_n}{\partial x}(t, x_t)\sigma(x_t, u_t)\, dw_t.$$

Bounds for the martingale terms are obtained as before, while $g(n)[\lambda+\phi(x_n)-\phi(x_{n+1}^*)]$ goes to zero, since ϕ is bounded. □

Proof of Theorem 3.2. Just note that, under the hypothesis of Theorem 3.2, the expressions in (5.3) and (5.12) are infinitesimal, although possibly not summable. □

6. Pathwise optimality for multivariate point processes. In this section we deal with stochastic processes taking value in $\mathbb{N}^d$, $\mathbb{N}$ being the set of positive integers, and generated by the family of operators

$$L^v = \sum_{i=1}^d \lambda^{(i)}(x, v)\nabla_i \tag{6.1}$$

with $x \in \mathbb{N}^d$; here $\nabla_i f(x) = f(x+e_i) - f(x)$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, the "1" being in the ith component. The functions $\lambda^{(i)}$ are assumed to be measurable and nonnegative.

Suppose u is an admissible control. Here P^u is meant to be defined on $D = D([0, +\infty), \mathbb{N}^d) \subset D([0, +\infty), \mathbb{R}^d)$. It is well known that, under P^u, the processes

$$M_t^{(i)} = x_t^{(i)} - \int_0^t \lambda^{(i)}(x_s, u_s)ds, \quad i = 1, \ldots, d, \tag{6.2}$$

are orthogonal local martingales. In (6.2), x_t denotes the canonical process.

In this context, the equations (HJB) and (SHJB) have the same form they had for diffusions, once the operator L^v is replaced by the one in (6.1).

As for diffusions, we now state two sets of assumptions.

Assumption A.

A1. For all $T > 0$ there exists a solution V_T of (HJB) that is continuously differentiable in t.

A2. There exists a solution (ϕ, λ) of (SHJB). The "inf" in (SHJB) is attained at $v = k(x)$, and the feedback $u_t^* = k(x_t)$ is an admissible control. Moreover, there exist constants $C, \epsilon > 0$ for which

$$E^{P^{u^*}}\left[\left(\phi^-(x_t)\right)^{2+\epsilon}\right] < C.$$

A3. The following inequality holds:

$$\limsup_{T\to+\infty} g(T)\left[\phi(x_0) + \lambda T - V_T(0, x_0)\right] \leq 0 \qquad \mu\text{-a.s.}$$

Moreover, there is an integer $m > 0$ such that the following further assumptions hold.

A4. There is a constant $B > 0$ such that for all $1 \le r \le 2^m$ and for all $i = 1, \dots, d$

$$|\nabla_i V_T(t,x)|^r \lambda^{(i)}(x,u) \le B(c(x,u)+1). \tag{6.3}$$

(Note that it is enough to check (6.3) for $r = 1$ and $r = 2^m$.)

A5. As A4, with ϕ replacing V_T.

A6. There is a constant $C > 0$ such that

$$E\left\{\left[\int_0^T c(x_t^*, u_t^*)dt\right]^{2^{m-1}}\right\} \le CT^{2^{m-1}}$$

for all $T > 0$.

A7. Define $q(t) = t^{1/2}g(t)$. Then

$$\int_1^{+\infty} q^{2^{m-1}}(t)dt < +\infty.$$

As we shall see later, the reason why these assumptions are slightly more complicated than the ones in section 3 is that it is harder to get good estimates for discontinuous martingales than for continuous ones.

Assumption B. There is a solution (ϕ, λ) of (SHJB), with ϕ bounded. Moreover, Assumptions A2, A5, A6, and A7 hold.

Under either Assumption A or B Theorems 3.1 and 3.2 hold true. Their proof, in the context of point processes, is given in the following section.

7. Proofs for point processes. Theorems 3.1 and 3.2 for point processes are proved using the same argument used for diffusions. We sketch here the proof, stressing the only important modification that is needed. For simplicity of notation we write P for P^{u,u^*}.

We again start by defining $h_n = nq^{-1}(n)$ and $\tau_n = \inf\{s : J_s(u) \ge h_n\} \wedge n$. The proof of

$$\limsup_{n\to\infty} g(n)[J_{n+1}(u^*) - J_n(u)] \le 0 \quad \text{a.s.} \tag{7.1}$$

on the set $\{\tau_n < n\}$ is identical to the one in section 5. For the case $\{\tau_n = n\}$ we need to obtain representations for the cost function J_T that correspond to (5.5) and (5.6). Observe that

$$\begin{aligned} V_T(T,x_T) - V_T(0,x_0) &= -V_T(0,x_0) \\ &= \sum_{0<t\le T}[V_T(t,x_t) - V_T(t,x_{t-})] + \int_0^T \frac{\partial V_T}{\partial t}(t,x_t)dt \\ &= \sum_{i=1}^d \int_0^T \nabla_i V_T(t,x_{t-})dx_t^{(i)} + \int_0^T \frac{\partial V_T}{\partial t}(t,x_t)dt \\ &= \int_0^T \Big(\frac{\partial V_T}{\partial t} + L^{u_t}V_T\Big)(t,x_t)dt + M_T^{V_T}(u), \end{aligned}$$

where

$$M_T^{V_T}(u) = \sum_{i=1}^d \int_0^T \nabla_i V_T(t,x_{t-})dM_t^{(i)},$$

with $M_t^{(i)}$ being the local martingale in (6.2). Thus, similarly to (5.5),

$$J_T(u) = V_T(0, x_0) + \int_0^T \Delta^{V_T}(t, x_t, u_t)dt + M_T^{V_T}(u), \tag{7.2}$$

where

$$\Delta^{V_T}(t, x_t, u_t) = L^{u_t}V_T(t, x_t) + c(x_t, u_t) - \inf_{v \in U}[L^v V_T(t, x_t) + c(x_t, v)] \geq 0.$$

A similar representation for the cost is obtained if one uses the solution of the (SHJB). The key to the proof of Theorem 3.1 is to show that for every $\epsilon > 0$, under Assumption A,

$$\sum_n P\{g(n)|M_n^{V_n}(u)| > \epsilon, \tau_n = n\} < +\infty \tag{7.3}$$

together with the analogous estimate for the martingale with ϕ replacing V_T and u^* replacing u. We prove only (7.3); the rest of the proof is identical to the diffusion case.

We begin by noting that the quadratic variation of the local martingale $M_T^{V_T}(u)$ is given by (see [17, Theorem 22])

$$[M_T^{V_T}(u), M_T^{V_T}(u)] = \sum_{i=1}^d \int_0^T [\nabla_i V_T(t, x_{t-})]^2 dx_t^{(i)}.$$

It follows, by Burkholder inequality, that

$$\begin{aligned} E\left\{|M_n^{V_{\tau_n}}(u)|^{2^m}\right\} &\leq C\sum_{i=1}^d E\left\{\left[\int_0^{\tau_n}[\nabla_i V_n(t, x_{t-})]^2 dx_t^{(i)}\right]^{2^{m-1}}\right\} \\ &\leq C'\sum_{i=1}^d \left(E\left\{\left[\int_0^{\tau_n}[\nabla_i V_{\tau_n}(t, x_{t-})]^2 dM_t^{(i)}\right]^{2^{m-1}}\right\}\right. \\ &\qquad \left. + E\left\{\left[\int_0^{\tau_n}[\nabla_i V_{\tau_n}(t, x_t)]^2 \lambda^{(i)}(x_t, u_t)dt\right]^{2^{m-1}}\right\}\right) \end{aligned}$$

for some constants C, C' depending on m. Now we can apply again Burkholder inequality to the local martingale

$$\int_0^{\tau_n} [\nabla_i V_n(t, x_{t-})]^2 dM_t^{(i)}$$

in order to bound the term

$$E\left\{\left[\int_0^{\tau_n}[\nabla_i V_n(t, x_{t-})]^2 dM_t^{(i)}\right]^{2^{m-1}}\right\}.$$

Iterating this procedure and noting that, for the last iteration

$$E\left\{\int_0^{\tau_n}[\nabla_i V_n(t, x_{t-})]^{2^m} dx_t^{(i)}\right\} = E\left\{\int_0^{\tau_n}[\nabla_i V_n(t, x_{t-})]^{2^m} \lambda^{(i)}(x_t, u_t)dt\right\},$$

we obtain the following estimate:

$$E\Big\{|M^{V_n}_{\tau_n}(u)|^{2^m}\Big\} \le C\sum_{i=1}^{d}\sum_{r=1}^{m} E\left\{\left[\int_0^{\tau_n}[\nabla_i V_{\tau_n}(t,x_{t-})]^{2^r}\lambda^{(i)}(x_t,u_t)dt\right]^{2^{m-r}}\right\}. \tag{7.4}$$

It follows, using Chebyshev's inequality, Assumption A4, and the definition of τ_n that

$$P\{g(n)|M^{V_n}_n(u)| > \epsilon, \tau_n = n\} \le Cg^{2^m}(n)E\Big\{|M^{V_n}_{\tau_n}(u)|^{2^m}\Big\}$$

$$\le Cg^{2^m}(n)\sum_{i=1}^{d}\sum_{r=1}^{m} B[J_{\tau_n}(u)+n]^{2^{m-r}} \le C'g^{2^m}(n)\sum_{r=1}^{m}[h_n+n]^{2^{m-r}}.$$

The summability over n of this last expression is straightforward. The proof of (7.3) is therefore completed.

The modifications needed for the proof of Theorem 3.2 are identical to the case of diffusions. □

8. Pathwise optimality for Markov chains with finite state space. Let X be a finite set and assume U is a compact metric space. For $x, y \in X$, with $x \neq y$, let

$$l_{x,y} : U \to \mathbb{R}^+$$

be strictly positive functions. The X-valued controlled processes we consider in this section correspond to the family of operators L^v, $v \in U$, defined by

$$L^v f(x) = \sum_{y\neq x} l_{x,y}(v)[f(y) - f(x)]. \tag{8.1}$$

To complete the assumptions on the model we assume that the running cost function $c(x, v)$ is continuous in v.

LEMMA 8.1. *There exists a solution* (ϕ, λ) *of* (SHJB).

We have not been able to find a proper reference for Lemma 8.1. Its proof, obtained by adapting that of [4, Theorem 6.1], is given in the appendix. Note that, due to the finiteness of X, compactness of U, and continuity of $l_{x,y}(\cdot), c(x,\cdot)$, there is a feedback $k(x)$ that minimizes the "inf" in (SHJB); moreover, the associated feedback control is admissible.

We can now prove the main result of this section.

THEOREM 8.2. *The control* $u^*(x_t)$ *is g-optimal a.s. for every g satisfying Assumption* A7*, and it is g-optimal in probability if* $g(t)t^{1/2} = o(1)$.

Proof. The proof is a simple modification of the one for point processes. As for both diffusions and point processes, the proof starts by defining the stopping time τ_n, and showing that

$$\limsup_{n\to\infty} g(n)[J_{n+1}(u^*) - J_n(u)] \le 0 \tag{8.2}$$

on the set $\{\tau_n < n\}$, a.s. or in probability according to which assumptions one chooses. For the case $\tau_n = n$ we obtain a representation of the cost function in terms of the solution of (SHJB).

For $y \in X$, let $N_t(y)$ be the process that counts the number of jumps to the state y. It is known that

$$M_t(y) = N_t(y) - \int_0^t l_{x_s,y}(u_s)ds$$

is a P^u-martingale, with quadratic variation $N_t(y)$. Thus, if (ϕ, λ) is the solution of (SHJB), we get

$$\phi(x_T) - \phi(x_0) = \sum_{0 \leq t \leq T} [\phi(x_t) - \phi(x_{t-})]$$

$$= \int_0^T \sum_{y \neq x_{t-}} [\phi(y) - \phi(x_{t-})]dN_t(y) = \int_0^T L^{u_t}\phi(x_t)dt + \int_0^T \sum_{y \neq x_{t-}} [\phi(y) - \phi(x_{t-})]dM_t(y).$$

It follows that

$$g(n)[J_{n+1}(u^*) - J_n(u)] \leq g(n)[\lambda + \phi(x_n)]$$

$$+g(n) \int_0^{n+1} \sum_{y \neq x_{t-}} [\phi(y) - \phi(x^*_{t-})]dM^*_t(y) - g(n) \int_0^n \sum_{y \neq x_{t-}} [\phi(y) - \phi(x_{t-})]dM_t(y). \tag{8.3}$$

Expression (8.3) is estimated as in the proof for point processes. It is clear that here $\phi(x)$ and $l_{x,y}(u)$ are bounded, which makes the proof work as the one for point processes under Assumption B. □

Appendix A.

Proof of Lemma 8.1.

For $\alpha > 0$ define the function

$$V_\alpha(x) = \inf_{u \in \mathcal{U}} E^u_x \left\{ \int_0^{+\infty} e^{-\alpha t} c(x_t, u_t)dt \right\}, \tag{A.1}$$

where E^u_x is the expectation with respect to P^u with initial condition $P^u \circ \Pi_0^{-1} = \delta_x$. It is easy to see that V_α can also be rewritten as

$$V_\alpha(x) = \inf_{u \in \mathcal{U}} E^u_x \left\{ \int_0^T e^{-\alpha t} c(x_t, u_t)dt + e^{-\alpha T} V_\alpha(x_T) \right\}$$

for any $T > 0$. Now define

$$\begin{aligned} V_{\alpha,T}(x,t) = & \inf_{u \in \mathcal{U}} E^u_x \left\{ \int_t^T e^{-\alpha s} c(x_s, u_s)ds + e^{-\alpha T} V_\alpha(x_T) \Big| x_t = x \right\} \\ = & \; e^{-\alpha t} V_\alpha(x). \end{aligned} \tag{A.2}$$

Clearly $V_{\alpha,T}$ is the value function of a stochastic control problem in the finite time horizon $[0, T]$. Thus it satisfies the (HJB) equation

$$\frac{\partial V_{\alpha,T}}{\partial t}(x,t) + \inf_{v \in U} \left[L^v V_{\alpha,T}(x,t) + e^{-\alpha t} c(x,v) \right] = 0$$

that, by (A.2), is equivalent to

$$\inf_{v \in U} \Big[L^v V_\alpha(x) + c(x,v) \Big] = \alpha V_\alpha(x). \tag{A.3}$$

Now let

$$\bar{V_\alpha} = \frac{1}{|X|} \sum_{x \in X} V_\alpha(x),$$

$$\phi_\alpha(x) = V_\alpha(x) - \bar{V_\alpha}.$$

Equation (A.3) can be rewritten in terms of ϕ_α as

$$\inf_{v \in U} \left[\sum_{y \neq x} l_{x,y}(v) \big(\phi_\alpha(y) - \phi_\alpha(x) \big) + c(x,v) \right] = \alpha \phi_\alpha(x) + \alpha \bar{V_\alpha}. \tag{A.4}$$

Let $k_\alpha(x)$ be the feedback that realizes the "inf" in (A.4) (or (A.3)), and let m_α be the invariant measure of the corresponding optimal process. The invariant measure satisfies the equation

$$\sum_{y \neq x} m_\alpha(x) l_{x,y}(k_\alpha(x)) = \sum_{y \neq x} m_\alpha(y) l_{y,x}(k_\alpha(y)) \tag{A.5}$$

from which one easily obtains

$$0 < A \leq m_\alpha(x) \leq B < +\infty \quad \text{for all } \alpha > 0, x \in X, \tag{A.6}$$

where

$$B = \frac{|X| \sup\{l_{x,y}(v) : x,y \in X, v \in U\}}{\inf\{l_{x,y}(v) : x,y \in X, v \in U\}},$$

$$A = \frac{\inf\{l_{x,y}(v) : x,y \in X, v \in U\} \wedge 1}{1 + |X| \sup\{l_{x,y}(v) : x,y \in X, v \in U\}}.$$

In particular, A and B do not depend on α.

Now consider (A.3). After multiplying it by $m_\alpha(x) V_\alpha(x)$, summing over x, and using (A.5), we obtain

$$\begin{aligned} &\frac{1}{2} \sum_x \sum_{y \neq x} m_\alpha(x) l_{x,y}(k_\alpha(x)) [V_\alpha(y) - V_\alpha(x)]^2 \\ &\quad = \sum_x m_\alpha(x) c(x, k_\alpha(x)) V_\alpha(x) - \alpha \sum_x m_\alpha(x) V_\alpha^2(x). \end{aligned} \tag{A.7}$$

On the other hand, if we multiply (A.3) by $m_\alpha(x)$ and sum over x, we get

$$0 = \alpha \sum_x m_\alpha(x) V_\alpha(x) - \sum_x m_\alpha(x) c(x, k_\alpha(x)). \tag{A.8}$$

By (A.7) and (A.8) we obtain

$$\begin{aligned} &\frac{1}{2} \sum_x \sum_{y \neq x} m_\alpha(x) l_{x,y}(k_\alpha(x)) [\phi_\alpha(y) - \phi_\alpha(x)]^2 \\ &\quad = \sum_x m_\alpha(x) c(x, k_\alpha(x)) \phi_\alpha(x) - \alpha \sum_x m_\alpha(x) V_\alpha(x) \phi_\alpha(x). \end{aligned} \tag{A.9}$$

Note now that, by (A.1),

$$\alpha V_\alpha(x) \le \sup_{x,v} c(x,v) < +\infty. \tag{A.10}$$

Thus (A.6), (A.9), and (A.10) yield

$$|\phi_\alpha(y) - \phi_\alpha(x)|^2 \le C \sup_z |\phi_\alpha(z)| \quad \text{for all } x, y \in X \tag{A.11}$$

for some constants $C > 0$. Inequality (A.11) implies that ϕ_α is uniformly bounded in α. Otherwise, for any $M > 0$ there is α, x such that $|\phi_\alpha(x)| > M$. Combining this with (A.11), we get

$$\left| \sum_y \phi_\alpha(y) \right| \ge M - |X|\sqrt{CM}.$$

It cannot be so for M large, since $\sum_y \phi_\alpha(y) = 0$.

Boundedness of both ϕ_α and αV_α guarantees that along some sequence $\alpha_n \to 0$

$$\phi_{\alpha_n} \to \phi, \quad \alpha_n \bar{V}_{\alpha_n} \to \lambda.$$

We may now pass to the limit as $\alpha_n \to 0$ in (A.4), and we see that (ϕ, λ) is a solution to (SHJB).

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[2] A. V. Asriev and V. I. Rotar, *On asymptotic optimality in probability and almost surely in dynamic control*, Stochastics Stochastic Rep., 33 (1990), pp. 1–16.

[3] T. A. Belkina, Y. M. Kabanov, and E. L. Presman, *Stochastic Linear Quadratic Regulator. Optimality Almost Sure and in Probability*, preprint.

[4] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, Gauthier Villars, New York, Paris, 1988.

[5] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1999.

[6] H. Brezis, *Analyse fonctionnelle, theorie et applications*, Masson, Paris, 1983.

[7] G. B. Di Masi and Y. M. Kabanov, *On sensitive probabilistic criteria in the linear regulator problem with the infinite horizon*, in Stochastic Processes and Optimal Control, A. A. Novikov, ed., TVP Publishing Company, Moscow, 1994.

[8] W. H. Fleming and W. M. McEneaney, *Risk-sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[9] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[10] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[11] T. A. Konyuhova and V. I. Rotar, *On optimal in probability solutions for a linear stochastic model*, in Probability Problems of Discrete Mathematics, Moscow Institute of Electronic Engineering, Moscow, 1987, pp. 131–135.

[12] T. A. Konyuhova and V. I. Rotar, *Optimal in probability stochastic control in a linear model with quadratic cost*, in Probability Problems of Discrete Mathematics, Moscow Institute of Electronic Engineering, Moscow, 1988, pp. 111–114.

[13] A. Leizarowitz, *On almost sure optimization for stochastic control systems*, Stochastics, 23 (1998), pp. 85–107.

[14] N. Y. Petrakov and V. I. Rotar, *Factor of the Uncertainty and the Control of Economical Systems*, Nauka, Moscow, 1985.

[15] È. L. Presman, *Optimality almost surely and in probability for a stochastic linear-quadratic regulator*, Theory Probab. Appl., 42 (1997), pp. 531–535.

[16] È. L. Presman, V. I. Rotar, and M. Taksar, *Optimality in probability and almost surely. The general scheme and a linear quadratic regulator problem*, Stochastics Stochastic Rep., 43 (1993), pp. 127–137.

[17] P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, Berlin, 1990.

[18] V. I. Rotar, *Some remarks about asymptotic optimality*, in Investigation of Stochastic Problems of Controlling Economical Systems, CEMI AN USSR, Moscow, 1996, pp. 93–116.

[19] V. I. Rotar, *Connectivity property and optimality almost surely and in probability*, in New Trends in Probability and Statistics, VSP, Utrecht, 1991, pp. 528–539.

[20] N. Shimakura, *Partial Differential Operators of Elliptic Type*, AMS, Providence, RI, 1992.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1558–1584

EXISTENCE AND CHARACTERIZATION OF AN OPTIMAL CONTROL FOR THE PROBLEM OF LONG WAVES IN A SHALLOW-WATER MODEL[*]

AZIZ BELMILOUDI[†]

Abstract. In this paper we present a method of optimal control developed in order to calculate the current corresponding to the observed sea level in a fluid domain Ω and during a time T. The control is the external stress $\underline{f}$. The cost function measures the distance between the observed and computed sea levels. The equations satisfied by the depth and the depth averaged velocity are of nonlinear shallow-water type. The existence and uniqueness of a solution for the direct problem are studied in the case of Dirichlet nonhomogeneous boundary conditions. We prove, by means of minimizing sequences, the existence of an optimal control $(\underline{f}, \underline{u})$ in the case of the small data and a very viscous fluid. To characterize it we build a sequence of problems corresponding to a linearization of the direct problem. We obtain the necessary conditions of optimality. The set of equations and the inequality characterizing the optimal control $(\underline{f}, \underline{u})$ is obtained as the limit of the penalization.

Key words. optimal control, shallow-water equations, nonlinear partial differential equations, minimizing sequences, penalization, altimetric measurements, oceanography

AMS subject classifications. 35Q30, 49J20, 49K20, 65J10, 76B15, 76D05, 76U05, 86A22

PII. S0363012999335961

1. Introduction and setting of the problem.

The mathematical method developed in this paper allows us to obtain the circulation in an oceanic domain Ω from satellite measurements which are surface observations. Altimetric measurements give the distance between the satellite and the sea surface. It is now possible to extract from these data the sea level topography with a precision in the order of centimeters. The sea level will act as the observation in our control model.

The phenomenon we are studying derives from the modelization of long waves in shallow-water zones (lochs, lagoons, etc.). In such regions currents are weak but play an important role in the biological and ecological equilibrium. The equations of motion are of Navier–Stokes type. Shallow-water domains are characterized by a small ratio between vertical and horizontal length scales. Then the friction on the bottom is important. Some assumptions of the model are imposed by physical features, such as the Boussinesq approximation: density variations are neglected except in the terms of gravity acceleration. Moreover, we assume that the pressure is hydrostatic, which is justified by the difference between horizontal and vertical scalings: the vertical component of the Navier–Stokes equations is simplified in order to express the balance between the gravity term and the vertical gradient of pressure.

The shallow-water equations are obtained by integrating the Navier–Stokes equations with respect to depth [5]. Moreover, we suppose that the velocity is small enough that $\frac{\partial \underline{\tilde{u}}}{\partial t} \gg (\underline{\tilde{u}}\nabla)\underline{\tilde{u}}$. This assumption is justified for shallow-water domains (see, for example, Pedlosky [17, pp. 67–69]) where currents remain weak, but it presents great variations with time. It allows one to neglect the advection term. Then the equations

[*] Received by the editors February 8, 1999; accepted for publication (in revised form) July 1, 2000; published electronically January 31, 2001.

http://www.siam.org/journals/sicon/39-5/33596.html

[†] IRMAR-Université Rennes1, LANS-INSA de Rennes, 20 avenue des Buttes de Coësmes, CS 14315, 35043 Rennes Cédex, France (Aziz.Belmiloudi@insa-rennes.fr).

verified by the depth averaged motion are the following:

$$\begin{aligned}&\frac{\partial \underline{\tilde{u}}}{\partial t} - \nu_1 \Delta \underline{\tilde{u}} + \frac{C}{D} \mid \underline{\tilde{u}} \mid_2 \underline{\tilde{u}} + \underline{F} \wedge \underline{\tilde{u}} + g\nabla \tilde{h} = \underline{\tilde{f}} \text{ on } \Omega,\\ &\frac{\partial \tilde{h}}{\partial t} - \nu_2 \Delta \tilde{h} + \operatorname{div}(\tilde{h}\underline{\tilde{u}}) = f_0 \text{ on } \Omega,\end{aligned} \tag{1.1}$$

where $\underline{\tilde{u}}: \ \Omega\times]0,T[\longrightarrow \mathbb{R}^2$ is the velocity and $\tilde{h}: \ \Omega\times]0,T[\longrightarrow \mathbb{R}$ is the depth of the studied layer and $\mid . \mid_2$ denotes the euclidean norm in $\mathbb{R}^2$.

The domain $\Omega \in \mathbb{R}^2$ is the projection of the oceanic domain on the horizontal plane. It corresponds to the undisturbed sea surface. Γ denotes its boundary. $\underline{F}$ is the Coriolis force defined by $(0, 0, 2\omega\text{sin}(\phi))$, where ω is the rotation rate of the earth and ϕ the latitude. g denotes the acceleration of gravity. The dissipative term corresponds to Reynolds stresses. ν_1 and ν_2 are eddy viscosity and diffusivity coefficients. C is the Chezy positive constant, D the averaged depth of the layer. The right-hand terms $\underline{\tilde{f}}$ and f_0 represent, respectively, the outside stress and the fluid exchanges (rain, evaporation, etc.). The shear effect on the bottom is represented by the term $\frac{C}{D} \mid \underline{\tilde{u}} \mid_2 \underline{\tilde{u}}$.

The total depth $\tilde{h}$ can be considered as the sum of two terms: the bottom topography $H(x, y)$ which is given and the topography $\xi(x, y, t)$ of the free sea surface. In order to solve (1.1), we have to set boundary and initial conditions:

$$\begin{aligned}&(\underline{\tilde{u}}, \tilde{h}) = (\underline{\tilde{u}}_B, \tilde{h}_B) \text{ on } \Gamma,\\ &(\underline{\tilde{u}}, \tilde{h})(t = 0) = (\underline{\tilde{u}}_0, \tilde{h}_0) \text{ in } \Omega.\end{aligned} \tag{1.2}$$

We are first going to prove that problem (1.1)–(1.2) is equivalent to another nonlinear problem with homogeneous boundary conditions.

Let $\underline{u}_L$ be the solution of the linear problem

$$\begin{aligned}&\frac{\partial \underline{u}_L}{\partial t} - \nu_1 \Delta \underline{u}_L = 0 \text{ in } \Omega,\\ &\underline{u}_L(t = 0) = \underline{u}_1 \text{ in } \Omega,\\ &\underline{u}_L = \underline{\tilde{u}}_B \text{ on } \Gamma\end{aligned} \tag{1.3}$$

and let h_L be the solution of the linear problem

$$\begin{aligned}&\frac{\partial h_L}{\partial t} - \nu_2 \Delta h_L = 0 \text{ in } \Omega,\\ &h_L(t = 0) = h_1 \text{ in } \Omega,\\ &h_L = \tilde{h}_B \text{ on } \Gamma,\end{aligned} \tag{1.4}$$

where $\underline{u}_1 \in H^1(\Omega) \cap L^\infty(\Omega)$, $\underline{\tilde{u}}_B \in L^2(0, T, H^{1/2}(\Gamma)) \cap L^\infty(0, T, L^\infty(\Gamma))$, $h_1 \in H^1(\Omega) \cap L^\infty(\Omega)$, and $\tilde{h}_B \in L^2(0, T, H^{1/2}(\Gamma)) \cap L^\infty(0, T, L^\infty(\Gamma))$.

REMARK 1. (i) *We can choose any* $(\underline{u}_1, h_1) \in H^1(\Omega) \cap L^\infty(\Omega)$. *For example, we can take* $\underline{u}_1 = \underline{\tilde{u}}_0$ *and* $h_1 = \tilde{h}_0$. $(\underline{u}_L, h_L)$ *is a lifting of the boundary conditions.*

(ii) *According to Fabre, Puel, and Zuazua* [7] *the function* $(\underline{u}_L, h_L)$ *solution of the problems* (1.3), (1.4) *will thus satisfy* $(\underline{u}_L, h_L) \in L^2(0, T, (H^1(\Omega))^3) \cap L^\infty(0, T, (L^\infty(\Omega))^3)$. □

Setting $\underline{u} = \underline{\tilde{u}} - \underline{u}_L$, $h = \tilde{h} - h_L$, $f_1 = f_0 - \operatorname{div}(h_L \underline{u}_L)$, $\underline{f} = \underline{\tilde{f}} - g\nabla h_L - \underline{F} \wedge \underline{u}_L$, problem (1.1)–(1.2) can be rewritten as follows:

$$
\begin{aligned}
&\textit{Find } (\underline{u}, h) \textit{ such that}\\
&\frac{\partial \underline{u}}{\partial t} - \nu_1 \Delta \underline{u} + \frac{C}{D} \mid (\underline{u} + \underline{u}_L) \mid_2 (\underline{u} + \underline{u}_L) + \underline{F} \wedge \underline{u} + g \nabla h = \underline{f} \textit{ in } \Omega,\\
&\frac{\partial h}{\partial t} - \nu_2 \Delta h + \operatorname{div}(h\underline{u}) + \operatorname{div}(h_L \underline{u}) + \operatorname{div}(h \underline{u}_L) = f_1 \textit{ in } \Omega,\\
&(\underline{u}, h) = (\underline{0}, 0) \textit{ on } \Gamma,\\
&(\underline{u}, h)(t = 0) = (\underline{u}_0, h_0) \textit{ in } \Omega,
\end{aligned}
\tag{1.5}
$$

where $(\underline{u}_0, h_0) = (\tilde{\underline{u}}_0 - \underline{u}_1, \tilde{h}_0 - h_1)$.

It has to be noted that the two linear problems (1.3), (1.4) do not depend on the forcing $\tilde{\underline{f}}$ but only on the boundary conditions $(\tilde{\underline{u}}_B, \tilde{h}_B)$. These two problems being solved, problem (1.1)–(1.2) is equivalent to the homogeneous problem (1.5).

Our purpose is to develop a control method in order to compute the depth averaged velocity $\tilde{\underline{u}}$. The surface topography $\xi(x, y, t)$ can be deduced from altimetric measurements. Therefore $\tilde{h} = H(x, y) + \xi(x, y, t)$ can be taken as the observation. We assume that the fluid exchanges f_0 are known; the model is then controlled by the external forcing $\tilde{\underline{f}}$.

The previous remark concerning the equivalence between problem (1.1)–(1.2) and problem (1.5) makes it possible to apply the control method to problem (1.5) instead of problem (1.1)–(1.2). Obviously we first have to solve problems (1.3) and (1.4) in order to obtain $(\underline{u}_L, h_L)$. Then $\underline{f}$ acts as the control and the observation is the depth h_{obs}. The optimal control is the forcing minimizing a cost function which measures the distance between the computed depth h and the observation h_{obs}. Precisely we will study the following optimal control problem: find $(\underline{u}, h, \underline{f})$ such that the cost function

$$
J(\underline{f}, \underline{u}) = \frac{1}{4} \parallel h - h_{obs} \parallel^4_{L^4(0,T,L^4(\Omega))} + \frac{\gamma}{2} \parallel \underline{f} \parallel^2_{L^2(0,T,L^2(\Omega))} \qquad (\mathcal{P})
$$

is minimized subject to the problem (1.5), with $\underline{f} \in K_c$, K_c (given) being a convex, closed, nonempty subset of $L^2(0, T, L^2(\Omega))$.

This paper is organized as follows: Section 2 is devoted to the study of problem (1.5). We give sufficient conditions on the surface forcing $\underline{f}$, on the initial situation $(\underline{u}_0, h_0)$, and on $(\underline{u}_L, h_L)$ in order to prove the existence of a solution. These conditions being satisfied, we prove the uniqueness of the solution.

Section 3 is devoted to the control problem associated with (1.5). We first deal with the homogeneous case $\tilde{\underline{u}}_B = \underline{0}$, $\tilde{h}_B = 0$. Then we have $\underline{u}_L = \underline{0}$, $h_L = 0$; problem (1.1)–(1.2) and problem (1.5) are identical. We prove the existence of an optimal control by means of minimizing sequences $(\underline{f}_k, \underline{u}_k)$. In order to characterize this optimal control, we have to introduce a penalized control problem $\mathcal{P}_\epsilon$: the cost function is penalized and adapted to the optimal control $(\underline{f}, \underline{u})$. This type of method was introduced by Lions [12], [14] to control singular distributed systems. More recent works have developed the method of control in mathematical physics and performed variational data assimilation [1], [2], [3], [4], [15], [16].

The specificity of this paper derives from the following features: the direct problem is time-dependent and of shallow-water type, the equations are coupled and nonlinear. Moreover, the altimetric measurements allow us to take the sea level topography as the observation. The penalized control problem $\mathcal{P}_\epsilon$ is defined in such a way that the direct problem is linear. In order to characterize $(\underline{f}_\epsilon, \underline{w}_\epsilon)$, solution of $\mathcal{P}_\epsilon$, we define a suitable adjoint problem. The solution $(\underline{f}, \underline{u})$ of our initial control problem is obtained as the limit of $(\underline{f}_\epsilon, \underline{w}_\epsilon)$ when $\epsilon \to 0$. The set of equations and inequality characterizing $(\underline{f}, \underline{u})$ is obtained by passing to the limit in the characterization of $(\underline{f}_\epsilon, \underline{w}_\epsilon)$. The same technique can be used in the nonhomogeneous case: $\tilde{\underline{u}}_B \neq \underline{0}$, $\tilde{h}_B \neq 0$ because the velocity $\underline{u}_L$ and the depth h_L do not depend on control $\underline{f}$. $(\underline{u}_L, h_L)$ is calculated first. Then the control method is applied to the homogeneous problem (1.5).

The main result of this paper is the following theorem:

> Under the assumptions of Proposition 1, the optimal control problem $(\mathcal{P})$ has at least one solution $(\underline{u}, h, \underline{f})$ such that $\underline{f} \in K_c$ and $(\underline{u}, h)$ is the solution of problem (2.1) with a forcing equal to $\underline{f}$. Moreover, there exists $(\underline{R}, S) \in L^2(0,T,(H^1_0(\Omega))^3) \cap L^\infty(0,T,(L^2(\Omega))^3)$ such that

$$-\left(\frac{\partial \underline{R}}{\partial t}, \underline{v}\right) + a_1(\underline{R}, \underline{v}) + \frac{C}{D}\left(\mid (\underline{u}+\underline{u}_L) \mid_2 \underline{R} + (\underline{u}+\underline{u}_L, \underline{R})_2 \frac{(\underline{u}+\underline{u}_L)}{\mid (\underline{u}+\underline{u}_L) \mid_2}, \underline{v}\right)$$
$$-(\underline{F} \wedge \underline{R}, \underline{v}) - ((h+h_L)\nabla S, \underline{v}) = 0,$$
$$-\left(\frac{\partial S}{\partial t}, \beta\right) + a_2(S,\beta) - (((\underline{u}+\underline{u}_L)\nabla)S, \beta) - g(\mathrm{div}(\underline{R}), \beta) = ((h-h_{obs})^3, \beta),$$
$$(\underline{R}, S)(t=T) = (\underline{0}, 0) \quad \forall(\underline{v}, \beta) \in (H^1_0(\Omega))^3 \quad \text{almost everywhere (a.e.) } t \in (0,T),$$

and

$$(\underline{R} + \gamma \underline{f},\ \underline{g} - \underline{f})_{L^2(0,T,L^2(\Omega))} \geq 0\ \forall \underline{g} \in K_c.$$

2. Existence and uniqueness conditions. Let Ω be a fixed bounded open domain of $\mathbb{R}^2$. Γ denotes its boundary and is supposed to be sufficiently regular.

We introduce the following functional spaces:
$V_1 = (H^1_0(\Omega))^2$, $H_1 = (L^2(\Omega))^2$, $V_2 = H^1_0(\Omega)$, $H_2 = L^2(\Omega), V = V_1 \times V_2$, and $H = H_1 \times H_2$.

The norm and the seminorm defined on $H^1(\Omega)$ are equivalent in V_1, V_2, and V. Then we set $\| \underline{v} \| = \| \underline{v} \|_{V_1}$, $\| \beta \| = \| \beta \|_{V_2}$, and $\| X \| = \| X \|_V$ for $\underline{v} \in V_1, \beta \in V_2$, and $X \in V$. $\mid . \mid$ denotes the norm in $L^2(\Omega)$, $\mid . \mid_2$ denotes the euclidean norm in $\mathbb{R}^2$, and $(.,.)_2$ denotes the scalar product in $\mathbb{R}^2$.

We define

$a_1(\underline{u}, \underline{v}) = \nu_1(\nabla \underline{u}, \nabla \underline{v})$,
$a_2(h, \beta) = \nu_2(\nabla h, \nabla \beta)$,
$a(X,Y) = a_1(\underline{u}, \underline{v}) + a_2(h,\beta)$ with $X = (\underline{u}, h)$ and $Y = (\underline{v}, \beta)$.

a_1, a_2, and a are bilinear continuous coercive forms, respectively, on V_1, V_2, and V and we denote by α the constant of coercivity of a.

We can now write the weak formulation of problem (1.5):

(2.1)
Find $(\underline{u}, h) \in L^2(0,T,V) \cap L^\infty(0,T,H)$ such that

$$\left(\frac{\partial \underline{u}}{\partial t}, \underline{v}\right) + a_1(\underline{u}, \underline{v}) + \frac{C}{D}(\mid (\underline{u}+\underline{u}_L) \mid_2 (\underline{u}+\underline{u}_L), \underline{v}) + (\underline{F} \wedge \underline{u}, \underline{v}) - g(\mathrm{div}(\underline{v}), h) = (\underline{f}, \underline{v}),$$
$$\left(\frac{\partial h}{\partial t}, \beta\right) + a_2(h,\beta) + (\mathrm{div}(h\underline{u}), \beta) + (\mathrm{div}(h_L\underline{u}), \beta) + (\mathrm{div}(h\underline{u}_L), \beta) = (f_1, \beta)\ \forall(\underline{v}, \beta) \in V,$$
$$(\underline{u}, h)(t=0) = (\underline{u}_0, h_0).$$

Before studying the existence and uniqueness conditions, we are first going to prove four lemmas.

Lemma 1. *The operator :* $\underline{u} \longrightarrow | \underline{u} |_2 \, \underline{u}$ *is a monotone operator.*

Proof. Let $(\underline{u}, \underline{v}) \in V_1 \times V_1$,

$$(| \underline{u} |_2 \, \underline{u} - | \underline{v} |_2 \, \underline{v}, \underline{u} - \underline{v}) = (| \underline{u} |_2^2, | \underline{u} |_2) - (| \underline{u} |_2 \, \underline{u}, \underline{v}) - (| \underline{v} |_2 \, \underline{v}, \underline{u}) + (| \underline{v} |_2^2, | \underline{v} |_2).$$

Since $(\underline{u}, \underline{v})_2 \leq | \underline{u} |_2 | \underline{v} |_2$ we have

$$(| \underline{u} |_2 \, \underline{u} - | \underline{v} |_2 \, \underline{v}, \underline{u} - \underline{v}) \geq (| \underline{u} |_2^2 - | \underline{v} |_2^2, | \underline{u} |_2 - | \underline{v} |_2) = ((| \underline{u} |_2 - | \underline{v} |_2)^2, | \underline{u} |_2 + | \underline{v} |_2).$$

This implies that

$$(| \underline{u} |_2 \, \underline{u} - | \underline{v} |_2 \, \underline{v}, \underline{u} - \underline{v}) \geq 0. \qquad \square$$

Lemma 2. *If* $(\underline{u}, \underline{v})$ *is given in* $V_1 \times V_1$, *we have the following results:*

(i) $| \underline{u} |_2 \, \underline{v} \in L^{3/2}(\Omega)$.

(ii) *There exists a positive constant* c *such that* $\| \quad | \underline{u} |_2 \, \underline{v} \|_{V_1'} \leq c \, | \underline{u} | \quad \| \underline{v} \|$.

Proof. (i) Using the Schwarz inequality, we obtain

$$\int_\Omega | \underline{u} |_2^{3/2} | \underline{v} |_2^{3/2} \, d\Omega \leq c \left(\int_\Omega | \underline{u} |_2^2 \, d\Omega \right)^{3/4} \left(\int_\Omega | \underline{v} |_2^6 \, d\Omega \right)^{1/4}.$$

By applying the Sobolev injections ([11], for example), we have $\| | \underline{u} |_2 \, \underline{v} \|_{L^{3/2}(\Omega)} \leq c \, | \underline{u} | \| \underline{v} \|$ and then $| \underline{u} |_2 \, \underline{v} \in L^{3/2}(\Omega)$.

(ii) Setting $\underline{w} \in V_1$, we have $| (| \underline{u} |_2 \, \underline{v}, \underline{w}) | \leq c \, \| | \underline{u} |_2 \, \underline{v} \|_{L^{3/2}(\Omega)} \| \underline{w} \|_{L^3(\Omega)}$. Using result (i), we obtain $| (| \underline{u} |_2 \, \underline{v}, \underline{w}) | \leq c \, | \underline{u} | \| \underline{v} \| \| \underline{w} \|$, from which we can deduce result (ii). $\square$

Lemma 3. (i) *If* $\underline{v} \in V_1$ *or* $h \in V_2$, *then* $| (\mathrm{div}(h\underline{v}), \beta) | \leq c \, \| h \|_{L^4(\Omega)} \| \beta \| \| \underline{v} \|_{L^4(\Omega)}$,

(ii) *If* $(\underline{v}, h) \in V$, *then we have (Gagliardo–Nirenberg's inequality)* $\| \underline{v} \|_{L^4(\Omega)} \leq c \, \| \underline{v} \|^{\frac{1}{2}} | \underline{v} |^{\frac{1}{2}}$ *and* $\| h \|_{L^4(\Omega)} \leq c \, \| h \|^{\frac{1}{2}} | h |^{\frac{1}{2}}$.

Proof. (i) By applying the Green formula and $\underline{v} \in V_1$ (or $h \in V_2$) we obtain $| (\mathrm{div}(h\underline{v}), \beta) | = | ((h\nabla)\beta), \underline{v}) |$. Thus $| (\mathrm{div}(h\underline{v}), \beta) | \leq c \, \| h \|_{L^4(\Omega)} \| \beta \| \| \underline{v} \|_{L^4(\Omega)}$. For the proof of (ii) see, for example, [6], [18], [19].

Lemma 4. *Let* $(\underline{u}_m, h_m)$ *be a sequence converging toward* $(\underline{u}, h)$ *in* $L^2(0, T, H)$ *strongly and in* $L^2(0, T, V)$ *weakly. Then for any vector function* $\varphi(t)(\underline{v}, \beta)$, $\varphi \in C^1([0, T])$ *and* $(\underline{v}, \beta) \in V$, *we have*

(i) $\lim_{m \to \infty} \int_0^T (| \underline{u}_m + \underline{u}_L |_2 \, (\underline{u}_m + \underline{u}_L), \varphi(t)\underline{v}) dt = \int_0^T (| (\underline{u} + \underline{u}_L) |_2 \, (\underline{u} + \underline{u}_L), \varphi(t)\underline{v}) dt$.

(ii) $\lim_{m \to \infty} \int_0^T (\mathrm{div}(\underline{u}_m h_m), \varphi(t)\beta) dt = \int_0^T (\mathrm{div}(\underline{u}h), \varphi(t)\beta) dt$.

Proof. (i) We write

$$\int_0^T (| \underline{u}_m + \underline{u}_L |_2 \, (\underline{u}_m + \underline{u}_L), \varphi(t)\underline{v}) dt = \int_0^T \langle | \underline{u}_m + \underline{u}_L |_2 \, \varphi(t)\underline{v}, \underline{u}_m + \underline{u}_L \rangle_{V_1', V_1} dt$$

and

$$\int_0^T (| \underline{u} + \underline{u}_L |_2 \, (\underline{u} + \underline{u}_L), \varphi(t)\underline{v}) dt = \int_0^T \langle | \underline{u} + \underline{u}_L |_2 \, \varphi(t)\underline{v}, \underline{u} + \underline{u}_L \rangle_{V_1', V_1} dt.$$

Subtracting the previous equalities gives

$$\left| \int_0^T (\mid \underline{u}_m + \underline{u}_L \mid_2 (\underline{u}_m + \underline{u}_L), \varphi(t)\underline{v}) dt - \int_0^T (\mid \underline{u} + \underline{u}_L \mid_2 (\underline{u} + \underline{u}_L), \varphi(t)\underline{v}) dt \right|$$

$$\begin{aligned}
&= \left| \int_0^T \langle \mid \underline{u}_m + \underline{u}_L \mid_2 \varphi(t)\underline{v}, \underline{u}_m + \underline{u}_L \rangle_{V_1', V_1} dt \right. \\
&\qquad \left. - \int_0^T \langle \mid \underline{u} + \underline{u}_L \mid_2 \varphi(t)\underline{v}, \underline{u} + \underline{u}_L \rangle_{V_1', V_1} dt \right| \\
&\leq \left| \int_0^T \langle (\mid \underline{u}_m + \underline{u}_L \mid_2 - \mid \underline{u} + \underline{u}_L \mid_2) \varphi(t)\underline{v}, \underline{u}_m + \underline{u}_L \rangle_{V_1', V_1} dt \right| \\
&\qquad + \left| \int_0^T \langle \mid \underline{u} + \underline{u}_L \mid_2 \varphi(t)\underline{v}, \underline{u}_m - \underline{u} \rangle_{V_1', V_1} dt \right| \\
&\leq \; \parallel (\mid \underline{u}_m - \underline{u} \mid_2) \varphi(t)\underline{v} \parallel_{L^2(0,T,V_1')} \parallel \underline{u}_m + \underline{u}_L \parallel_{L^2(0,T,V_1)} \\
&\qquad + \left| \int_0^T \langle \mid \underline{u} + \underline{u}_L \mid_2 \varphi(t)\underline{v}, \underline{u}_m - \underline{u} \rangle_{V_1', V_1} dt \right|.
\end{aligned}$$

Since $\underline{u}_m$ converges towards $\underline{u}$ strongly in $L^2(0, T, H_1)$ and weakly in $L^2(0, T, V_1)$, since φ is bounded, and by using lemma 2 we obtain result (i).

(ii) By applying the Green formula, we have

$$\begin{aligned}
\int_0^T (\operatorname{div}(\underline{u}_m h_m), \varphi(t)\beta) dt &= - \int_0^T (h_m \varphi(t) \nabla\beta, \underline{u}_m) dt \\
&= - \int_0^T \sum_{i=1}^{i=2} \int_\Omega h_m \varphi(t) \partial_i \beta (u_m)_i d\Omega dt.
\end{aligned}$$

Using Temam [18, Lemma 3.2, p. 289] these integrals converge to

$$- \int_0^T (h\varphi(t)\nabla\beta, \underline{u}) dt = \int_0^T (\operatorname{div}(h\underline{u}), \varphi(t)\beta) dt. \qquad \square$$

We can now state the result of existence and uniqueness.

Proposition 1. *We assume that* $F_{ex} = (\underline{f}, f_1) \in L^2(0, T, H)$, $X_0 = (\underline{u}_0, h_0) \in V \cap (L^\infty(\Omega))^3$, *and* $(\underline{u}_L, h_L) \in L^2(0, T, (H^1(\Omega))^3) \cap L^\infty(0, T, (L^\infty(\Omega))^3)$.
If the following conditions are satisfied,

(i) $K = (2\alpha - C_g - C_L(\parallel \underline{u}_L \parallel_{L^\infty(0,T,L^4(\Omega))} + \parallel h_L \parallel_{L^\infty(0,T,L^4(\Omega))}))/C_0 > 0$,

(ii) $\mid X_0 \mid < K$,

(iii) $\mid X_0 \mid^2 + C_I(\parallel F_{ex} \parallel^2_{L^2(0,T,L^2(\Omega))} + \parallel \underline{u}_L \parallel^3_{L^3(0,T,L^4(\Omega))}) < K^2$,

then problem (2.1) *admits one unique solution* $(\underline{u}, h)$ *in* $L^2(0, T, V) \cap L^\infty(0, T, H)$.

Remark 2. *If* $(\underline{u}_L, h_L) \in L^2(0, T, (H^1(\Omega))^3) \cap L^\infty(0, T, (L^\infty(\Omega))^3)$, *then* $\operatorname{div}(h_L \underline{u}_L) \in L^2(0, T, L^2(\Omega))$. *If* $(\underline{\tilde{f}}, f_0) \in L^2(0, T, H)$, *then* $(\underline{f}, f_1) \in L^2(0, T, H)$. $\square$

Proof of the existence.

Since V is separable, there exists a free and total sequence $(\underline{w}_1, p_1), \ldots, (\underline{w}_m, p_m), \ldots$ in V. We denote by V_m the space generated by $(\underline{w}_1, p_1), \ldots, (\underline{w}_m, p_m)$. For each m

we define an approach problem:

(2.2)

$$\begin{aligned}
&\text{Find } (\underline{u}_m, h_m) \in V_m \text{ such that}\\
&\left(\frac{\partial \underline{u}_m}{\partial t}, \underline{w}_k\right) + a_1(\underline{u}_m, \underline{w}_k) + \frac{C}{D}(\mid \underline{u}_m + \underline{u}_L \mid_2 (\underline{u}_m + \underline{u}_L), \underline{w}_k) + (\underline{F} \wedge \underline{u}_m, \underline{w}_k)\\
&-g(\mathrm{div}(\underline{w}_k), h_m) = (\underline{f}, \underline{w}_k) \ \forall k = 1, m,\\
&\left(\frac{\partial h_m}{\partial t}, p_k\right) + a_2(h_m, p_k) + (\mathrm{div}(h_m \underline{u}_m), p_k) + (\mathrm{div}(h_m \underline{u}_L), p_k)\\
&+(\mathrm{div}(h_L \underline{u}_m), p_k) = (f_1, p_k) \ \forall k = 1, m,\\
&(\underline{u}_m, h_m)(t = 0) = (\underline{u}_{0m}, h_{0m}),
\end{aligned}$$

where, $(\underline{u}_{0m}, h_{0m})$ is the orthogonal projection in H of $(\underline{u}_0, h_0)$ on V_m such that $(\underline{u}_{0m}, h_{0m})$ converges strongly towards $(\underline{u}_0, h_0)$ in H, $\mid \underline{u}_{0m} \mid \leq \mid \underline{u}_0 \mid$, and $\mid h_{0m} \mid \leq \mid h_0 \mid$. Since $(\underline{u}_m, h_m) \in V_m$, we have $(\underline{u}_m, h_m) = (\sum_{k=1}^m g_{km}(t)\underline{w}_k, \sum_{k=1}^m l_{km}(t)p_k)$. (g_{km}, l_{km}) are scalar functions defined on $[0, T]$.

Multiplying the first half of (2.2) by g_{km} and the second half of (2.2) by l_{km} and adding with respect to k, we obtain

$$\begin{aligned}
(2.3)\qquad &\frac{1}{2}\frac{\partial \mid \underline{u}_m \mid^2}{\partial t} + a_1(\underline{u}_m, \underline{u}_m) + \frac{C}{D} \parallel \underline{u}_m + \underline{u}_L \parallel^3_{L^3(\Omega)} - g(\mathrm{div}(\underline{u}_m), h_m)\\
&\qquad -\frac{C}{D}(\mid \underline{u}_m + \underline{u}_L \mid_2 (\underline{u}_m + \underline{u}_L), u_L) = (\underline{f}, \underline{u}_m),\\
&\frac{1}{2}\frac{\partial \mid h_m \mid^2}{\partial t} + a_2(h_m, h_m) + (\mathrm{div}(h_m \underline{u}_m), h_m) + (\mathrm{div}(h_m \underline{u}_L), h_m)\\
&\qquad +(\mathrm{div}(h_L \underline{u}_m), h_m) = (f_1, h_m).
\end{aligned}$$

From these two equations we deduce

$$\begin{aligned}
&\frac{1}{2}\frac{\partial \mid X_m \mid^2}{\partial t} + a(X_m, X_m) + \frac{C}{D} \parallel \underline{u}_m + \underline{u}_L \parallel^3_{L^3(\Omega)} = \frac{C}{D}(\mid \underline{u}_m + \underline{u}_L \mid_2 (\underline{u}_m + \underline{u}_L), \underline{u}_L)\\
&\quad + g(\mathrm{div}(\underline{u}_m), h_m) - (\mathrm{div}(h_m \underline{u}_m), h_m)\\
&\quad - (\mathrm{div}(h_m \underline{u}_L), h_m) - (\mathrm{div}(h_L \underline{u}_m), h_m) + (F_{ex}, X_m),
\end{aligned}$$

where $X_m = (\underline{u}_m, h_m)$ and $F_{ex} = (\underline{f}, f_1)$.

According to Green's formula and Gagliardo–Nirenberg's inequality we have

$$2(\mathrm{div}(h_m \underline{u}_m), h_m) = (h_m^2, \mathrm{div}(\underline{u}_m))$$

and

$$\mid 2(\mathrm{div}(h_m \underline{u}_m), h_m) \mid \leq C \parallel h_m \parallel^2_{L^4(\Omega)} \parallel \underline{u}_m \parallel \leq C_0 \parallel X_m \parallel^2 \mid X_m \mid .$$

According to Lemma 3 we have

$$\begin{aligned}
\mid 2(\mathrm{div}(h_m \underline{u}_L), h_m) \mid &\leq C_L \parallel \underline{u}_L \parallel_{L^4(\Omega)} \parallel X_m \parallel^2,\\
\mid 2(\mathrm{div}(h_L \underline{u}_m), h_m) \mid &\leq C_L \parallel h_L \parallel_{L^4(\Omega)} \parallel X_m \parallel^2 .
\end{aligned}$$

Moreover,

$$\left|\frac{2C}{D}(\mid \underline{u}_m + \underline{u}_L \mid_2 (\underline{u}_m + \underline{u}_L), \underline{u}_L)\right| \leq C_L(\parallel \underline{u}_L \parallel_{L^4(\Omega)} \parallel X_m \parallel^2 + \parallel \underline{u}_L \parallel^3_{L^4(\Omega)})$$

and

$$| g(\mathrm{div}(\underline{u}_m), h_m) | \leq C_g \parallel X_m \parallel^2 .$$

Using the previous inequalities and the coercivity of a we deduce

$$\frac{\partial \mid X_m \mid^2}{\partial t} + \psi_m(t) \parallel X_m \parallel^2 \leq C_I(\mid F_{ex} \mid^2 + \parallel \underline{u}_L \parallel^3_{L^4(\Omega)}), \tag{2.4}$$

where $\psi_m(t) = 2\alpha - C_0 \mid X_m(t) \mid - C_g - C_L(\parallel \underline{u}_L \parallel_{L^\infty(0,T,L^4(\Omega))} + \parallel h_L \parallel_{L^\infty(0,T,L^4(\Omega))})$.

Let us prove that $\psi_m(t) > 0 \ \forall t \in [0,T]$.

According to condition (ii), we have $\psi_m(0) > 0$. Let us suppose that there exists $t \in [0,T]$ such that $\psi_m(t) \leq 0$. Since function ψ_m is continuous, there exists $T_0 \in [0,T[$ such that $\forall t \in [0, T_0[, \psi_m(t) > 0$ and $\psi_m(T_0) = 0$.

Integrating (2.4) from 0 to T_0 we obtain

$$\begin{aligned} &| X_m(T_0) |^2 + \int_0^{T_0} \psi_m(t) \parallel X_m(t) \parallel^2 dt \\ &\quad \leq C_I(\parallel F_{ex} \parallel^2_{L^2(0,T,L^2(\Omega))} + \parallel \underline{u}_L \parallel^3_{L^3(0,T,L^4(\Omega))}) + | X_0 |^2 . \end{aligned}$$

Since $\psi_m > 0$ on $[0, T_0[$ and $\psi_m(T_0) = 0$, we obtain

$$K^2 \leq C_I(\parallel F_{ex} \parallel^2_{L^2(0,T,L^2(\Omega))} + \parallel \underline{u}_L \parallel^3_{L^3(0,T,L^4(\Omega))}) + | X_0 |^2;$$

thus condition (iii) cannot be satisfied and we conclude that

$$\psi_m(t) > 0 \ \forall t \in [0,T].$$

We can now deduce from inequality (2.4) that X_m is uniformly bounded in $L^\infty(0,T,H) \cap L^2(0,T,V)$. This result makes it possible to extract from X_m a sub sequence also denoted by X_m which converges towards X in $L^\infty(0,T,H)$ weakly* and $L^2(0,T,V)$ weakly. By using (2.2), the continuity of the operator a, Hölder's inequality, and Lemma 3, we obtain

$$\left(\frac{\partial X_m}{\partial t}, w \right) \leq c(\parallel X_m \parallel + \parallel X_m \parallel^2 + | F_{ex} | + \parallel \underline{u}_L \parallel^2_{L^4(\Omega)} + \parallel h_L \parallel^2_{L^4(\Omega)}) \parallel w \parallel,$$

where $w = (\underline{v}, \beta)$.

Since X_m is uniformly bounded in $L^2(0,T,V)$, F_{ex} is in $L^2(0,T,L^2(\Omega))$ and $(\underline{u}_L, h_L)$ is in $L^2(0,T,L^4(\Omega))$, we can deduce that $\partial X_m / \partial t$ is uniformly bounded in $L^1(0,T,V')$.

Let us introduce space $Y = \{w \in L^2(0,T,V), \partial w/\partial t \in L^1(0,T,V')\}$. According to Temam [18], [19] the injection of Y into $L^2(0,T,H)$ is compact. Sequence X_m is uniformly bounded in Y. Then we can extract from X_m a subsequence also denoted by X_m and such that

$$\begin{array}{ll} X_m \rightharpoonup X & \text{weakly in } L^2(0,T,V), \\ X_m \longrightarrow X & \text{strongly in } L^2(0,T,H). \end{array}$$

We are going to prove that $X = (\underline{u}, h)$ is solution of problem (2.1).

In order to pass to the limit in (2.2), let us consider $\varphi \in C^1([0,T])$ such that $\varphi(T) = 0$.

Multiplying (2.2) by φ and integrating with respect to time, we obtain

(2.5)

$$\begin{aligned}
&-\int_0^T (\underline{u}_m, \varphi'(t)\underline{w}_k)dt + \int_0^T a_1(\underline{u}_m, \varphi(t)\underline{w}_k)dt + \frac{C}{D}\int_0^T (| \underline{u}_m + \underline{u}_L |_2 (\underline{u}_m + \underline{u}_L), \varphi(t)\underline{w}_k)dt \\
&+\int_0^T (\underline{F} \wedge \underline{u}_m, \varphi(t)\underline{w}_k)dt + g\int_0^T (\nabla h_m, \varphi(t)\underline{w}_k)dt + (u_{0m}, \varphi(0)\underline{w}_k) = \int_0^T (\underline{f}, \varphi(t)\underline{w}_k)dt, \\
&-\int_0^T (h_m, \varphi'(t)p_k)dt + \int_0^T a_2(\underline{u}_m, \varphi(t)p_k)dt + \int_0^T (\mathrm{div}(h_m\underline{u}_m), \varphi(t)p_k)dt + (h_{0m}, \varphi(0)p_k) \\
&+\int_0^T (\mathrm{div}(h_m\underline{u}_L), \varphi(t)p_k)dt + \int_0^T (\mathrm{div}(h_L\underline{u}_m), \varphi(t)p_k)dt = \int_0^T (f_1, \varphi(t)p_k)dt.
\end{aligned}$$

It is easy to pass to the limit in the linear terms. For the nonlinear terms we apply the result of Lemma 4.

Since (u_{0m}, h_{0m}) converges towards $(\underline{u}_0, h_0)$, the limit $(\underline{u}, h)$ verifies the equation

(2.6)

$$\begin{aligned}
&-\int_0^T (\underline{u}, \varphi'(t)\underline{v})dt + \int_0^T a_1(\underline{u}, \varphi(t)\underline{v})dt + \frac{C}{D}\int_0^T (| (\underline{u} + \underline{u}_L) |_2 (\underline{u} + \underline{u}_L), \varphi(t)\underline{v})dt, \\
&+\int_0^T (\underline{F} \wedge \underline{u}, \varphi(t)\underline{v})dt + g\int_0^T (\nabla h, \varphi(t)\underline{v})dt + (\underline{u}_0, \varphi(0)\underline{v}) = \int_0^T (\underline{f}, \varphi(t)\underline{v})dt, \\
&-\int_0^T (h, \varphi'(t)\beta)dt + \int_0^T a_2(h, \varphi(t)\beta)dt + \int_0^T (\mathrm{div}(h\underline{u}), \varphi(t)\beta)dt + (h_0, \varphi(0)\beta) \\
&+\int_0^T (\mathrm{div}(h_L\underline{u}), \varphi(t)\beta)dt + \int_0^T (\mathrm{div}(h\underline{u}_L), \varphi(t)\beta)dt = \int_0^T (f_1, \varphi(t)\beta)dt
\end{aligned}$$

for $\underline{v} = \underline{w}_1, \underline{w}_2, \dots$ and $\beta = p_1, p_2, \dots$. By linearity and a continuity argument this equation holds for any $(\underline{v}, \beta) \in V$.

Assuming $\varphi \in \mathcal{D}(0, T)$, we deduce that $(\underline{u}, h)$ verifies (2.1) in the distribution sense on $(0, T)$.

In order to prove the initial conditions, we multiply (2.1) by a function $\varphi \in C^1([0, T])$ such that $\varphi(T) = 0$. Integrating with respect to t and comparing with (2.7), we obtain

$$\underline{u}(0) = \underline{u}_0 \text{ and } h(0) = h_0.$$

Since $(\underline{u}, h) \in L^2(0, T, V)$ and $(\underline{u}', h') \in L^1(0, T, V')$, we conclude that $(\underline{u}, h)$ is weakly continuous from $[0, T]$ into H [13].

Remark 3. (i) *Since $(\underline{u}, h)$ is in $L^\infty(0, T, H) \cap L^2(0, T, V)$, we obtain that $(\underline{u}, h)$ is in $L^4(0, T, L^4(\Omega))$ by applying result* (ii) *of Lemma* 3.

(ii) *If we a priori impose the regularity $h \in L^4(0, T, L^4(\Omega))$, then we obtain the existence of a solution $(\underline{u}, h)$ without any conditions.*

Proof of the uniqueness.

Let us suppose that problem (2.1) admits two solutions $(\underline{u}_1, h_1)$ and $(\underline{u}_2, h_2)$. Set $\underline{u} = \underline{u}_1 - \underline{u}_2, h = h_1 - h_2$, and $X = (\underline{u}, h)$.

$(\underline{u},h)$ is the solution of the problem

(2.7)

$$\begin{aligned}
&\left(\frac{\partial \underline{u}}{\partial t},\underline{v}\right)+a_1(\underline{u},\underline{v})+\frac{C}{D}(\mid \underline{u}_1+\underline{u}_L \mid_2 (\underline{u}_1+\underline{u}_L)- \mid \underline{u}_2+\underline{u}_L \mid (\underline{u}_2+\underline{u}_L),\underline{v})\\
&+(\underline{F}\wedge \underline{u},\underline{v})-g(\mathrm{div}(\underline{v}),h)=0,\\
&\left(\frac{\partial h}{\partial t},\beta\right)+a_2(h,\beta)+(\mathrm{div}(h_1\underline{u}_1-h_2\underline{u}_2),\beta)+(\mathrm{div}(h_L\underline{u}+h\underline{u}_L),\beta)=0\ \forall(\underline{v},\beta)\in V,\\
&(\underline{u},h)(t=0)=0.
\end{aligned}$$

Setting $(\underline{v},\beta)=(\underline{u},h)$ in (2.7), we obtain

$$\begin{aligned}
\frac{1}{2}\frac{\partial \mid X\mid^2}{\partial t}+a(X,X)+\frac{C}{D}(\mid \underline{u}_1 \mid_2 \underline{u}_1- \mid \underline{u}_2 \mid_2 \underline{u}_2,\underline{u}) &= g(\mathrm{div}(\underline{u}),h)\\
&-(\mathrm{div}((h_1+h_L)\underline{u}),h)-(\mathrm{div}(h(\underline{u}_2+\underline{u}_L)),h).
\end{aligned} \tag{2.8}$$

Using the coercivity of operator a and Lemmas 1, 3 we obtain

$$\begin{aligned}
&\frac{\partial \mid X\mid^2}{\partial t}+2\alpha \parallel X\parallel^2\\
&\quad\le c_1 \mid X \mid\parallel X\parallel +c_2 \mid X\mid^{1/2}\parallel X\parallel^{3/2}(\parallel \underline{u}_2+\underline{u}_L\parallel_{L^4(\Omega)}+\parallel h_1+h_L\parallel_{L^4(\Omega)}).
\end{aligned}$$

By applying Hölder's inequality we have

$$\frac{\partial \mid X\mid^2}{\partial t}+\alpha\parallel X\parallel^2\le c(1+\parallel \underline{u}_2+\underline{u}_L\parallel^4_{L^4(\Omega)}+\parallel h_1+h_L\parallel^4_{L^4(\Omega)})\mid X\mid^2 .$$

The functions $\parallel \underline{u}_2+\underline{u}_L\parallel^4_{L^4(\Omega)}$ and $\parallel h_1+h_L\parallel^4_{L^4(\Omega)}$ being integrable with respect to time, we deduce by applying Gronwall's lemma that

$$\mid X(t)\mid^2\le c\ \mid X(0)\mid^2 \exp\left(\int_0^t(\parallel \underline{u}_2(s)+\underline{u}_L(s)\parallel^4_{L^4(\Omega)}+\parallel h_1(s)+h_L(s)\parallel^4_{L^4(\Omega)})ds\right).$$

Since $\mid X\mid(0)=0$, then $\mid X(t)\mid^2\le 0\ \forall t\in[0,T]$.

Hence

$$\underline{u}_1=\underline{u}_2 \quad\text{and}\quad h_1=h_2. \qquad \square$$

Now we give a physical interpretation of conditions (i) and (iii).

(i) $K=(2\alpha-C_g-C_L(\parallel \underline{u}_L\parallel_{L^\infty(0,T,L^4(\Omega))}+\parallel h_L\parallel_{L^\infty(0,T,L^4(\Omega))}))/C_0>0$, where $\alpha=\min(\nu_1,\nu_2)$.

$(\underline{u}_L,h_L)$ corresponds to a lifting of the boundary conditions $(\underline{\tilde{u}}_B,\tilde{h}_B)$. Therefore condition (i) expresses a relationship between the eddy viscosity coefficients ν_1, ν_2 and the boundary conditions. ν_1 and ν_2 have to be chosen large enough to verify (i).

(iii) $\mid X_0\mid^2+C_I(\parallel F_{ex}\parallel^2_{L^2(0,T,L^2(\Omega))}+\parallel \underline{u}_L\parallel^3_{L^3(0,T,L^4(\Omega))})<K^2$, where $X_0=(\underline{u}_0,h_0)$.

The initial conditions $(\underline{u}_0,h_0)$, the boundary conditions, and the eddy viscosity coefficients being fixed, condition (iii) applies to the variability of the external forcing F_{ex} which has to be small enough. Physically, this condition is realistic and not too restrictive.

3. Existence and characterization of the optimal control.

The problem is controlled by the external stress $\underline{f}$. The depth of the layer is $h = \xi + H - h_L$. The bottom topography H is known and the surface topography ξ can be deduced from altimetric measurements. Therefore h will be the observation for the control model.

Thus $U_c = L^2(0, T, L^2(\Omega))$ will be the control space; we choose $L^4(0, T, L^4(\Omega))$ as the observation space; $h_{obs} \in L^4(0, T, L^4(\Omega))$ denotes the observation. In order to obtain the existence of a solution for the direct problem, we assume that the data are small and the fluid is viscous enough.

The cost function J is defined by

$$J(\underline{g}, \underline{v}) = \frac{1}{4} \parallel h - h_{obs} \parallel^4_{L^4(0,T,L^4(\Omega))} + \frac{1}{2}\gamma \parallel \underline{g} \parallel^2_{U_c} \quad \text{with } \gamma > 0,$$

where $(\underline{v}, h)$ is the solution of the direct problem (2.1), the external forcing being equal to $\underline{g}$.

Let K_c be a convex, closed, nonempty subset of U_c. The optimal control problem can be written as follows:

$$\begin{aligned} &\text{Find } (\underline{f}, \underline{u}) \in U_{ad} \text{ such that} \\ &J(\underline{f}, \underline{u}) = \inf_{(\underline{g},\underline{v}) \in U_{ad}} J(\underline{g}, \underline{v}), \end{aligned} \tag{3.1}$$

$U_{ad} = \{(\underline{g}, \underline{v}) \in K_c \times L^2(0, T, V_1) \cap L^\infty(0, T, H_1)/J(\underline{g}, \underline{v}) < \infty$, and there exists $h \in L^2(0, T, V_2) \cap L^\infty(0, T, H_2)$ such that $(\underline{v}, h, \underline{g})$ is solution of problem (2.1),with a forcing equal to $\underline{g}\}$ is the admissibility set.

REMARK 4. *The space U_{ad} is not empty. Let $\underline{g} = 0$ and then let $(\underline{v}, h) \in L^2(0, T, V) \cap L^\infty(0, T, H)$ be the solution of* (2.1). *According to Proposition* 1 $(\underline{v}, h)$ *exists. Moreover we have $J(0, \underline{v}) \le c(\parallel h \parallel^{1/2}_{L^2(0,T,V_1)} \parallel h \parallel^{1/2}_{L^\infty(0,T,H_1)} + \parallel h_{obs} \parallel_{L^4(0,T,L^4(\Omega))})^4 < \infty$. Thus $(0, \underline{v}, h) \in U_{ad}$.*

3.1. Existence and characterization of an optimal control in the case of homogeneous boundary conditions.

We first deal with the case of homogeneous boundary conditions $\tilde{\underline{u}}_B = \underline{0}, \tilde{h}_B = 0$. Then we have $\underline{u}_L = 0, h_L = 0$ and the initial problem (1.1)–(1.2) is identical to problem (1.5).

THEOREM 1. *Under the assumptions of Proposition* 1, *the optimal control problem* (3.1) *has at least one solution $(\underline{f}, \underline{u}) \in U_{ad}$.*

Proof. Let $(\underline{f}_k, \underline{u}_k)$ be a minimizing sequence in U_{ad} for the function J such that

$$\liminf_{k \to \infty} J(\underline{f}_k, \underline{u}_k) = \inf_{(\underline{g},\underline{v}) \in U_{ad}} J(\underline{g}, \underline{v}).$$

Since $J(\underline{f}_k, \underline{u}_k)$ is bounded for a minimizing sequence $(\underline{f}_k, \underline{u}_k) \in U_{ad}$, then $(\underline{f}_k, h_k)$ is uniformly bounded in $U_c \times L^4(0, T, L^4(\Omega))$ according to the definition of J.

Setting $(\underline{v}, \beta) = (\underline{u}_k, h_k)$ in (2.1) gives

$$\begin{aligned} \frac{1}{2}\frac{\partial \mid \underline{u}_k \mid^2}{\partial t} + a_1(\underline{u}_k, \underline{u}_k) + \frac{C}{D}(\mid \underline{u}_k \mid_2 \underline{u}_k, \underline{u}_k) - g(\mathrm{div}(\underline{u}_k), h_k) &= (\underline{f}_k, \underline{u}_k), \\ \frac{1}{2}\frac{\partial \mid h_k \mid^2}{\partial t} + a_2(h_k, h_k) + (\mathrm{div}(h_k \underline{u}_k), h_k) &= (f_1, h_k). \end{aligned}$$

Adding these two equations, we obtain

$$\frac{1}{2}\frac{\partial \mid X_k \mid^2}{\partial t} + a(X_k, X_k) + \frac{C}{D} \parallel \underline{u}_k \parallel^3_{L^3(\Omega)} - g(\mathrm{div}(\underline{u}_k), h_k) + (\mathrm{div}(h_k \underline{u}_k), h_k) = (\underline{F}_k, X_k),$$

where $X_k = (\underline{u}_k, h_k)$ and $\underline{F}_k = (\underline{f}_k, f_1)$.

According to Green's formula we have $(\mathrm{div}(h_k\underline{u}_k), h_k) = \frac{1}{2}(\mathrm{div}(\underline{u}_k), h_k^2)$ and therefore $\mid (\mathrm{div}(h_k\underline{u}_k), h_k) \mid \leq c \parallel \underline{u}_k \parallel \quad \parallel h_k \parallel^2_{L^4(\Omega)}$.

Applying the coercivity of the bilinear form a and Hölder's inequality, we obtain

$$\frac{1}{2}\frac{\partial \mid X_k \mid^2}{\partial t} + \alpha \parallel X_k \parallel^2 + \frac{C}{D} \parallel \underline{u}_k \parallel^3_{L^3(\Omega)} \leq c_0 \mid \underline{F}_k \mid^2 + c_1 \mid X_k \mid^2 + \frac{\alpha}{2} \parallel X_k \parallel^2 + c_2 \parallel h_k \parallel^4_{L^4(\Omega)}$$

which implies

$$\frac{\partial \mid X_k \mid^2}{\partial t} + \alpha \parallel X_k \parallel^2 \leq 2c_0 \mid \underline{F}_k \mid^2 + 2c_1 \mid X_k \mid^2 + 2c_2 \parallel h_k \parallel^4_{L^4(\Omega).} \tag{3.2}$$

Applying Gronwall's lemma now gives

$$\mid X_k(t) \mid^2 \leq c_4(1 + \parallel \underline{F}_k \parallel^2_{L^2(0,T,H)} + \parallel h_k \parallel^4_{L^4(0,T,L^4(\Omega))}) \ \forall t \in [0,T].$$

Since $(h_k, \underline{f}_k)$ is uniformly bounded in $L^4(0,T,L^4(\Omega)) \times L^2(0,T,L^2(\Omega))$ and f_1 is independent of k, we have

$$\mid X_k(t) \mid^2 \leq c_5 \ \forall t \in [0,T].$$

By integrating inequality (3.2) with respect to time, we now obtain

$$\parallel X_k(t) \parallel^2_{L^2(0,T,V)} \leq c.$$

Hence

$$X_k \text{ is uniformly bounded in } L^2(0,T,V) \cap L^\infty(0,T,H). \tag{3.3}$$

By using (2.1), the continuity of a, Hölder's inequality, and Lemma 3 we obtain

$$\left(\frac{\partial X_k}{\partial t}, \underline{w}\right) \leq c_6(\parallel X_k \parallel + \parallel X_k \parallel^2_{L^4(\Omega)} + \mid \underline{F}_k \mid) \parallel \underline{w} \parallel \tag{3.4}$$

and then, according to the continuous injection of $H^1(\Omega)$ in $L^4(\Omega)$,

$$\left(\frac{\partial X_k}{\partial t}, \underline{w}\right) \leq c_7(\parallel X_k \parallel + \parallel X_k \parallel^2 + \mid \underline{F}_k \mid) \parallel \underline{w} \parallel .$$

Since $\parallel X_k \parallel_{L^2(0,T,V)}$ and $\parallel \underline{F}_k \parallel_{L^2(0,T,H)}$ are uniformly bounded, we can conclude that

$$\frac{\partial X_k}{\partial t} \text{ is uniformly bounded in } L^1(0,T,V'). \tag{3.5}$$

Let us introduce space $Y = \{\underline{w} \in L^2(0,T,V), \partial\underline{w}/\partial t \in L^1(0,T,V')\}$. According to Temam [18], [19] the injection of Y into $L^2(0,T,H)$ is compact. We have proved that X_k is uniformly bounded in Y. $\underline{f}_k$ being uniformly bounded in U_c, we can extract from $(X_k, \underline{F}_k)$ a subsequence also denoted by $(X_k, \underline{F}_k)$ and such that

$$\begin{array}{lc} \underline{f}_k \rightharpoonup \underline{f} & \text{weakly in } U_c, \\ X_k \rightharpoonup X & \text{weakly in } L^2(0,T,V), \\ X_k \longrightarrow X & \text{strongly in } L^2(0,T,H). \end{array}$$

We now have to verify that $X = (\underline{u}, h)$ is the solution of problem (2.1), $\underline{f}$ being the external forcing. For this, we refer to a similar result occurring in the proof of Proposition 1.

Since J is weakly lower semicontinuous, we have

$$J(\underline{f}, \underline{u}) \leq \liminf_{k\to\infty} J(\underline{f}_k, \underline{u}_k),$$

and then $J(\underline{f}, \underline{u}) = \inf_{(\underline{g},\underline{v})\in U_{ad}} J(\underline{g}, \underline{v})$, which achieves the proof. □

Remark 5. *The question of the uniqueness is not treated; in general (in nonlinear problems) the answer is expected to be negative (for the same remark see, for example, Fattorini and Sritharan* [8], [9] *and Lions* [12]*).* □

In order to characterize the optimal control, we introduce a cost function J_ϵ which is a penalization of J corresponding to a linearization of problem (2.1). Then we will study the limit when ϵ tends to 0.

Let $(\underline{f}, \underline{u})$ be the solution of the control problem (3.1). Function J_ϵ is defined by

$$\begin{aligned} J_\epsilon(\underline{g}, \underline{w}) \quad &= \frac{1}{4} \parallel H(\underline{g}, \underline{w}) - h_{obs} \parallel^4_{L^4(0,T,L^4(\Omega))} + \frac{1}{2}\gamma \parallel \underline{g} \parallel^2_{U_c} \\ &\quad + \frac{1}{2\epsilon}\int_0^T a_1(\underline{U}(\underline{g}, \underline{w}) - \underline{w}, \underline{U}(\underline{g}, \underline{w}) - \underline{w})dt \\ &\quad + \frac{1}{2} \parallel \underline{U}(\underline{g}, \underline{w}) - \underline{u} \parallel^2_{L^2(0,T,H_1)} + \frac{1}{2} \parallel \underline{g} - \underline{f} \parallel^2_{U_c}, \end{aligned} \tag{3.6}$$

where $(\underline{U}(\underline{g}, \underline{w}), H(\underline{g}, \underline{w}))$ is the solution of the following problem:

$$\begin{aligned} &\left(\frac{\partial \underline{U}}{\partial t}, \underline{v}\right) + a_1(\underline{U}, \underline{v}) + \frac{C}{D}(\mid \underline{w} \mid_2 \underline{U}, \underline{v}) + (\underline{F} \wedge \underline{U}, \underline{v}) - g(\mathrm{div}(\underline{v}), H) = (\underline{g}, \underline{v}), \\ &\left(\frac{\partial H}{\partial t}, \beta\right) + a_2(H, \beta) + (\mathrm{div}(H\underline{w}), \beta) = (f_1, \beta) \ \forall(\underline{v}, \beta) \in V, \\ &(\underline{U}, H)(t = 0) = (\underline{u}_0, h_0). \end{aligned} \tag{3.7}$$

Lemma 5. *$\underline{U}(\underline{f},\underline{u})=\underline{u}$ and $H(\underline{f},\underline{u})=h$, where $(\underline{u},h)$ is the solution of problem* (2.1)*, the external forcing being equal to $\underline{f}$.*

Proof. Let $\underline{u}_1 = \underline{U}(\underline{f}, \underline{u}) - \underline{u}$ and $h_1 = H(\underline{f}, \underline{u}) - h$.

Subtracting (2.1) from (3.7) gives

$$\begin{aligned} &\left(\frac{\partial \underline{u}_1}{\partial t}, \underline{v}\right) + a_1(\underline{u}_1, \underline{v}) + \frac{C}{D}(\mid \underline{u} \mid_2 \underline{u}_1, \underline{v}) + (\underline{F} \wedge \underline{u}_1, \underline{v}) - g(\mathrm{div}(\underline{v}), h_1) = 0, \\ &\left(\frac{\partial h_1}{\partial t}, \beta\right) + a_2(h_1, \beta) + (\mathrm{div}(h_1\underline{u}), \beta) = 0 \ \forall(\underline{v}, \beta) \in V, \\ &(\underline{u}_1, h_1)(t = 0) = (\underline{0}, 0). \end{aligned} \tag{3.8}$$

Setting $(\underline{v}, \beta) = (\underline{u}_1, h_1)$ in (3.8) and summing the first and second parts of (3.8), we obtain

$$\frac{1}{2}\frac{\partial \mid Y \mid^2}{\partial t} + a(Y, Y) + \frac{C}{D}\int_\Omega \mid \underline{u} \mid_2 \quad \mid \underline{u}_1 \mid_2^2 d\Omega = -(\mathrm{div}(h_1\underline{u}), h_1) + g(\mathrm{div}(\underline{u}_1), h_1),$$

where $Y = (\underline{u}_1, h_1)$.

According to Green's formula we have $(\mathrm{div}(h_1\underline{u}), h_1) = \frac{1}{2}(\mathrm{div}(\underline{u}), h_1^2)$ and therefore

$$\mid (\mathit{div}(h_1\underline{u}), h_1) \mid \leq c \parallel \underline{u} \parallel \quad \parallel h_1 \parallel^2_{L^4(\Omega)}.$$

Using the coercivity of operator a, Gagliardo–Nirenberg's inequality and Hölder's inequality we obtain

$$\frac{\partial \mid Y \mid^2}{\partial t} + \alpha \parallel Y \parallel^2 \leq c(1 + \parallel \underline{u} \parallel^2) \mid Y \mid^2 .$$

By now applying Gronwall's lemma, we can deduce

$$\mid Y(t) \mid^2 \leq c \mid Y(0) \mid^2 \left(\exp \left(\int_0^t \parallel \underline{u} \parallel^2 \, dt \right) \right) \forall t \in [0, T].$$

Since $\mid Y(0) \mid = 0$ we have $Y = 0 \ \forall t$ and then $(\underline{u}_1, h_1) = (\underline{0}, 0)$. □

The control problem associated with direct problem (3.7) and cost function J_ϵ defined by (3.6) is then the following:

$$\text{Find } (\underline{f}_\epsilon, \underline{w}_\epsilon) \in U^\epsilon_{ad} \text{ such that} \tag{3.9}$$
$$J_\epsilon(\underline{f}_\epsilon, \underline{w}_\epsilon) = \inf_{(\underline{g}, \underline{v}) \in U^\epsilon_{ad}} J_\epsilon(\underline{g}, \underline{v}),$$

where the admissibility set is $U^\epsilon_{ad} = K_c \times L^2(0, T, V_1)$ and K_c is a convex, closed, nonempty subset of U_c.

3.1.1. Study of the penalized control problem.

In this section we prove the existence of an optimal control solution of problem (3.9). We then characterize this control by means of direct problem (3.7) and its adjoint.

Proposition 2. *For any $\epsilon > 0$, control problem* (3.9) *has at least one solution $(\underline{f}_\epsilon, \underline{w}_\epsilon)$ in U^ϵ_{ad}.*

Proof. This result is proved in the same way as in Theorem 1. □

From now on, $(\underline{f}_\epsilon, \underline{w}_\epsilon)$ denotes the optimal control solution of problem (3.9). We set $\underline{U}_\epsilon = \underline{U}(\underline{f}_\epsilon, \underline{w}_\epsilon), H_\epsilon = H(\underline{f}_\epsilon, \underline{w}_\epsilon)$. $(\underline{U}_\epsilon, H_\epsilon)$ is the solution of the following problem:

$$\tag{3.10} \begin{aligned} &\left(\frac{\partial \underline{U}_\epsilon}{\partial t}, \underline{v} \right) + a_1(\underline{U}_\epsilon, \underline{v}) + \frac{C}{D}(\mid \underline{w}_\epsilon \mid_2 \underline{U}_\epsilon, \underline{v}) + (\underline{F} \wedge \underline{U}_\epsilon, \underline{v}) - g(\operatorname{div}(\underline{v}), H_\epsilon) = (\underline{f}_\epsilon, \underline{v}), \\ &\left(\frac{\partial H_\epsilon}{\partial t}, \beta \right) + a_2(H_\epsilon, \beta) + (\operatorname{div}(H_\epsilon \underline{w}_\epsilon), \beta) = (f_1, \beta) \ \forall (\underline{v}, \beta) \in V, \\ &(\underline{U}_\epsilon, H_\epsilon)(t = 0) = (\underline{u}_0, h_0). \end{aligned}$$

We are now going to characterize the optimal control $(\underline{f}_\epsilon, \underline{w}_\epsilon)$.

Theorem 2. *Let $(\underline{f}, \underline{u})$ be a solution of control problem* (3.1) *and $(\underline{f}_\epsilon, \underline{w}_\epsilon)$ be a* *solution of control problem* (3.9)*. Then there exists* $(\underline{U}_\epsilon, H_\epsilon)$ *and* $(\underline{R}_\epsilon, S_\epsilon)$ *such that*

$$\begin{aligned}
&\left(\frac{\partial \underline{U}_\epsilon}{\partial t}, \underline{v}\right) + a_1(\underline{U}_\epsilon, \underline{v}) + \frac{C}{D}(\mid \underline{w}_\epsilon \mid_2 \underline{U}_\epsilon, \underline{v}) + (\underline{F} \wedge \underline{U}_\epsilon, \underline{v}) - g(\mathrm{div}(\underline{v}), H_\epsilon) = (\underline{f}_\epsilon, \underline{v}),\\
&\left(\frac{\partial H_\epsilon}{\partial t}, \beta\right) + a_2(H_\epsilon, \beta) + (\mathrm{div}(H_\epsilon \underline{w}_\epsilon), \beta) = (f_1, \beta)\ \forall(\underline{v}, \beta) \in V,\\
&(\underline{U}_\epsilon, H_\epsilon)(t=0) = (\underline{u}_0, h_0),
\end{aligned}$$

$$\begin{aligned}
&\left(-\frac{\partial \underline{R}_\epsilon}{\partial t}, \underline{v}\right) + a_1(\underline{R}_\epsilon, \underline{v}) + \frac{C}{D}(\mid \underline{w}_\epsilon \mid_2 \underline{R}_\epsilon + (\underline{R}_\epsilon, \underline{U}_\epsilon)_2 \frac{\underline{w}_\epsilon}{\mid \underline{w}_\epsilon \mid_2}, \underline{v}) - (\underline{F} \wedge \underline{R}_\epsilon, \underline{v})\\
&-((H_\epsilon \nabla) S_\epsilon, \underline{v}) = (\underline{U}_\epsilon - \underline{u}, \underline{v}),\\
&\left(-\frac{\partial S_\epsilon}{\partial t}, \beta\right) + a_2(S_\epsilon, \beta) + (\mathrm{div}(\beta \underline{w}_\epsilon), S_\epsilon) - g(\mathrm{div}(\underline{R}_\epsilon), \beta) = ((H_\epsilon - h_{obs})^3, \beta)\ \forall(\underline{v}, \beta) \in V,\\
&(\underline{R}_\epsilon, S_\epsilon)(t=T) = (\underline{0}, 0),
\end{aligned}$$

and
$$(\underline{R}_\epsilon + \gamma \underline{f}_\epsilon + \underline{f}_\epsilon - \underline{f},\ \underline{g} - \underline{f}_\epsilon)_{U_c} \geq 0\ \forall \underline{g} \in K_c.$$

Proof. The partial derivative of (3.7), at point $(\underline{f}_\epsilon, \underline{w}_\epsilon)$, with respect to $\underline{g}$ at constant $\underline{w}$ is

(3.11)
$$\begin{aligned}
&\left(\frac{\partial \underline{U}_g}{\partial t}, \underline{v}\right) + a_1(\underline{U}_g, \underline{v}) + \frac{C}{D}(\mid \underline{w}_\epsilon \mid_2 \underline{U}_g, \underline{v}) + (\underline{F} \wedge \underline{U}_g, \underline{v}) - g(\mathrm{div}(\underline{v}), H_g) = (\underline{g}, \underline{v}),\\
&\left(\frac{\partial H_g}{\partial t}, \beta\right) + a_2(H_g, \beta) + (\mathrm{div}(H_g \underline{w}_\epsilon), \beta) = 0\ \forall(\underline{v}, \beta) \in V,\\
&(\underline{U}_g, H_g) = (\underline{0}, 0)
\end{aligned}$$

and the partial derivative of (3.7), at point $(\underline{f}_\epsilon, \underline{w}_\epsilon)$, with respect to $\underline{w}$ at constant $\underline{g}$ is

(3.12)
$$\begin{aligned}
&\left(\frac{\partial \underline{U}_w}{\partial t}, \underline{v}\right) + a_1(\underline{U}_w, \underline{v}) + \frac{C}{D}\left(\mid \underline{w}_\epsilon \mid_2 \underline{U}_w, \underline{v}\right) + \frac{C}{D}\left((\underline{w}, \underline{w}_\epsilon)_2 \frac{\underline{U}_\epsilon}{\mid \underline{w}_\epsilon \mid_2}, \underline{v}\right) + (\underline{F} \wedge \underline{U}_w, \underline{v})\\
&-g(\mathrm{div}(\underline{v}), H_w) = 0,\\
&\left(\frac{\partial H_w}{\partial t}, \beta\right) + a_2(H_w, \beta) + (\mathrm{div}(H_w \underline{w}_\epsilon) + \mathrm{div}(H_\epsilon \underline{w}), \beta) = 0\ \forall(\underline{v}, \beta) \in V,\\
&(\underline{U}_w, H_w) = (\underline{0}, 0).
\end{aligned}$$

The existence of $(\underline{U}_\epsilon, H_\epsilon)$, the solution of the direct problem, follows from the definition of J_ϵ. On the other hand, we have to prove the existence of $(\underline{R}_\epsilon, S_\epsilon)$ verifying the adjoint problem.

First $(\underline{R}_\epsilon, S_\epsilon)$ will denote the solution of the following problem:

$$\begin{aligned}
&\left(-\frac{\partial \underline{R}_\epsilon}{\partial t}, \underline{v}\right) + a_1(\underline{R}_\epsilon, \underline{v}) + \frac{C}{D}(\mid \underline{w}_\epsilon \mid_2 \underline{R}_\epsilon, \underline{v}) - (\underline{F} \wedge \underline{R}_\epsilon, \underline{v}) - ((H_\epsilon \nabla) S_\epsilon, \underline{v}) \\
&= (\underline{U}_\epsilon - \underline{u}, \underline{v}) + \frac{1}{\epsilon} a_1(\underline{U}_\epsilon - \underline{w}_\epsilon, \underline{v}), \\
&\left(-\frac{\partial S_\epsilon}{\partial t}, \beta\right) + a_2(S_\epsilon, \beta) + (\mathrm{div}(\beta \underline{w}_\epsilon), S_\epsilon) - g(\mathrm{div}(\underline{R}_\epsilon), \beta) \\
&= ((H_\epsilon - h_{obs})^3, \beta) \ \forall (\underline{v}, \beta) \in V, \\
&(\underline{R}_\epsilon, S_\epsilon)(t = T) = (\underline{0}, 0).
\end{aligned} \tag{3.13}$$

This linear problem has one unique solution $(\underline{R}_\epsilon, S_\epsilon)$ in $L^2(0, T, V)$.

Setting $(\underline{v}, \beta) = (\underline{R}_\epsilon, S_\epsilon)$ in (3.12) and integrating with respect to time gives

$$\begin{aligned}
&\int_0^T \left(\frac{\partial \underline{U}_w}{\partial t}, \underline{R}_\epsilon\right) dt + \int_0^T a_1(\underline{U}_w, \underline{R}_\epsilon) dt + \frac{C}{D} \int_0^T (\mid \underline{w}_\epsilon \mid_2 \underline{U}_w, \underline{R}_\epsilon) dt \\
&\quad + \quad \frac{C}{D} \int_0^T \left((\underline{w}, \underline{w}_\epsilon)_2 \frac{\underline{U}_\epsilon}{\mid \underline{w}_\epsilon \mid_2}, \underline{R}_\epsilon\right) dt \\
&\quad + \int_0^T (\underline{F} \wedge \underline{U}_w, \underline{R}_\epsilon) dt - g \int_0^T (\mathrm{div}(\underline{R}_\epsilon), H_w) dt = 0, \\
&\int_0^T \left(\frac{\partial H_w}{\partial t}, S_\epsilon\right) dt + \int_0^T a_2(H_w, S_\epsilon) dt + \int_0^T (\mathrm{div}(H_w \underline{w}_\epsilon) + \mathrm{div}(H_\epsilon \underline{w}), S_\epsilon) dt = 0.
\end{aligned}$$

After integration by parts we obtain

$$\begin{aligned}
&-\int_0^T \left(\frac{\partial \underline{R}_\epsilon}{\partial t}, \underline{U}_w\right) dt + \int_0^T a_1(\underline{R}_\epsilon, \underline{U}_w) dt + \frac{C}{D} \int_0^T (\mid \underline{w}_\epsilon \mid_2 \underline{R}_\epsilon, \underline{U}_w) dt \\
&+ \frac{C}{D} \int_0^T \left((\underline{w}, \underline{w}_\epsilon)_2 \frac{\underline{U}_\epsilon}{\mid \underline{w}_\epsilon \mid_2}, \underline{R}_\epsilon\right) dt \\
&- \int_0^T (\underline{F} \wedge \underline{R}_\epsilon, \underline{U}_w) dt - g \int_0^T (\mathrm{div}(\underline{R}_\epsilon), H_w) dt = 0, \\
&- \int_0^T \left(\frac{\partial S_\epsilon}{\partial t}, H_w\right) dt + \int_0^T a_2(S_\epsilon, H_w) dt + \int_0^T (\mathrm{div}(H_w \underline{w}_\epsilon) + \mathrm{div}(H_\epsilon \underline{w}), S_\epsilon) dt = 0.
\end{aligned}$$

Since $(\underline{R}_\epsilon, S_\epsilon)$ verifies (3.13), this implies

$$\begin{aligned}
&\frac{C}{D} \int_0^T \left((\underline{w}, \underline{w}_\epsilon)_2 \frac{\underline{U}_\epsilon}{\mid \underline{w}_\epsilon \mid_2}, \underline{R}_\epsilon\right) dt - g \int_0^T (\mathrm{div}(\underline{R}_\epsilon), H_w) dt + \int_0^T (\underline{U}_\epsilon - \underline{u}, \underline{U}_w) dt \\
&\quad + \frac{1}{\epsilon} \int_0^T a_1(\underline{U}_\epsilon - \underline{w}_\epsilon, \underline{U}_w) dt = 0, \\
&\int_0^T (\mathrm{div}(H_\epsilon \underline{w}), S_\epsilon) dt + g \int_0^T (\mathrm{div}(\underline{R}_\epsilon), H_w) dt + \int_0^T ((H_\epsilon - h_{obs})^3, H_w) dt = 0;
\end{aligned}$$

hence

(3.14)

$$\begin{aligned}&g\int_0^T(\mathrm{div}(\underline{R}_\epsilon),H_w)dt=\frac{C}{D}\int_0^T\left((\underline{U}_\epsilon,\underline{R}_\epsilon)_2\frac{\underline{w}_\epsilon}{\mid\underline{w}_\epsilon\mid_2},\underline{w}\right)dt+\int_0^T(\underline{U}_\epsilon-\underline{u},\underline{U}_w)dt\\&\quad+\frac{1}{\epsilon}\int_0^T a_1(\underline{U}_\epsilon-\underline{w}_\epsilon,\underline{U}_w)dt,\\&g\int_0^T(\mathrm{div}(\underline{R}_\epsilon),H_w)dt=\int_0^T(H_\epsilon\nabla S_\epsilon,\underline{w})dt-\int_0^T((H_\epsilon-h_{obs})^3,H_w)dt.\end{aligned}$$

The partial derivative of J_ϵ, at point $(\underline{f}_\epsilon,\underline{w}_\epsilon)$, with respect to $\underline{w}$ at constant $\underline{g}$ is

$$\frac{\partial J_\epsilon}{\partial\underline{w}}(\underline{f}_\epsilon,\underline{w}_\epsilon).\underline{w}=\int_0^T((H_\epsilon-h_{obs})^3,H_w)dt+\frac{1}{\epsilon}\int_0^T a_1(\underline{U}_\epsilon-\underline{w}_\epsilon,\underline{U}_w-\underline{w})dt+\int_0^T(\underline{U}_\epsilon-\underline{u},\underline{U}_w)dt.$$

Control problem (3.9) includes no constraint concerning $\underline{w}$, so

$$\frac{\partial J_\epsilon}{\partial\underline{w}}(\underline{f}_\epsilon,\underline{w}_\epsilon).\underline{w}=0\ \forall\underline{w}\in L^2(0,T,V_1).$$

From this and (3.14) we deduce

$$\frac{1}{\epsilon}\int_0^T a_1(\underline{U}_\epsilon-\underline{w}_\epsilon,\underline{w})dt=\int_0^T\left(H_\epsilon\nabla S_\epsilon-\frac{C}{D}(\underline{U}_\epsilon,\underline{R}_\epsilon)_2\frac{\underline{w}_\epsilon}{\mid\underline{w}_\epsilon\mid_2},\underline{w}\right)dt\quad\forall\underline{w}\in L^2(0,T,V_1).$$

In particular, we have

$$\begin{gathered}\frac{1}{\epsilon}\int_0^T a_1(\underline{U}_\epsilon-\underline{w}_\epsilon,\varphi(t)\underline{v})dt=\int_0^T\left(H_\epsilon\nabla S_\epsilon-\frac{C}{D}(\underline{U}_\epsilon,\underline{R}_\epsilon)_2\frac{\underline{w}_\epsilon}{\mid\underline{w}_\epsilon\mid_2},\varphi(t)\underline{v}\right)dt\\\forall(\underline{v},\varphi)\in V_1\times\mathcal{D}(0,T).\end{gathered}$$

Thus, in the distribution sense,

$$\frac{1}{\epsilon}a_1(\underline{U}_\epsilon-\underline{w}_\epsilon,\underline{v})=\left(H_\epsilon\nabla S_\epsilon-\frac{C}{D}(\underline{U}_\epsilon,\underline{R}_\epsilon)_2\frac{\underline{w}_\epsilon}{\mid\underline{w}_\epsilon\mid_2},\underline{v}\right)\quad\forall\underline{v}\in V_1.\tag{3.15}$$

Since $(\underline{R}_\epsilon,S_\epsilon)$ is the solution of problem (3.13), we obtain by using (3.15) that $(\underline{R}_\epsilon,S_\epsilon)$ verifies the problem

(3.16)

$$\begin{aligned}&\left(-\frac{\partial\underline{R}_\epsilon}{\partial t},\underline{v}\right)+a_1(\underline{R}_\epsilon,\underline{v})+\frac{C}{D}\left(\mid\underline{w}_\epsilon\mid_2\underline{R}_\epsilon+(\underline{R}_\epsilon,\underline{U}_\epsilon)_2\frac{\underline{w}_\epsilon}{\mid\underline{w}_\epsilon\mid_2},\underline{v}\right)-(\underline{F}\wedge\underline{R}_\epsilon,\underline{v})\\&-((H_\epsilon\nabla)S_\epsilon,\underline{v})=(\underline{U}_\epsilon-\underline{u},\underline{v}),\\&\left(-\frac{\partial S_\epsilon}{\partial t},\beta\right)+a_2(S_\epsilon,\beta)+(\mathrm{div}(\beta\underline{w}_\epsilon),S_\epsilon)-g(\mathrm{div}(\underline{R}_\epsilon),\beta)=((H_\epsilon-h_{obs})^3,\beta)\quad\forall(\underline{v},\beta)\in V,\end{aligned}$$

which is the adjoint problem occuring in Theorem 2.

We now are going to prove the inequality which characterizes the optimal control $\underline{f}_\epsilon$.

Setting $(\underline{v},\beta)=(\underline{R}_\epsilon,S_\epsilon)$ in (3.11) and integrating with respect to time give

$$\int_0^T\left(\frac{\partial\underline{U}_g}{\partial t},\underline{R}_\epsilon\right)dt+\int_0^T a_1(\underline{U}_g,\underline{R}_\epsilon)dt+\int_0^T\frac{C}{D}(\mid\underline{w}_\epsilon\mid_2\underline{U}_g,\underline{R}_\epsilon)dt+\int_0^T(\underline{F}\wedge\underline{U}_g,\underline{R}_\epsilon)dt$$
$$-g\int_0^T(\mathrm{div}(\underline{R}_\epsilon),H_g)dt=\int_0^T(\underline{g},\underline{R}_\epsilon)dt,$$
$$\int_0^T\left(\frac{\partial H_g}{\partial t},S_\epsilon\right)dt+\int_0^T a_2(H_g,S_\epsilon)dt+\int_0^T(\mathrm{div}(H_g\underline{w}_\epsilon),S_\epsilon)dt=0$$

and then

$$-\int_0^T\left(\frac{\partial\underline{R}_\epsilon}{\partial t},\underline{U}_g\right)dt+\int_0^T a_1(\underline{R}_\epsilon,\underline{U}_g)dt+\int_0^T\frac{C}{D}(\mid\underline{w}_\epsilon\mid_2\underline{R}_\epsilon,\underline{U}_g)dt$$
$$-\int_0^T(\underline{F}\wedge\underline{R}_\epsilon,\underline{U}_g)dt$$
$$-g\int_0^T(\mathrm{div}(\underline{R}_\epsilon),H_g)dt=\int_0^T(\underline{g},\underline{R}_\epsilon)dt,$$
$$-\int_0^T\left(\frac{\partial S_\epsilon}{\partial t},H_g\right)dt+\int_0^T a_2(S_\epsilon,H_g)dt+\int_0^T(\mathrm{div}(H_g\underline{w}_\epsilon),S_\epsilon)dt=0.$$

Since $(\underline{R}_\epsilon,S_\epsilon)$ is the solution of problem (3.16), we obtain

$$(3.17)\qquad\begin{aligned}&-\frac{C}{D}\int_0^T\left((\underline{R}_\epsilon,\underline{U}_\epsilon)_2\frac{\underline{w}_\epsilon}{\mid\underline{w}_\epsilon\mid_2},\underline{U}_g\right)dt-g\int_0^T(\mathrm{div}(\underline{R}_\epsilon),H_g)dt\\&+\int_0^T(\underline{U}_\epsilon-\underline{u},\underline{U}_g)dt+\int_0^T(H_\epsilon\nabla S_\epsilon,\underline{U}_g)dt=\int_0^T(\underline{g},\underline{R}_\epsilon)dt,\\&g\int_0^T(\mathrm{div}(\underline{R}_\epsilon),H_g)dt+\int_0^T((H_\epsilon-h_{obs})^3,H_g)dt=0.\end{aligned}$$

The partial derivative of J_ϵ, at point $(\underline{f}_\epsilon,\underline{w}_\epsilon)$, with respect to $\underline{g}$ at constant $\underline{w}$ is

$$\begin{aligned}\frac{\partial J_\epsilon}{\partial\underline{g}}(\underline{f}_\epsilon,\underline{w}_\epsilon).\underline{g}\;&=\;\int_0^T((H_\epsilon-h_{obs})^3,H_g)dt+\gamma(\underline{f}_\epsilon,\underline{g})_{U_c}+(\underline{f}_\epsilon-\underline{f},\underline{g})_{U_c}\\&\quad+\frac{1}{\epsilon}\int_0^T a_1(\underline{U}_\epsilon-\underline{w}_\epsilon,\underline{U}_g)dt+\int_0^T(\underline{U}_\epsilon-\underline{u},\underline{U}_g)dt.\end{aligned}$$

Applying relationship (3.15), we deduce

$$\frac{\partial J_\epsilon}{\partial\underline{g}}(\underline{f}_\epsilon,\underline{w}_\epsilon).\underline{g}=\int_0^T((H_\epsilon-h_{obs})^3,H_g)dt+\gamma(\underline{f}_\epsilon,\underline{g})_{U_c}+(\underline{f}_\epsilon-\underline{f},\underline{g})_{U_c}$$
$$+\int_0^T\left(H_\epsilon\nabla S_\epsilon-\frac{C}{D}(\underline{U}_\epsilon,\underline{R}_\epsilon)_2\frac{\underline{w}_\epsilon}{\mid\underline{w}_\epsilon\mid_2},\underline{U}_g\right)dt+\int_0^T(\underline{U}_\epsilon-\underline{u},\underline{U}_g)dt.$$

$(\underline{R}_\epsilon,S_\epsilon)$ verifies equations (3.17). We then obtain

$$\frac{\partial J_\epsilon}{\partial\underline{g}}(\underline{f}_\epsilon,\underline{w}_\epsilon).\underline{g}=\gamma(\underline{f}_\epsilon,\underline{g})_{U_c}+(\underline{f}_\epsilon-\underline{f},\underline{g})_{U_c}+\int_0^T(\underline{g},\underline{R}_\epsilon)dt.$$

Since $U_c = L^2(0,T,L^2(\Omega))$ we have

$$\frac{\partial J_\epsilon}{\partial \underline{g}}(\underline{f}_\epsilon, \underline{w}_\epsilon).\underline{g} = \int_0^T (\underline{R}_\epsilon + \gamma \underline{f}_\epsilon + \underline{f}_\epsilon - \underline{f}, \underline{g})dt.$$

Control problem (3.9) includes a constraint on $\underline{g}$: $\underline{g} \in K_c \subset U_c$. Therefore the optimal control $\underline{f}_\epsilon$ has to verify the inequality

$$\frac{\partial J_\epsilon}{\partial \underline{g}}(\underline{f}_\epsilon, \underline{w}_\epsilon).(\underline{g} - \underline{f}_\epsilon) \geq 0 \quad \forall \underline{g} \in K_c$$

and we have

$$\int_0^T (\underline{R}_\epsilon + \gamma \underline{f}_\epsilon + \underline{f}_\epsilon - \underline{f}, \underline{g} - \underline{f}_\epsilon)dt \geq 0 \quad \forall \underline{g} \in K_c.$$

This achieves the proof of Theorem 2. □

3.1.2. Characterization of the optimal control. The penalized control problem (3.9) has a solution which is characterized by means of the direct and adjoint penalized equations.

We are now going to characterize the optimal control $(\underline{f}, \underline{u})$, the solution of problem (3.1). $(\underline{u}, h)$ denotes the solution of direct problem (2.1). The adjoint problem is obtained by passing to the limit in the adjoint penalized equations introduced in Theorem 2.

THEOREM 3. *Let* $(\underline{f}_\epsilon, \underline{w}_\epsilon)$ *be a solution of control problem* (3.9). *The following convergence results are verified as* $\epsilon \longrightarrow 0$:

- $\underline{f}_\epsilon$ *converges to* $\underline{f}$ *in* U_c *strongly,*
- $\lim_{\epsilon \longrightarrow 0} \frac{1}{\epsilon} \parallel \underline{U}_\epsilon - \underline{w}_\epsilon \parallel^2_{L^2(0,T,H^1(\Omega))} = 0,$
- $\underline{U}_\epsilon \rightharpoonup \underline{u}$ *weakly in* $L^2(0,T,H^1(\Omega))$,
- $\underline{U}_\epsilon \longrightarrow \underline{u}$ *strongly in* $L^2(0,T,L^2(\Omega))$,
- $H_\epsilon \rightharpoonup h$ *weakly in* $L^2(0,T,H^1(\Omega))$,
- $H_\epsilon \longrightarrow h$ *strongly in* $L^2(0,T,L^2(\Omega))$,
- $\underline{w}_\epsilon \rightharpoonup \underline{u}$ *weakly in* $L^2(0,T,H^1(\Omega))$,
- $J_\epsilon(\underline{w}_\epsilon, \underline{f}_\epsilon)$ *converges to* $J(\underline{u}, \underline{f})$,

Proof. $(\underline{f}, \underline{u}) \in U_{ad}$ denotes a solution of control problem (3.1). $(\underline{f}_\epsilon, \underline{w}_\epsilon) \in U^\epsilon_{ad}$ denotes a solution of the penalized control problem (3.9). Since $J_\epsilon(\underline{f}, \underline{u}) = J(\underline{f}, \underline{u})$ and $U_{ad} \subset U^\epsilon_{ad}$, we have

$$J_\epsilon(\underline{f}_\epsilon, \underline{w}_\epsilon) \leq J(\underline{f}, \underline{u}).$$

According to the definition of J_ϵ, this implies that

$$\begin{array}{ll} & \parallel \underline{f}_\epsilon - \underline{f} \parallel^2_{U_c} \leq 2J(\underline{f}, \underline{u}), \\ (3.18) & \displaystyle\int_0^T a_1(\underline{U}_\epsilon - \underline{w}_\epsilon, \underline{U}_\epsilon - \underline{w}_\epsilon)dt \leq 2\epsilon J(\underline{f}, \underline{u}), \\ & \parallel \underline{U}_\epsilon - \underline{u} \parallel^2_{L^2(0,T,H_1)} \leq 2J(\underline{f}, \underline{u}), \\ & \parallel H_\epsilon - h_{obs} \parallel^4_{L^4(0,T,L^4(\Omega))} \leq 4J(\underline{f}, \underline{u}). \end{array}$$

Using the coercivity of a, we deduce from (3.18) the following results:

$$\begin{array}{ll} & \parallel \underline{f}_\epsilon \parallel^2_{U_c} \leq 2 \parallel \underline{f} \parallel^2_{U_c} + 4J(\underline{u}, \underline{f}), \\ (3.19) & \displaystyle\lim_{\epsilon \rightarrow 0} \parallel \underline{U}_\epsilon - \underline{w}_\epsilon \parallel_{L^2(0,T,H^1(\Omega))} = 0, \\ & \parallel \underline{U}_\epsilon \parallel_{L^2(0,T,H_1)} \leq c, \\ & \parallel H_\epsilon \parallel_{L^4(0,T,L^4(\Omega))} \leq c, \end{array}$$

where c is a positive constant.

Equation (3.19) implies that $\underline{U}_\epsilon$ converges towards $\underline{w}_\epsilon$ strongly in $L^2(0,T,H^1(\Omega))$ and that $(\underline{U}_\epsilon, H_\epsilon)$ is uniformly bounded in $L^2(0,T,L^2(\Omega)) \times L^4(0,T,L^4(\Omega))$. For ϵ sufficiently small, $\| \underline{U}_\epsilon - \underline{w}_\epsilon \|_{L^2(0,T,H^1(\Omega))}$ is uniformly bounded. Therefore there exists a positive constant $c > 0$ independent of ϵ such that

$$\| \underline{w}_\epsilon \|_{L^2(0,T,H^1(\Omega))} \leq c(1 + \| \underline{U}_\epsilon \|_{L^2(0,T,H^1(\Omega))})$$

and

$$\| \underline{w}_\epsilon \|_{L^2(0,T,L^2(\Omega))} \leq c(1 + \| \underline{U}_\epsilon \|_{L^2(0,T,L^2(\Omega))}).$$

Since $\underline{U}_\epsilon$ is uniformly bounded in $L^2(0,T,L^2(\Omega))$, we have

$$\| \underline{w}_\epsilon \|_{L^2(0,T,L^2(\Omega))} \leq c. \tag{3.20}$$

We can now pass to the limit in direct problem (3.10).

Setting $(\underline{v}, \beta) = (\underline{U}_\epsilon, H_\epsilon)$ in (3.10) and adding the first and second parts of (3.10) gives

$$\begin{aligned} \frac{1}{2}\frac{\partial \mid X_\epsilon \mid^2}{\partial t} + a(X_\epsilon, X_\epsilon) + \frac{C}{D}\int_\Omega \mid \underline{w}_\epsilon \mid_2 \mid \underline{U}_\epsilon \mid_2^2 \, d\Omega = (\underline{F}_\epsilon, X_\epsilon) + g(\operatorname{div}(\underline{U}_\epsilon), H_\epsilon) \\ - (\operatorname{div}(H_\epsilon \underline{w}_\epsilon), H_\epsilon), \end{aligned}$$

where $X_\epsilon = (\underline{U}_\epsilon, H_\epsilon)$ et $\underline{F}_\epsilon = (\underline{f}_\epsilon, f_1)$.

According to Green's formula we have $(\operatorname{div}(H_\epsilon \underline{w}_\epsilon), H_\epsilon) = \frac{1}{2}(\operatorname{div}(\underline{w}_\epsilon), H_\epsilon^2)$ and therefore

$$\mid (\operatorname{div}(H_\epsilon \underline{w}_\epsilon), H_\epsilon) \mid \leq c(\| \underline{U}_\epsilon - \underline{w}_\epsilon \| + \| \underline{U}_\epsilon \|) \ \| H_\epsilon \|_{L^4(\Omega)}^2 .$$

Applying the coercivity of a, the positivity of $\int_\Omega \mid \underline{w}_\epsilon \mid_2 \mid \underline{U}_\epsilon \mid_2^2 \, d\Omega$, and Hölder's inequality we obtain

$$\frac{\partial \mid X_\epsilon \mid^2}{\partial t} + \alpha \| X_\epsilon \|^2 \leq c_2 \mid \underline{F}_\epsilon \mid^2 + c_3 \mid X_\epsilon \mid^2 + c_4(\| H_\epsilon \|_{L^4(\Omega)}^4 + \| \underline{U}_\epsilon - \underline{w}_\epsilon \|^2). \tag{3.21}$$

Gronwall's lemma now gives

$$\mid X_\epsilon \mid^2 \leq c(1 + \| \underline{F}_\epsilon \|_{L^2(0,T,L^2(\Omega))}^2 + \| H_\epsilon \|_{L^4(0,T,L^4(\Omega))}^4 + \| \underline{U}_\epsilon - \underline{w}_\epsilon \|_{L^2(0,T,H^1(\Omega))}^2).$$

According to (3.19) and (3.21), we deduce that sequence X_ϵ is uniformly bounded in $L^2(0,T,V) \cap L^\infty(0,T,H)$; $(\underline{f}_\epsilon, \underline{w}_\epsilon)$ being uniformly bounded in $U_c \times L^2(0,T,H^1(\Omega))$, we can extract from $(\underline{f}_\epsilon, X_\epsilon, \underline{w}_\epsilon)$ a subsequence also denoted by $(\underline{f}_\epsilon, X_\epsilon, \underline{w}_\epsilon)$, which verifies

$$\begin{array}{rcl} \underline{f}_\epsilon \rightharpoonup \underline{f}_0 & & \text{weakly in } U_c, \\ X_\epsilon \rightharpoonup X_0 & & \text{weakly in } L^2(0,T,H^1(\Omega)), \\ X_\epsilon \longrightarrow X_0 & & \text{strongly in } L^2(0,T,L^2(\Omega)), \\ \underline{w}_\epsilon \rightharpoonup \underline{U}_0 & & \text{weakly in } L^2(0,T,H^1(\Omega)). \end{array} \tag{3.22}$$

We now have to verify that $X_0 = (\underline{U}_0, H_0)$ is the solution of (2.1), $\underline{f}_0$ being the external forcing. For this, we refer to a similar result occurring in the proof of Proposition 1.

Now applying the property of weak lower semicontinuity of J, we obtain [6]

(3.23)

$J(\underline{f}_0, \underline{U}_0) \leq \liminf_{\epsilon \to 0} J(\underline{f}_\epsilon, \underline{U}_\epsilon)$

and

$\frac{1}{2}(\| \underline{U}_0 - \underline{u} \|^2_{L^2(0,T,L^2(\Omega))} + \| \underline{f}_0 - \underline{f} \|^2_{U_c})$

$\leq \liminf_{\epsilon \to 0} \frac{1}{2}(\| \underline{U}_\epsilon - \underline{u} \|^2_{L^2(0,T,L^2(\Omega))} + \| \underline{f}_\epsilon - \underline{f} \|^2_{U_c})$.

From (3.23), we deduce

$$J(\underline{f}_0, \underline{U}_0) \leq \liminf_{\epsilon \to 0} J_\epsilon(\underline{f}_\epsilon, \underline{w}_\epsilon).$$

Since $(\underline{f}_\epsilon, \underline{w}_\epsilon)$ (resp., $(\underline{f}, \underline{u})$) is the optimal control solution of (3.1) (resp., (3.9)), we thus have

$$J(\underline{f}_0, \underline{U}_0) \leq \liminf_{\epsilon \to 0} J_\epsilon(\underline{f}_\epsilon, \underline{w}_\epsilon) \leq J(\underline{f}, \underline{u}) \leq J(\underline{f}_0, \underline{U}_0).$$

We conclude that

$$J(\underline{f}_0, \underline{U}_0) = J(\underline{f}, \underline{u});$$

$(\underline{f}_0, \underline{U}_0)$ is a solution of control problem (3.1).

Applying Lemma 5, we also have

$$J(\underline{f}, \underline{u}) \leq \liminf_{\epsilon \to 0} J_\epsilon(\underline{f}_\epsilon, \underline{w}_\epsilon) \leq \limsup_{\epsilon \to 0} J_\epsilon(\underline{f}_\epsilon, \underline{w}_\epsilon) \leq J(\underline{f}, \underline{u}),$$

which implies

$$\lim_{\epsilon \to 0} J_\epsilon(\underline{f}_\epsilon, \underline{w}_\epsilon) = J(\underline{f}, \underline{u}) = J(\underline{f}_0, \underline{U}_0).$$

$\underline{U}_\epsilon$ converges towards $\underline{U}_0$ in $L^2(0, T, L^2(\Omega))$ strongly; therefore

$$\begin{aligned} 0 \leq \limsup_{\epsilon \to 0} &\left(\frac{1}{2} \left(\frac{1}{\epsilon} \int_0^T a_1(\underline{U}_\epsilon - \underline{w}_\epsilon - \underline{U}_\epsilon - \underline{w}_\epsilon) dt + \| \underline{f}_\epsilon - \underline{f} \|^2_{U_c} \right) \right) \\ &+ \frac{1}{2} \| \underline{U}_0 - \underline{u} \|^2_{L^2(0,T,L^2(\Omega))} \\ &\leq \limsup_{\epsilon \to 0} J_\epsilon(\underline{f}_\epsilon, \underline{w}_\epsilon) - \liminf_{\epsilon \to 0} J(\underline{f}_\epsilon, \underline{U}_\epsilon). \end{aligned}$$

According to (3.23) this implies

$$\begin{aligned} &\lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_0^T a_1(\underline{U}_\epsilon - \underline{w}_\epsilon, \underline{U}_\epsilon - \underline{w}_\epsilon) dt = 0, \\ &\lim_{\epsilon \to 0} \| \underline{f}_\epsilon - \underline{f} \|^2_{U_c}, \\ &(\underline{U}_0, H_0, \underline{f}_0) = (\underline{u}, h, \underline{f}). \qquad \square \end{aligned}$$

These convergence results will allow us to pass to the limit in the adjoint penalized problem and thus to characterize the optimal control $(\underline{f}, \underline{u})$.

THEOREM 4. *Let $(\underline{f}, \underline{u})$ be a solution of control problem* (3.1) *and* $(\underline{u}, h)$ *be the solution of direct problem* (2.1). *Then there exists* $(\underline{R}, S) \in L^2(0, T, V)$ *such that*

$$\left(-\frac{\partial \underline{R}}{\partial t}, \underline{v}\right) + a_1(\underline{R}, \underline{v}) + \frac{C}{D}\left(\mid \underline{u} \mid_2 \underline{R} + (\underline{R}, \underline{u})_2 \frac{\underline{u}}{\mid \underline{u} \mid_2}, \underline{v}\right) - (\underline{F} \wedge \underline{R}, \underline{v}) - (h\nabla S, \underline{v}) = 0,$$
$$\left(-\frac{\partial S}{\partial t}, \beta\right) + a_2(S, \beta) - ((\underline{u}\nabla)S, \beta) - g(\mathrm{div}(\underline{R}), \beta) = ((h - h_{obs})^3, \beta) \quad \forall(\underline{v}, \beta) \in V,$$
$$(\underline{R}, S)(t = T) = (\underline{0}, 0)$$
and
$$(\underline{R} + \gamma\underline{f},\ \underline{g} - \underline{f})_{U_c} \geq 0\ \forall \underline{g} \in K_c.$$

Proof. $(\underline{R}_\epsilon, S_\epsilon)$ being the solution of the adjoint penalized problem (3.16), we introduce the sequence d_ϵ by setting

$$d_\epsilon = \parallel S_\epsilon \parallel^{-1}_{L^4(0,T,L^4(\Omega))} \text{ if } S_\epsilon \text{ is not uniformly bounded in } L^4(0, T, L^4(\Omega)),$$
$$d_\epsilon = 1 \text{ if } S_\epsilon \text{ is uniformly bounded in } L^4(0, T, L^4(\Omega)).$$

REMARK 6. *Sequence d_ϵ converges towards d with $d = 0$ or 1.*

We multiply (3.16) by d_ϵ and we thus obtain a couple $(\widetilde{\underline{R}_\epsilon}, \widetilde{S}_\epsilon)$ verifying

(3.24)
$$\left(-\frac{\partial \widetilde{\underline{R}_\epsilon}}{\partial t}, \underline{v}\right) + a_1(\widetilde{\underline{R}_\epsilon}, \underline{v}) + \frac{C}{D}\left(\mid \underline{w}_\epsilon \mid_2 \widetilde{\underline{R}_\epsilon} + (\widetilde{\underline{R}_\epsilon}, \underline{U}_\epsilon)_2 \frac{\underline{w}_\epsilon}{\mid \underline{w}_\epsilon \mid_2}, \underline{v}\right) - (\underline{F} \wedge \widetilde{\underline{R}_\epsilon}, \underline{v})$$
$$-((H_\epsilon\nabla)\widetilde{S}_\epsilon, \underline{v}) = d_\epsilon(\underline{U}_\epsilon - \underline{u}, \underline{v}),$$
$$\left(-\frac{\partial \widetilde{S}_\epsilon}{\partial t}, \beta\right) + a_2(\widetilde{S}_\epsilon, \beta) + (\mathrm{div}(\beta\underline{w}_\epsilon), \widetilde{S}_\epsilon) - g(\mathrm{div}(\widetilde{\underline{R}_\epsilon}), \beta) = d_\epsilon((H_\epsilon - h_{obs})^3, \beta)$$
$$\forall(\underline{v}, \beta) \in V,$$
$$(\widetilde{\underline{R}_\epsilon}, \widetilde{S}_\epsilon)(t = T) = (\underline{0}, 0)$$
and the inequality
$$(\widetilde{\underline{R}_\epsilon} + d_\epsilon(\gamma\underline{f}_\epsilon + \underline{f}_\epsilon - \underline{f}), \underline{g} - \underline{f}_\epsilon)_{U_c} \geq 0 \quad \forall\underline{g} \in K_c.$$

Setting $(\underline{v}, \beta) = (\widetilde{\underline{R}_\epsilon}, \widetilde{S}_\epsilon)$ in (3.24), we obtain

$$-\frac{1}{2}\frac{\partial \mid \widetilde{\underline{R}_\epsilon} \mid^2}{\partial t} + a_1(\widetilde{\underline{R}_\epsilon}, \widetilde{\underline{R}_\epsilon}) + \frac{C}{D}\left(\mid \underline{w}_\epsilon \mid_2 \widetilde{\underline{R}_\epsilon} + (\widetilde{\underline{R}_\epsilon}, \underline{U}_\epsilon)_2 \frac{\underline{w}_\epsilon}{\mid \underline{w}_\epsilon \mid_2}, \widetilde{\underline{R}_\epsilon}\right) - ((H_\epsilon\nabla)\widetilde{S}_\epsilon, \widetilde{\underline{R}_\epsilon})$$
$$= d_\epsilon(\underline{U}_\epsilon - \underline{u}, \widetilde{\underline{R}_\epsilon}),$$
$$-\frac{1}{2}\frac{\partial \mid \widetilde{S}_\epsilon \mid^2}{\partial t} + a_2(\widetilde{S}_\epsilon, \widetilde{S}_\epsilon) + (\mathrm{div}(\widetilde{S}_\epsilon\underline{w}_\epsilon), \widetilde{S}_\epsilon) - g(\mathrm{div}(\widetilde{\underline{R}_\epsilon}), \widetilde{S}_\epsilon) = d_\epsilon((H_\epsilon - h_{obs})^3, \widetilde{S}_\epsilon).$$

Adding these two equations gives

$$-\frac{1}{2}\frac{\partial \mid Y_\epsilon \mid^2}{\partial t} + a(Y_\epsilon, Y_\epsilon) + \frac{C}{D}\int_\Omega \mid \underline{w}_\epsilon \mid_2 \quad \mid \widetilde{\underline{R}_\epsilon} \mid_2^2 d\Omega + \frac{C}{D}\int_\Omega (\widetilde{\underline{R}_\epsilon}, \underline{U}_\epsilon)_2 \frac{(\underline{w}_\epsilon, \widetilde{\underline{R}_\epsilon})_2}{\mid \underline{w}_\epsilon \mid_2} d\Omega$$
$$-((H_\epsilon\nabla)\widetilde{S}_\epsilon, \widetilde{\underline{R}_\epsilon}) + (\mathrm{div}(\widetilde{S}_\epsilon\underline{w}_\epsilon), \widetilde{S}_\epsilon) - g(\mathrm{div}(\widetilde{\underline{R}_\epsilon}), \widetilde{S}_\epsilon)$$
$$= d_\epsilon(\underline{U}_\epsilon - \underline{u}, \widetilde{\underline{R}_\epsilon}) + d_\epsilon((H_\epsilon - h_{obs})^3, \widetilde{S}_\epsilon),$$

where $Y_\epsilon = (\widetilde{\underline{R}_\epsilon}, \widetilde{S}_\epsilon)$.

Hölder's inequality implies that

$$d_\epsilon((H_\epsilon - h_{obs})^3, \widetilde{S}_\epsilon) \leq c_0(d_\epsilon^{4/3} \parallel H_\epsilon - h_{obs} \parallel^4_{L^4(\Omega)} + \parallel \widetilde{S}_\epsilon \parallel^4_{L^4(\Omega)}).$$

According to Green's formula we have

$$2(\text{div}(\widetilde{S}_\epsilon \underline{w}_\epsilon), \widetilde{S}_\epsilon) = (\widetilde{S}_\epsilon^{\ 2}, \text{div}(\underline{w}_\epsilon)) \text{ and } \mid 2(\text{div}(\widetilde{S}_\epsilon \underline{w}_\epsilon), \widetilde{S}_\epsilon) \mid \leq c \parallel \widetilde{S}_\epsilon \parallel^2_{L^4(\Omega)} \parallel \underline{w}_\epsilon \parallel.$$

Now applying the coercivity of a and Lemma 3 we obtain

$$\begin{aligned} -\frac{1}{2}\frac{\partial \mid Y_\epsilon \mid^2}{\partial t} + \alpha \parallel Y_\epsilon \parallel^2 + \frac{C}{D}\int_\Omega \mid \underline{w}_\epsilon \mid_2 \quad \mid \widetilde{\underline{R}_\epsilon} \mid_2^2 \, d\Omega \\ \leq \frac{C}{D}\int_\Omega \mid \underline{U}_\epsilon \mid_2 \quad \mid \widetilde{\underline{R}_\epsilon} \mid_2^2 \, d\Omega + c_1 \parallel \widetilde{\underline{R}_\epsilon} \parallel \mid \widetilde{S}_\epsilon \mid \\ + c_2(\parallel H_\epsilon \parallel_{L^4(\Omega)} \parallel \widetilde{\underline{R}_\epsilon} \parallel_{L^4(\Omega)} \parallel \widetilde{S}_\epsilon \parallel + \parallel \widetilde{S}_\epsilon \parallel^2_{L^4(\Omega)} \parallel \underline{w}_\epsilon \parallel) \\ + d_\epsilon \mid \underline{U}_\epsilon - \underline{u} \mid \mid \widetilde{\underline{R}_\epsilon} \mid + c_0(d_\epsilon^{4/3} \parallel H_\epsilon - h_{obs} \parallel^4_{L^4(\Omega)} + \parallel \widetilde{S}_\epsilon \parallel^4_{L^4(\Omega)}), \end{aligned}$$

which implies

$$\begin{aligned} -\frac{1}{2}\frac{\partial \mid Y_\epsilon \mid^2}{\partial t} + \alpha \parallel Y_\epsilon \parallel^2 + \frac{C}{D}\int_\Omega \mid \underline{w}_\epsilon \mid_2 \quad \mid \widetilde{\underline{R}_\epsilon} \mid_2^2 \, d\Omega \leq c_3 \mid \underline{U}_\epsilon \mid \parallel Y_\epsilon \parallel^2_{L^4(\Omega)} \\ + c_4(\parallel H_\epsilon \parallel_{L^4(\Omega)} \parallel Y_\epsilon \parallel_{L^4(\Omega)} \parallel Y_\epsilon \parallel + \parallel \underline{w}_\epsilon \parallel \parallel \widetilde{S}_\epsilon \parallel^2_{L^4(\Omega)}) + c_5 \parallel Y_\epsilon \parallel \mid Y_\epsilon \mid \\ + c_6 d_\epsilon \mid \underline{U}_\epsilon - \underline{u} \mid \mid Y_\epsilon \mid + c_0(d_\epsilon^{4/3} \parallel H_\epsilon - h_{obs} \parallel^4_{L^4(\Omega)} + \parallel \widetilde{S}_\epsilon \parallel^4_{L^4(\Omega)}). \end{aligned}$$

Applying Gagliardo–Nirenberg's inequality, Hölder's inequality, and the boundedness of $\underline{U}_\epsilon - \underline{u}$ and $\underline{U}_\epsilon$ in $L^\infty(0, T, H)$, we obtain

$$\begin{aligned} -\frac{\partial \mid Y_\epsilon \mid^2}{\partial t} + \alpha \parallel Y_\epsilon \parallel^2 + \frac{2C}{D}\int_\Omega \mid \underline{w}_\epsilon \mid_2 \quad \mid \widetilde{\underline{R}_\epsilon} \mid_2^2 \, d\Omega \leq c_7 \mid Y_\epsilon \mid^2 + c_8 \parallel H_\epsilon \parallel^4_{L^4(\Omega)} \\ + c_9 \parallel \underline{w}_\epsilon \parallel^2 + c_{10} d_\epsilon \mid Y_\epsilon \mid + c_{11}(d_\epsilon^{4/3} \parallel H_\epsilon - h_{obs} \parallel^4_{L^4(\Omega)} + \parallel \widetilde{S}_\epsilon \parallel^4_{L^4(\Omega)}). \end{aligned} \tag{3.25}$$

In particular, we have

$$\begin{aligned} -\frac{\partial \mid Y_\epsilon \mid^2}{\partial t} \leq c_{12} \mid Y_\epsilon \mid^2 + c_8 \parallel H_\epsilon \parallel^4_{L^4(\Omega)} + c_9 \parallel \underline{w}_\epsilon \parallel^2 + c_{13} d_\epsilon^2 \\ + c_{10}(d_\epsilon^{4/3} \parallel H_\epsilon - h_{obs} \parallel^4_{L^4(\Omega)} + \parallel \widetilde{S}_\epsilon \parallel^4_{L^4(\Omega)}). \end{aligned}$$

Integrating with respect to time gives

$$\begin{aligned} \mid Y_\epsilon \mid^2 \leq c_{11}\int_t^T \mid Y_\epsilon \mid^2 \, dt + c_8 \parallel H_\epsilon \parallel^4_{L^4(0,T,L^4(\Omega))} + c_9 \parallel \underline{w}_\epsilon \parallel^2_{L^2(0,T,H^1(\Omega))} + c_{12} T d_\epsilon^2 \\ + c_{10}(d_\epsilon^{4/3} \parallel H_\epsilon - h_{obs} \parallel^4_{L^4(0,T,L^4(\Omega))} + \parallel \widetilde{S}_\epsilon \parallel^4_{L^4(0,T,L^4(\Omega))}). \end{aligned}$$

Since $H_\epsilon - h_{obs}$ is uniformly bounded in $L^4(0, T, L^4(\Omega))$, H_ϵ in $L^4(0, T, L^4(\Omega))$, and $\underline{w}_\epsilon$ in $L^2(0, T, H^1(\Omega))$, then by Gagliardo–Nirenberg's inequality we deduce

$$\mid Y_\epsilon \mid^2 \leq c_{11}\int_t^T \mid Y_\epsilon \mid^2 \, dt + c_{13}(1 + d_\epsilon^{4/3} + d_\epsilon^2 + \parallel \widetilde{S}_\epsilon \parallel^4_{L^4(0,T,L^4(\Omega))}).$$

Since d_ϵ is bounded and $\widetilde{S}_\epsilon$ is uniformly bounded in $L^4(0,T,L^4(\Omega))$, by using Gronwall's lemma we prove that Y_ϵ is uniformly bounded in $L^\infty(0,T,L^2(\Omega))$.

From (3.25) we deduce that

$$Y_\epsilon \text{ is uniformly bounded in } L^2(0,T,H^1(\Omega)) \cap L^\infty(0,T,L^2(\Omega)).$$

Using (3.24), we have

$$-\left(\frac{\partial Y_\epsilon}{\partial t}, W\right) = -a(Y_\epsilon, W) - \frac{C}{D}\int_\Omega \mid \underline{w}_\epsilon \mid_2 \ (\underline{\widetilde{R}_\epsilon}, \underline{v})_2 d\Omega - \frac{C}{D}\int_\Omega (\underline{\widetilde{R}_\epsilon}, \underline{U}_\epsilon)_2 \frac{(\underline{w}_\epsilon, \underline{v})_2}{\mid \underline{w}_\epsilon \mid_2} d\Omega$$
$$+((H_\epsilon \nabla)\widetilde{S}_\epsilon, \underline{v}) + ((\widetilde{S}_\epsilon \nabla)\beta, \underline{w}_\epsilon) + g(\mathrm{div}(\underline{\widetilde{R}_\epsilon}), \beta) + d_\epsilon(\underline{U}_\epsilon - \underline{u}, \underline{v}) + d_\epsilon((H_\epsilon - h_{obs})^3, \beta),$$

where $W = (\underline{v}, \beta)$.

According to the continuity of the bilinear form a, Lemma 3, and Hölder's inequality, we obtain

$$\left|\left(\frac{\partial Y_\epsilon}{\partial t}, W\right)\right| \leq c(\parallel Y_\epsilon \parallel + \parallel \underline{w}_\epsilon \parallel_{L^4(\Omega)} \parallel Y_\epsilon \parallel_{L^4(\Omega)} + \parallel \underline{U}_\epsilon \parallel_{L^4(\Omega)} \parallel Y_\epsilon \parallel_{L^4(\Omega)}$$
$$+ \parallel H_\epsilon \parallel_{L^4(\Omega)} \parallel Y_\epsilon \parallel + d_\epsilon \mid \underline{U}_\epsilon - \underline{u} \mid + d_\epsilon \parallel H_\epsilon - h_{obs} \parallel^3_{L^4(\Omega)}) \parallel W \parallel,$$

which implies

$$\left\|\frac{\partial Y_\epsilon}{\partial t}\right\|_{V'} \leq c(\parallel Y_\epsilon \parallel + \parallel \underline{w}_\epsilon \parallel_{L^4(\Omega)} \parallel Y_\epsilon \parallel_{L^4(\Omega)} + \parallel \underline{U}_\epsilon \parallel_{L^4(\Omega)} \parallel Y_\epsilon \parallel_{L^4(\Omega)}$$
$$+ \parallel H_\epsilon \parallel_{L^4(\Omega)} \parallel Y_\epsilon \parallel + d_\epsilon \mid \underline{U}_\epsilon - \underline{u} \mid + d_\epsilon \parallel H_\epsilon - h_{obs} \parallel^3_{L^4(\Omega)}).$$

Integrating with respect to time and using Hölder's inequality, we obtain

$$\left\|\frac{\partial Y_\epsilon}{\partial t}\right\|_{L^1(0,T,V')} \leq c(1 + \parallel Y_\epsilon \parallel^2_{L^2(0,T,H^1(\Omega))} + \parallel \underline{w}_\epsilon \parallel^2_{L^2(0,T,H^1(\Omega))} + \parallel \underline{U}_\epsilon \parallel^2_{L^2(0,T,H^1(\Omega))}$$
$$+ \parallel H_\epsilon \parallel^2_{L^2(0,T,H^1(\Omega))} + d^2_\epsilon + \parallel \underline{U}_\epsilon - \underline{u} \parallel^2_{L^2(0,T,L^2(\Omega))} + d^4_\epsilon + \parallel H_\epsilon - h_{obs} \parallel^4_{L^4(0,T,L^4(\Omega))}).$$

Applying (3.18), (3.19), and the boundedness of Y_ϵ,$\underline{U}_\epsilon$,H_ϵ and $\underline{w}_\epsilon$ in $L^2(0,T,H^1(\Omega))$, we deduce that $\partial Y_\epsilon / \partial t$ is uniformly bounded in $L^1(0,T,V')$.

We can then extract from (Y_ϵ) a subsequence also denoted (Y_ϵ) and such that

$$\begin{aligned} Y_\epsilon &\rightharpoonup Y \quad \text{weakly in } L^2(0,T,H^1(\Omega)), \\ Y_\epsilon &\longrightarrow Y \quad \text{strongly in } L^2(0,T,L^2(\Omega)). \end{aligned}$$

Since $(\underline{F}_\epsilon, \underline{U}_\epsilon, H_\epsilon, \underline{w}_\epsilon)$ verifies the convergence results proved in Theorem 3, we can pass to the limit in (3.24). The method is the same as in the proof of Proposition 1. We thus obtain that $Y = (\underline{R}, S)$ verifies the equations

$$\left(-\frac{\partial \underline{R}}{\partial t}, \underline{v}\right) + a_1(\underline{R}, \underline{v}) + \frac{C}{D}\left(\mid \underline{u} \mid_2 \underline{R} + (\underline{R}, \underline{u})_2 \frac{u}{\mid \underline{u} \mid_2}, \underline{v}\right) - (\underline{F} \wedge \underline{R}, \underline{v})$$
$$-((h\nabla)S, \underline{v}) = 0,$$
$$\left(-\frac{\partial S}{\partial t}, \beta\right) + a_2(S, \beta) + (\mathrm{div}(\beta \underline{u}), S) - g(\mathrm{div}(\underline{R}), \beta) = d((h - h_{obs})^3, \beta) \quad \forall (\underline{v}, \beta) \in V,$$
$$(\underline{R}, S)(T) = (\underline{0}, 0),$$

and

$$\int_0^T (\underline{R} + d\gamma \underline{f},\ \underline{g} - \underline{f}) dt \geq 0 \ \forall \underline{g} \in K_c, \text{where } d = 0 \ \mathit{or} \ 1.$$

We are now going to prove that $d = 1$ in order to achieve the proof of Theorem 4. Let us suppose that $d = 0$; then $(\underline{R}, S)$ verifies the system

$$\begin{aligned}
&\left(-\frac{\partial \underline{R}}{\partial t}, \underline{v}\right) + a_1(\underline{R}, \underline{v}) + \frac{C}{D}\left(|\,\underline{u}\,|_2\ \underline{R} + (\underline{R}, \underline{u})_2 \frac{\underline{u}}{|\,\underline{u}\,|_2}, \underline{v}\right) - (\underline{F} \wedge \underline{R}, \underline{v})\\
&-((h\nabla)S, \underline{v}) = 0,\\
&\left(-\frac{\partial S}{\partial t}, \beta\right) + a_2(S, \beta) + (\mathrm{div}(\beta \underline{u}), S) - g(\mathrm{div}(\underline{R}), \beta) = 0 \quad \forall(\underline{v}, \beta) \in V,\\
&(\underline{R}, S)(T) = (\underline{0}, 0).
\end{aligned} \tag{3.26}$$

Setting $(\underline{v}, \beta) = (\underline{R}, S)$ in (3.26) and adding the first and second parts of (3.26), we obtain

$$\begin{aligned}
&-\frac{1}{2}\frac{\partial\,|\,Y\,|^2}{\partial t} + a(Y, Y) + \frac{C}{D}\int_\Omega |\,\underline{u}\,|_2 \quad |\,\underline{R}\,|_2^2\, d\Omega + \frac{C}{D}\int_\Omega \frac{(\underline{u}, \underline{R})_2^2}{|\,\underline{u}\,|_2} d\Omega\\
&= ((h\nabla)S, \underline{R}) - (\mathrm{div}(S\underline{u}), S) + g(\mathrm{div}(\underline{R}), S),
\end{aligned}$$

where $Y = (\underline{R}, S)$.

By using the coercivity of operator a, Lemma 3, and Hölder's inequality, we obtain

$$-\frac{\partial\,|\,Y\,|^2}{\partial t} + \alpha \parallel Y \parallel^2 \leq c(1 + \parallel h \parallel_{L^4(\Omega)}^4 + \parallel \underline{u} \parallel_{L^4(\Omega)}^4)\,|\,Y\,|^2\ .$$

The functions $\parallel h \parallel_{L^4(\Omega)}^4$ and $\parallel \underline{u} \parallel_{L^4(\Omega)}^4$ are integrable with respect to time, so by applying Gronwall's lemma, we deduce

$$|\,Y\,|^2 \leq c\,|\,Y(T)\,|^2\ exp\left(\int_t^T (\parallel h(s) \parallel_{L^4(\Omega)}^4 + \parallel \underline{u}(s) \parallel_{L^4(\Omega)}^4) ds\right) \quad \forall t \in [0, T].$$

Since $Y(T) = 0$, we thus obtain $Y = 0$.

Since the solution S is the limit with $\epsilon \to 0$ of $\widetilde{S}_\epsilon$ and the sequence $\parallel \widetilde{S}_\epsilon \parallel_{L^4(0,T,L^4(\Omega))}$ is equal to 1, then $\parallel S \parallel_{L^4(0,T,L^4(\Omega))}$ is equal to 1 and we have a contradiction. So we can conclude that $d = 1$.

This achieves the proof of the theorem. □

3.2. Study of the optimal control problem with nonhomogeneous boundary conditions. The results obtained in section 3.1 can easily be extended to the general nonhomogeneous case. Then the first step consists of solving linear problems (1.3) and (1.4) whose solution $(\underline{u}_L, h_L)$ doesn't depend on the control $\underline{f}$. We have proved the equivalence between initial problem (1.1)–(1.2) and homogeneous problem (1.5). Therefore the control method is applied to problem (1.5), the penalization method works in the same way as in section 3.1, and we obtain the two following theorems.

Theorem 5. *Under the assumptions of Proposition* 1, *the optimal control problem* (3.9) *has at least one solution* $(\underline{f}, \underline{u})$ *in* U_{ad}.

THEOREM 6. *Optimal control* $(\underline{f}, \underline{u})$ *is characterized by*

$$\begin{aligned}
&\left(\frac{\partial \underline{u}}{\partial t}, \underline{v}\right) + a_1(\underline{u}, \underline{v}) + \frac{C}{D}(\mid (\underline{u} + \underline{u}_L) \mid_2 (\underline{u} + \underline{u}_L), \underline{v}) + (\underline{F} \wedge \underline{u}, \underline{v}) - g(\mathrm{div}(\underline{v}), h) = (\underline{f}, \underline{v}),\\
&\left(\frac{\partial h}{\partial t}, \beta\right) + a_2(h, \beta) + (\mathrm{div}(h\underline{u}), \beta) + (\mathrm{div}(h_L\underline{u}), \beta) + (\mathrm{div}(h\underline{u}_L), \beta) = (f_1, \beta)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall(\underline{v}, \beta) \in V,\\
&(\underline{u}, h)(t = 0) = (\underline{u}_0, h_0),
\end{aligned}$$

$$\begin{aligned}
&\left(-\frac{\partial \underline{R}}{\partial t}, \underline{v}\right) + a_1(\underline{R}, \underline{v}) + \frac{C}{D}\left(\mid (\underline{u} + \underline{u}_L) \mid_2 \underline{R} + (\underline{u} + \underline{u}_L, \underline{R})_2 \frac{(\underline{u} + \underline{u}_L)}{\mid (\underline{u} + \underline{u}_L) \mid_2}, \underline{v}\right)\\
&-(\underline{F} \wedge \underline{R}, \underline{v}) - ((h + h_L)\nabla S, \underline{v}) = 0,\\
&\left(-\frac{\partial S}{\partial t}, \beta\right) + a_2(S, \beta) - (((\underline{u} + \underline{u}_L)\nabla)S, \beta) - g(\mathrm{div}(\underline{R}), \beta) = ((h - h_{obs})^3, \beta)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall(\underline{v}, \beta) \in V,\\
&(\underline{R}, S)(t = T) = (\underline{0}, 0),
\end{aligned}$$

and

$$(\underline{R} + \gamma \underline{f},\ \underline{g} - \underline{f})_{U_c} \geq 0\ \forall \underline{g} \in K_c.$$

4. Conclusion.

We have developed a control method in order to calculate the velocity $\underline{u}$ and the depth h. The sea level, deduced from altimetric measurements, constitutes the observation in our model. The fluid exchanges are supposed to be known; the outside stress $\underline{f}$ is unknown. We take then $\underline{f}$ as the control. $(\underline{u}(\underline{f}), h(\underline{f}))$ are the velocity and depth corresponding to any control $\underline{f}$. The optimal control is defined as the outside stress minimizing a given cost function which measures the distance between the observed depth and the depth $h(\underline{f})$. The observed depth is driven by an outside stress $\underline{f}_r$. $\underline{f}$ is an approximation of the unknown real outside stress $\underline{f}_r$ and then $(\underline{u}(\underline{f}), h(\underline{f}))$ induced by the optimal control $\underline{f}$ has to approach the real circulation $(\underline{u}_r, h_r)$ observed by the satellite.

The existence of optimal control is proved by means of minimizing sequences. The question of the uniqueness is not treated; in general the answer is expected to be negative. To characterize the control, we introduce a family of penalized control problems; we obtain a set of equations characterizing these problems. Finally we demonstrate the convergence of the penalized problems and obtain the optimality conditions from which the solution of the initial nonlinear control problem and states may be determinated.

We can use this theoretical study to develop a numerical scheme: let a fixed ϵ be small enough, we may choose $\underline{f}_1$ and so $(\underline{u}_1, h_1)$ is the solution of the nonlinear shallow-water problem driven by $\underline{f}_1$. Then we solve the linear penalized control problem numerically, for example, by using a quasi-Newton method. We obtain then the solution $(\underline{f}_\epsilon, \underline{w}_\epsilon, \underline{U}_\epsilon, H_\epsilon)$ and we compare $(\underline{f}_\epsilon, \underline{U}_\epsilon, H_\epsilon)$ to $(\underline{f}_1, \underline{u}_1, h_1)$. If the error is not sufficiently small, we initialize the algorithm with $(\underline{f}_\epsilon, \underline{U}_\epsilon, H_\epsilon)$ and we iterate the process. The computed solution $(\underline{f}_\epsilon, \underline{U}_\epsilon, H_\epsilon)$ approximates the solution of the initial nonlinear control problem. The approximate solution $(\underline{f}_\epsilon, \underline{U}_\epsilon, H_\epsilon)$ depends on the small parameter ϵ. The circulation $(\underline{U}_\epsilon, H_\epsilon)$ is driven by the outside stress $\underline{f}_\epsilon$. In order to confirm and to validate numerically the convergence results and the control method, we may compare $(\underline{f}_\epsilon, \underline{U}_\epsilon, H_\epsilon)$ and $(\underline{f}_r, \underline{u}_r, h_r)$ for different values of ϵ.

We can thus compute the velocity and depth in a shallow-water domain Ω, during a time T, from satellite observations of the surface topography.

Acknowledgment. The author is grateful to the referee for many useful comments and suggestions which have improved the presentation of this paper.

REFERENCES

[1] A. Belmiloudi, *Resolution of optimal control problem for a perturbation linearized Navier-Stokes type equations*, in Proceedings of the 22nd Summer School, Application of Mathematics in Engineering and Business, Tech. Univ. of Sofia, Sozopol, Bulgaria, 1996, pp. 39–51.

[2] A. Belmiloudi, *A nonlinear optimal control problem for assimilation of surface data in Navier-Stokes type equations related to oceanography,* Numer. Funct. Anal. Optim., 20 (1999), pp. 1–26.

[3] A. Belmiloudi, *Regularity results and optimal control problems for the perturbation of Boussinesq equations of the ocean,* Numer. Funct. Anal. Optim., 21 (2000), pp. 623–651.

[4] A. Belmiloudi and F. Brossier, *A control method for assimilation of surface data in a linearized Navier-Stokes-type problem related to oceanography,* SIAM J. Control Optim., 35 (1997), pp. 2183–2197.

[5] C. Bernardi and O. Pironneau, *On the shallow water equations at low Reynolds number,* Comm. Partial Differential Equations, 16 (1991), pp. 59–104.

[6] H. Brezis, *Analyse fonctionnelle. Théorie et Application*, Masson, Paris, 1983.

[7] C. Fabre, J.P. Puel, and E. Zuazua, *Approximate controllability for the semilinear heart equation,* Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[8] H.O. Fattorini and S.S. Sritharan, *Existence of optimal controls for viscous flows problems,* Proc. Roy. Soc. London Ser. A, 439 (1992), pp. 81–102.

[9] H.O. Fattorini and S.S. Sritharan, *Necessary and sufficient conditions for optimal controls in viscous flow,* Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 211–251.

[10] M.D. Gunzburger, L.S. Hou, and TH.P. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls,* RAIRO Math. Modél. Anal. Numér., 25 (1991), pp. 711–748.

[11] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications,* tome 1 et 2, Dunod, Paris, 1968.

[12] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations,* Springer-Verlag, New York, 1971.

[13] J.-L. Lions, *Equations differentielles operationnelles,* Springer-Verlag, New York, 1961.

[14] J.-L. Lions, *Contrôle des systèmes distribués singuliers,* Gauthier-Villars, Paris, 1983.

[15] G.I. Marchuk, *Mathematical Models in Environmental Problems,* North-Holland, Amsterdam, 1986.

[16] G.I. Marchuk, *Adjoint Equations and Analysis of Complex systems,* Kluwer, Dordrecht, The Netherlands, 1995.

[17] J. Pedlosky, *Geophysical Fluid Dynamics,* Springer-Verlag, New York, 1979.

[18] R. Temam, *Navier-Stokes Equations,* North-Holland, Amsterdam, 1977.

[19] R. Temam, *Navier-Stokes Equations Theory and Nonlinear Functional Analysis,* CBMS-NSF Regional Conf. Ser. in Appl. Math 41, SIAM, Philadelphia, 1983.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1585–1614

GENERIC SIMPLICITY OF THE SPECTRUM AND STABILIZATION FOR A PLATE EQUATION*

JAIME H. ORTEGA† AND ENRIQUE ZUAZUA‡

Abstract. In this work we prove the generic simplicity of the spectrum of the clamped plate equation in a bounded regular domain of $\mathbb{R}^d$. That is, given $\Omega \subset \mathbb{R}^d$, we show that there exists an arbitrarily small deformation of the domain u, such that all the eigenvalues of the plate system in the deformed domain $\Omega + u$ are simple. To prove this result we first prove a nonstandard unique continuation property for this system that also holds generically with respect to the perturbations of the domain. Both the proof of this generic uniqueness result and the generic simplicity of the spectrum use Baire's lemma and shape differentiation. Finally, we show an application of this unique continuation property to a result of generic stabilization for a plate system with one dissipative boundary condition.

Key words. spectral theory, plate equation, unique continuation property, stabilization

AMS subject classifications. 35P05, 35J40, 93D15

PII. S0363012900358483

1. Introduction and main results.

In this work we are interested in the study of the spectral properties for the plate system

$$
\left\{
\begin{array}{rcll}
\triangle^2 y & = & \lambda y & \text{in } \Omega, \\
y & = & 0 & \text{on } \partial\Omega, \\
\dfrac{\partial y}{\partial n} & = & 0 & \text{on } \partial\Omega,
\end{array}
\right. \tag{1.1}
$$

where $\Omega \subseteq \mathbb{R}^d$ is a bounded domain with boundary of class C^4.

Problem (1.1) admits a sequence of eigenvalues

$$0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n \le \cdots \longrightarrow \infty,$$

which have finite multiplicity. The eigenfunctions $\{y_n\}_n \subset H_0^2(\Omega)$ of (1.1) can be chosen to form an orthonormal basis of $H_0^2(\Omega)$.

On the other hand, it is well known that the eigenvalues of (1.1) are not always simple. For instance, in [8], it is shown that the first eigenvalue is not simple in a suitable annular domain.

The problem of the simplicity of the spectrum arises in many contexts. This is, for instance, the case when analyzing stabilizability and controllability issues for evolution systems. When the spectrum is simple one can often reduce these problems to the analysis of suitable properties of eigenfunctions, which is an easier problem to deal with because of the lack of dependence with respect to time. We refer to J. L.

*Received by the editors June 13, 2000; accepted for publication (in revised form) September 20, 2000; published electronically January 31, 2001.
http://www.siam.org/journals/sicon/39-5/35848.html

†Universidad del Bío-Bío, Facultad de Ciencias, Departamento de Matemática, Casilla 5-C, Concepción, Chile (jortega@ubiobio.cl). The research of this author was supported by the Fellowship Programs of the Instituto de Cooperación Iberoamericana (ICI) (1993–1996) and of the Presidente de la República de Chile (1996–1997).

‡Universidad Complutense de Madrid, Departamento de Matemática Aplicada, 28040 Madrid, Spain (zuazua@eucmax.sim.ucm.es). The research of this author was supported by grants PB96-0663 of the DGES and FMRX-CT960033 of the European Union.

Lions and E. Zuazua [21] for an example of a control problem where this strategy is successfully applied.

Concerning the simplicity of the spectrum, the problem we address is as follows. *Are there arbitrarily small deformations of the domain* $u \in W^{5,\infty}(\Omega, \mathbb{R}^d)$, *such that the spectrum of* (1.1) *in the deformed domain* $\Omega + u$ *is simple?*

In this paper we give a positive answer to this question. We show that the set of deformations of the domain $u \in W^{5,\infty}(\Omega, \mathbb{R}^d)$, such that the spectrum of the plate system is simple, is a dense subset of $W^{5,\infty}(\Omega, \mathbb{R}^d)$.

The generic simplicity of the spectrum of second order elliptic operators is by now well known. We refer to J. H. Albert [1] for perturbations of the coefficients of the operator and to A. M. Micheletti [22] and K. Uhlenbeck [29] for perturbations of the domain.

The result we prove in this paper and the methods we employ are inspired in [25] by the authors, where a similar result was proved for the $2-D$ Stokes system. The proof combines Baire's lemma and shape differentiation. These two tools reduce the problem to the obtainment of a suitable unique continuation property for the eigenfunctions. In the context of second order problems this uniqueness problem can be dealt with by means of Holmgrem's theorem. However, this is not the case when working with the plate equation. In this case the uniqueness problem cannot be analyzed in the context of the classical Cauchy problems since only three boundary conditions are known to vanish. We then proceed as in [25], showing that the uniqueness property holds generically with respect to the perturbations of the domain. But this turns out to be sufficient to complete the proof of the generic simplicity of the spectrum.

Our generic unique continuation property refers to the following uniqueness problem.

If y *solves* (1.1) *for some* $\lambda > 0$ *and*

$$\frac{\partial^2 y}{\partial n^2} = 0 \qquad \text{on } \Gamma_0,$$

then, necessarily, $y \equiv 0$.

This property will be referred to as *spectral uniqueness.*

THEOREM 1.1. *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded domain with boundary of class* C^4. *Let* Γ_0 *be an open nonempty subset of* $\partial\Omega$.

Then, the set of deformations of the domain $u \in W^{5,\infty}(\Omega, \mathbb{R}^d)$, *such that* $u = 0$ *on* $\partial\Omega \setminus \Gamma_0$ *and for which spectral uniqueness holds when* Ω *and* Γ_0 *are replaced by* $\Omega + u$ *and* $\Gamma_0 + u$, *is residual in*

$$W_0 = \{u \in W^{5,\infty}(\Omega, \mathbb{R}^d) : u = 0 \text{ on } \partial\Omega \backslash \Gamma_0\}.$$

In other words, it is a countable intersection of dense open sets of W_0. *In particular, it is dense in* W_0. □

Remark 1.1. (1) As far as we know, there is no example in the literature of bounded domain Ω and open subset Γ_0 of $\partial\Omega$ for which there exists a nontrivial eigenfunction of (1.1) such that the further condition

$$\frac{\partial^2 y}{\partial n^2} = 0 \quad \text{on } \Gamma_0 \tag{1.2}$$

is satisfied.

However, our result is of a generic nature, and, therefore, it does not apply to any domain Ω and open subset Γ_0 of $\partial\Omega$.

Therefore, whether this unique continuation property holds for any Ω and $\Gamma_0 \subset \partial\Omega$ remains an open problem.

(2) Using multiplier techniques (see Appendix I of [18], [19], and [13]), one can show that this unique continuation holds for any domain Ω provided Γ_0 is a subset of the boundary of the form

$$\Gamma_0 = \Gamma(x_0) = \{x \in \partial\Omega \,:\, (x - x_0) \cdot n(x) > 0\}.$$

In fact, multiplier methods allow us to show that uniqueness holds for the dynamic plate model, but always for subsets of the boundary of this particular form.

Note, however, that these subsets of the boundary are always large. In other words, there are many small subsets of the boundary Γ_0 that cannot be written in this form.

Consequently, our result is, as far we know, the first one that applies to arbitrarily small subsets of boundary, but it is of generic nature.

(3) Note also that the deformations we apply do deform the subset Γ_0 itself. Whether this unique continuation result applies when u preserves Γ_0 or not is an open problem. □

With the aid of this result we prove the main result on the generic simplicity of the spectrum.

THEOREM 1.2. *Let Ω be a bounded domain of $\mathbb{R}^d$ of class C^4. Let Γ_0 be an open nonempty subset of $\partial\Omega$.*

Then the set

$$A = \left\{u \in W^{5,\infty}(\Omega, \mathbb{R}^d) : u = 0 \text{ on } \partial\Omega \setminus \Gamma_0 \text{ and the spectrum of } (1.1) \text{ is simple}\right\}$$

is residual in

$$W_0 = \left\{u \in W^{5,\infty}(\Omega, \mathbb{R}^d) : u = 0 \text{ on } \partial\Omega \setminus \Gamma_0\right\}. \qquad □$$

We also show an application of Theorem 1.1 to the study of the stabilization of the plate system with dissipative boundary conditions.

Let us consider the system

$$(1.3) \qquad \begin{cases} y_{tt} + \triangle^2 y &= 0 & \text{in } \Omega \times (0,\infty), \\ y &= 0 & \text{on } \partial\Omega \times (0,\infty), \\ \dfrac{\partial y}{\partial n} &= 0 & \text{on } \partial\Omega \setminus \Gamma_0 \times (0,\infty), \\ \triangle y &= -\dfrac{\partial y_t}{\partial n} & \text{on } \Gamma_0 \times (0,\infty), \\ y(x,0) &= y_0 & \text{in } \Omega, \\ y_t(x,0) &= y_1 & \text{in } \Omega. \end{cases}$$

It is easy to see that for any $(y_0, y_1) \in X = X_1 \times X_2$ system (1.3) admits a unique solution $y \in C([0,\infty); X_1) \cap C^1([0,\infty); X_2)$. Here and in what follows $X_1 = \{\varphi \in H^2(\Omega) \cap H_0^1(\Omega) : \frac{\partial \varphi}{\partial n} = 0 \text{ on } \partial\Omega \setminus \Gamma_0\}$ and $X_2 = L^2(\Omega)$.

We define the energy of the system as

$$(1.4) \qquad E(t) = \frac{1}{2} \int_\Omega \left[|y_t|^2 + |\triangle y|^2\right].$$

From the dissipative boundary conditions of (1.3) we have that

$$\frac{d}{dt}E(t) = -\int_{\Gamma_0} |\triangle y(x,t)|^2 \, d\Gamma \leq 0. \tag{1.5}$$

Thus, we deduce that the energy decreases for any solution y of (1.3) as $t \to \infty$. Therefore, the following question arises naturally.

Does the energy of any solution of (1.3) *tend to zero as* $t \to \infty$*? In other words, does*

$$E(t) \to 0 \qquad \text{as} \quad t \to \infty \tag{1.6}$$

for every solution of (1.3)*?*

We will see that the stabilization property (1.6) holds generically with respect to the domain Ω, as in the following theorem.

THEOREM 1.3. *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set of class C^4, and let Γ_0 be an open nonempty subset of the boundary, such that the spectral uniqueness property holds.*

Then, every solution of (1.3) *verifies that*

$$E(t) \to 0 \qquad \text{as } t \to +\infty. \qquad \square \tag{1.7}$$

To prove this result we use La Salle's invariance principle, which allows us to reduce our stabilization problem to a unique continuation property. This unique continuation problem turns out to be the one we have solved generically in Theorem 1.1, i.e., the spectral uniqueness problem.

We could also consider system (1.3) with two dissipative boundary conditions instead of one. Consider, for instance,

$$\begin{cases} y = \dfrac{\partial y}{\partial n} = 0 & \text{on } \partial\Omega \setminus \Gamma_0 \times (0,\infty), \\ \triangle y = -\dfrac{\partial y_t}{\partial n} & \text{on } \Gamma_0 \times (0,\infty), \\ \dfrac{\partial \triangle y}{\partial n} = y_t & \text{on } \Gamma_0 \times (0,\infty). \end{cases} \tag{1.8}$$

Then the stabilization result above holds for every domain Ω and for every nonempty open subset $\Gamma_0 \subset \partial\Omega$. In this case, by means of La Salle's invariance principle, we can reduce our problem to a unique continuation one that may be solved by applying the classical uniqueness theorem by Holmgrem. Indeed, in this case, we add the extra boundary conditions $\triangle y = \frac{\partial y}{\partial n} = 0$ on Γ_0 to the solutions of (1.1), in which case we are in the context of the classical Cauchy problem.

Remark 1.2. The methods developed in this paper may be applied to other plate systems, such as when the plate equation is replaced by the one taking into account the rotational inertia term, i.e.,

$$y_{tt} - \gamma \triangle y_{tt} + \triangle^2 y = 0,$$

with $\gamma > 0$. Other boundary conditions may also be considered. For instance, the condition $\frac{\partial y}{\partial n} = 0$ on $\partial\Omega \setminus \Gamma_0$ may be replaced by $\triangle y = 0$ on $\partial\Omega \setminus \Gamma_0$.

The same result applies as well to plate equations with nonlinear monotone boundary damping. $\square$

Remark 1.3. When Γ_0 is a subset of the boundary of the form $\Gamma(x_0)$ we indicated in Remark 1.1, uniform stability properties may be proved for the systems under consideration, i.e.,

$$E(t) \leq Ce^{-\alpha t}E(0)$$

for suitable $C, \alpha > 0$ (see, for instance, [15]). We do not address this problem here. In any case, one does not expect, in general, exponential decay to hold when the dissipative term acts on a small subset of the boundary. □

The rest of this work is organized as follows. In section 2 we present some preliminary results on the variational formulation of the plate equation and shape differentiation. In section 3 we prove a result of existence and regular dependence of the branches of eigenvalues and eigenfunctions of the bilaplacian with respect to the perturbation of the domain. In section 4 we compute the local variations of the eigenvalues and eigenfunctions of the bilaplacian. In section 5 we prove the unique continuation property of Theorem 1.1. In section 6 we prove the simplicity of the spectrum of Theorem 1.2. Finally, in section 7 we prove the stabilization result of Theorem 1.3.

2. Preliminaries.

2.1. Baire's lemma.

First we remember the Baire's lemma, which will be a useful tool.

LEMMA 2.1 (Baire's lemma). *Let X be a complete metric space, and let A_n be an open dense subset of X for all $n \in \mathbb{N}$.*

Then $\cap_{n\in\mathbb{N}} A_n$ is dense in X. □

A direct consequence of Baire's lemma is the following result.

LEMMA 2.2. *Let X be a complete metric space, and let $\{A_n\}_{n\geq 0}$ be a sequence of open subsets of X such that*

1. *$A_0 = X$, and*
2. *A_{n+1} is a dense subset of A_n for all $n \geq 0$.*

Then $\cap_{n=1}^{\infty} A_n$ is dense in X. □

2.2. Variational formulation of the plate system.

The variational formulation of the eigenvalue problem (1.1) is as follows: to find $y \in H_0^2(\Omega)$ and $\lambda \in \mathbb{R}$ such that

$$\int_\Omega \triangle y \triangle \varphi = \lambda \int_\Omega y\varphi \qquad \forall \varphi \in H_0^2(\Omega). \tag{2.1}$$

This variational eigenvalue problem can be handled in a standard way.

It is well known that there exists a positive constant $c > 0$ such that

$$\|f\|^2_{H^2(\Omega)} \leq c \int_\Omega |\triangle f|^2 \quad \forall f \in H_0^2(\Omega). \tag{2.2}$$

Then

$$|f|_2 = \left(\int_\Omega |\triangle f|^2\right)^{\frac{1}{2}} \tag{2.3}$$

defines a norm in $H_0^2(\Omega)$, equivalent to the one induced by the norm of $H^2(\Omega)$.

Thus

$$b : H_0^2(\Omega) \times H_0^2(\Omega) \to \mathbb{R}, \qquad b(\phi,\varphi) = \int_\Omega \triangle\phi\,\triangle\varphi \tag{2.4}$$

is a coercive and continuous bilinear form. Then, for each $f \in H^{-2}(\Omega)$, there exists a unique solution $y \in H_0^2(\Omega)$ of the problem

$$\int_\Omega \triangle y \triangle\varphi = \langle f,\varphi\rangle_{H^{-2}\times H_0^2} \qquad \forall\varphi \in H_0^2(\Omega).$$

This shows the existence and uniqueness of weak solutions of the elliptic problem

$$\begin{cases} \triangle^2 y &= f & \text{in } \Omega, \\ y &= 0 & \text{on } \partial\Omega, \\ \dfrac{\partial y}{\partial n} &= 0 & \text{on } \partial\Omega. \end{cases} \tag{2.5}$$

Using the compactness of the imbedding $H_0^2(\Omega) \hookrightarrow L^2(\Omega)$, one can show that the map $f \in L^2(\Omega) \to y \in L^2(\Omega)$ is compact. It is also easy to see that it is self-adjoint. Applying the classical spectral theory for compact, self-adjoint operators, we deduce that the eigenvalue problem (2.1) admits a sequence of eigenvalues

$$0 < \lambda_1 \le \lambda_2 \le \lambda_3 \le \cdots \to +\infty.$$

Moreover, each eigenvalue has finite multiplicity, and the eigenfunctions $y_i \in H^4(\Omega) \cap H_0^2(\Omega)$ can be chosen to form an orthonormal basis of $L^2(\Omega)$.

2.3. Shape differentiation. An important tool for the study of the generic properties is the *shape differentiation.* For more details about this technique, we refer to [4], [27], [28], and the bibliographies therein.

Given a domain Ω and a function $u : \Omega \to \mathbb{R}^d$, we define the new domain $\Omega + u$ by

$$\Omega + u = \{y \in \mathbb{R}^d : y = x + u(x),\ x \in \Omega\}. \tag{2.6}$$

Let us consider perturbations u in the space $W^{k,\infty}(\Omega,\mathbb{R}^d)$ with norm

$$\|u\|_{k,\infty} = \underset{0\le|\alpha|\le k,\, x\in\Omega}{\text{sup ess}} |D^\alpha u(x)|.$$

The following results are well known.

LEMMA 2.3 (see [28]). *Let $u \in W^{k,\infty}(\Omega,\mathbb{R}^d)$, and let $k \ge 1$ be such that $\|u\|_{k,\infty} \le \frac{1}{2}$. Then the map $(I+u) : \Omega \to \Omega + u$ is invertible. Furthermore, there exists $w \in W^{k,\infty}(\Omega,\mathbb{R}^d)$ such that $(I+u)^{-1} = I + w$ and $\|w\|_{k,\infty} \le C_k \|u\|_{k,\infty}$, where C_k is a constant independent on u.* □

Remark 2.1. According to this result, if Ω is of class C^j, we can choose $k = j+1$ (and therefore the perturbation space $W^{k,\infty}(\Omega,\mathbb{R}^d)$) such that our new domain $\Omega + u$ is also of class C^j. In particular, if Ω is of class C^4, then $\Omega + u$ is also of class C^4, and the solutions of the eigenvalue problem for the bilaplacian in the new domain $\Omega + u$ satisfy $y(u) \in H^4(\Omega+u) \cap H_0^2(\Omega+u)$ for every $u \in W^{5,\infty}(\Omega,\mathbb{R}^d)$ small enough. This is the functional framework we shall work in. □

LEMMA 2.4 (see [28]). *Let $k \geq 1$, and consider the function*

$$\begin{array}{rrcl} \gamma : & W^{k,\infty}(\Omega, \mathbb{R}^d) & \rightarrow & W^{k-1,\infty}(\Omega, \mathbb{R}), \\ & u & \rightarrow & \gamma(u) = Jac\,(I+u) = \left|\det\left[\partial_j\,(I+u)_i\right]\right|. \end{array}$$

This function γ is differentiable at $u = 0$. Furthermore, the directional derivative in the direction w at the point $u = 0$ is $\operatorname{div} w$; *that is,*

$$\langle D\gamma(0), w\rangle = \operatorname{div}\, w \quad \forall\, w \,\in\, W^{k,\infty}(\Omega, \mathbb{R}^d). \qquad \square$$

LEMMA 2.5 (see [28]). *Let $k \geq 1$. The map*

$$\begin{array}{rrcl} \beta : & \mathcal{W} \subset W^{k,\infty}(\Omega, \mathbb{R}^d) & \rightarrow & \mathcal{M}_{d\times d}\left(W^{k-1,\infty}(\Omega, \mathbb{R})\right), \\ & u & \rightarrow & {}^t\left[\partial_j\,(I+u)_i\right]^{-1}, \end{array}$$

where $\mathcal{W}$ is a neighborhood of $u = 0$ on $W^{k,\infty}(\Omega, \mathbb{R}^d)$, is differentiable on $u = 0$. Its directional derivative on $u = 0$ in the direction w is given by the matrix $-^t\left[\partial_j w_i\right]$, where ${}^t\left[\partial_j w_i\right]$ denotes the adjoint of $\left[\partial_j w_i\right]$. In other words,

$${}^t\left[\partial_j\,(I+u)_i\right]^{-1} = [I] -^t \left[\partial_j u_i\right] + \theta(u),$$

where the matrix $\theta(u)$ satisfies

$$\frac{\|\theta(u)\|_{k-1,\infty}}{\|u\|_{k,\infty}} \rightarrow 0 \quad \text{as} \quad \|u\|_{k,\infty} \rightarrow 0. \qquad \square$$

Now, we consider a function

$$\begin{array}{rrcl} v : & W^{k,\infty}(\Omega, \mathbb{R}^d) & \rightarrow & W^{m,r}(\Omega+u), \\ & u & \rightarrow & v(u), \end{array}$$

where $1 \leq r < \infty$ and $m \leq k$ are integer numbers. In practice, $v(u)$ will be the solution of a suitable problem, which depends on the perturbation function u (for instance, a solution of our eigenvalue problem (1.1)).

We are interested in the study of the regularity of the function $v(u)$ with respect to the perturbation parameter u.

DEFINITION 2.1 (see [28]). *Let $k \geq m \geq 1$, $1 \leq r < \infty$. We say that the function $v(u)$ has a first order local variation at $u = 0$ on $W^{m-1,r}_{loc}(\Omega)$ if $v(u) \in W^{m,r}(\Omega+u)$ for all $u \in W^{k,\infty}(\Omega, \mathbb{R}^d)$ and there exists a linear map $v'(\Omega; u)$ defined from $u \in W^{k,\infty}(\Omega, \mathbb{R}^d)$ to $W^{m-1,r}_{loc}(\Omega)$ such that, for each open set $\omega \subset\subset \Omega$,*

$$v(u) = v(0) + v'(\Omega; u) + \widehat{\theta}(u) \ \text{ in } \omega,$$

when $\|u\|_{k,\infty}$ is small enough and

$$\frac{\widehat{\theta}(u)}{\|u\|_{k,\infty}} \rightarrow 0 \quad \text{in } W^{m-1,r}(\omega) \quad \text{as } \|u\|_{k,\infty} \rightarrow 0. \qquad \square$$

Remark 2.2. From Definition 2.1 it follows that the *first local variation* can be defined as

$$v'(\Omega; u) = \lim_{t\rightarrow 0} \frac{v(tu)\,|_{\omega} - v(0)\,|_{\omega}}{t} \quad \text{ in } \ \omega, \tag{2.7}$$

where $\omega \subset\subset \Omega$ and $v(tu)|_{\omega}$, $v(0)|_{\omega}$ are the restrictions of the functions $v(tu)$ and $v(0)$ to ω. □

In what follows, to simplify the notation, we will write $v'(u) = v'(\Omega; u)$.

The following theorem provides sufficient conditions for the existence of the first local variation for functions which depend on the deformation u. Furthermore, it provides an expression for the local variation on the boundary in terms of the normal derivative of $v(0)$.

THEOREM 2.6 (see [28]). *Let Ω be a $C^{0,1}$ domain. Consider a map $u \to v(u) \in W^{m,r}(\Omega + u)$ defined on a neighborhood of $u = 0$ in $W^{k,\infty}(\Omega, \mathbb{R}^d)$, with $k \geq m \geq 1$ and $1 \leq r < \infty$.*

Let us assume that there exists a linear continuous map $u \to \dot{v}(u)$ defined on $W^{k,\infty}(\Omega, \mathbb{R}^d)$ with values in $W^{m,r}(\Omega)$, such that

$$v(u) \circ (I + u) = v(0) + \dot{v}(u) + \theta(u) \quad \text{in } W^{m,r}(\Omega)$$

for all $u \in W^{k,\infty}(\Omega, \mathbb{R}^d)$ small enough, where

$$\frac{\theta(u)}{\|u\|_{k,\infty}} \to 0 \quad \text{on } W^{m,r}(\Omega) \quad \text{as } \|u\|_{k,\infty} \to 0.$$

Furthermore, assume that for each $u \in W^{k,\infty}(\Omega, \mathbb{R}^d)$ small enough,

$$v(u) = 0 \quad \text{on} \quad \partial(\Omega + u).$$

Then, for each $\omega \subset\subset \Omega$, the function $u \to v_\omega(u) = v(u)|_\omega$, defined on a neighborhood of $u = 0$ in $W^{k,\infty}(\Omega, \mathbb{R}^d)$ with values in $W^{m-1,r}(\omega)$, is differentiable at $u = 0$.

Moreover, the map $u \to v(u)$ has a local derivative at $u = 0$ (see Definition 2.1) and the local derivative at $u = 0$, in the direction u, denoted by $v'(u)$, verifies $v'(u) \in W^{m-1,r}(\Omega)$ and

$$v'(u) = -(u \cdot n) \frac{\partial v(0)}{\partial n} \quad \text{on} \quad \partial\Omega,$$

where n is the unit outward normal vector to Ω. □

In what follows we will use the notation

$$\mathcal{W} = \{u \in W^{k,\infty}(\Omega, \mathbb{R}^d) : \|u\|_{k,\infty} < c_\Omega\},$$

where $k \geq 1$ and $c_\Omega < 1/2$ is small enough such that all the previous results hold.

LEMMA 2.7 (see [4, Lemma 9]). *Let $u \in \mathcal{W}$. If $f \in H_0^1(\Omega + u)$, there exists a unique $g \in H_0^1(\Omega)$ such that $f \circ (I + u) = g$. Moreover,*

$$\left(\frac{\partial f}{\partial z_i}\right) \circ (I + u) = \sum_j M_{ij}(u) \frac{\partial g}{\partial x_j} = D_i(u) g, \tag{2.8}$$

where the matrix $M(u)$ is defined as

$$M(u) = [M_{ij}(u)] =^t \left[\frac{\partial}{\partial x_j}(I + u)_i\right]^{-1},$$

and

$$z_i = x_i + u_i(x) \quad \forall\, x \in \Omega. \quad □$$

3. Regularity of the eigenvalues and eigenfunctions.

3.1. Some results of spectral theory. To prove the existence and regularity of the eigenvalues and eigenfunctions of the plate system with respect to the perturbation parameter u, we will use the Lyapunov–Schmidt method (see [30], [7, p. 30]).

LEMMA 3.1 (see [7, Lemma 4.1, p. 31]). *Suppose that X and Z are Hilbert spaces and $A : X \longrightarrow Z$ is a continuous linear operator. Let $U : X \longrightarrow N(A)$, $E : Z \longrightarrow R(A)$ be the orthogonal projections from X and Z on the kernel and range of A, respectively.*

Then, there exists a bounded linear operator $Q : R(A) \longrightarrow N(A)^{\perp}$, called the right inverse of A, such that

$$AQ = I : R(A) \longrightarrow R(A), \qquad QA = I - U : Z \longrightarrow N(A)^{\perp}.$$

Let Λ be a closed subset of a Banach space, such that $Int\Lambda \neq \emptyset$. If $N : \Lambda \times X \longrightarrow Z$ is a continuous operator, then the problem

$$Ax - N(x, \lambda) = 0 \tag{3.1}$$

is equivalent to the equations

$$z - QEN(y + z, \lambda) = 0, \tag{3.2}$$

$$(I - E)N(y + z, \lambda) = 0, \tag{3.3}$$

where $x = y + z$, $y \in N(A)$, and $z \in N(A)^{\perp}$. □

Assume that the operator N verifies that

$$N(0, 0) = 0, \qquad \frac{\partial N}{\partial x}(0, 0) = 0,$$

and consider (3.2) for (x, λ) in a neighborhood of $(0, 0)$ in $X \times \Lambda$. Applying the implicit function theorem to (3.2), we deduce the existence of a neighborhood $V \subset N(A) \times \Lambda$ of $(0, 0)$ and a function $z^* : V \longrightarrow N(A)^{\perp}$ with the same regularity of N providing the solution of (3.1). Therefore, if $\{y_1, \ldots , y_h\}$ is an orthonormal basis of $N(A)$, the solution $x(\lambda)$ of (3.1) satisfies

$$x(\lambda) = \sum_{i=1}^{h} c_i(\lambda) y_i + z^* \left(\sum_{i=1}^{h} c_i(\lambda) y_i, \lambda \right) = 0 \tag{3.4}$$

for suitable coefficients $c_1, \ldots , c_h$. Then, $(x, \lambda) \in V$ satisfy (3.1) iff

$$(I - E)\, N \left(\sum_{i=1}^{h} c_i(\lambda) y_i + z^* \left(\sum_{i=1}^{h} c_i(\lambda) y_i, \lambda \right), \lambda \right) = 0, \tag{3.5}$$

which is a finite dimensional system of equations on the constants $c_1, \ldots , c_h$.

Now we have the following result, which is a slight variation of a theorem due to J. H. Albert [2].

THEOREM 3.2. *Let E be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and let Λ be a Banach space. Let $P : D(P) \subset E \to E$ be a self-adjoint operator densely defined* *in E. Assume that λ is an eigenvalue of multiplicity h of P, and let $\phi_1, \ldots, \phi_h$ be the orthonormal eigenfunctions associated to λ. Moreover, assume that there exists a bounded linear operator $Q : E \rightarrow E$, such that $Q\Pi_N = 0$ and $Q(P+\lambda) = I - \Pi_N$, Π_N being the orthogonal projection in $N = \operatorname{Ker}(P+\lambda)$.*

Let $R(u)$ be an analytic self-adjoint map in $B(E,F)$ for every u in a neighborhood of $u = 0$ in Λ, such that $R(0) = 0$ and $P(u) = P + R(u)$.

Then there exist h analytic functions defined in a neighborhood of $u = 0$ in Λ with values in $\mathbb{R}$, $u \rightarrow \lambda_i(u)$, and h analytic functions $u \rightarrow \phi_i(u)$, with values in E, $i = 1, \ldots, h$, defined in a neighborhood of $u = 0$ in Λ, such that the following hold.

1. $\lambda_j(0) = \lambda, \qquad j = 1, \ldots, h.$
2. *For all u small enough, $(\lambda_j(u), \phi_j(u))$ is a solution of the eigenvalue problem $P(u)\phi_j(u) = \lambda_j(u)\phi_j(u)$.*
3. *For all u small enough the set $\{\phi_1(u), \ldots, \phi_h(u)\}$ is orthonormal in E.*
4. *For each interval $I \subset \mathbb{R}$ such that $\overline{I}$ contains only the eigenvalue λ of (P), there exists a neighborhood U of $u = 0$ such that there are exactly h eigenvalues (counting the multiplicity) $\lambda_1(u), \ldots, \lambda_h(u)$ of (P_u) contained on I.* □

Remark 3.1. To prove Theorem 3.2, we prove first that we can find functions $u \rightarrow \lambda(u) \in \mathbb{R}$ and $u \longrightarrow \phi(u)$ such that $\phi(u)$ is an eigenfunction of (P_u) associated to the eigenvalue $\lambda(u)$. To find the other $h-1$ branches of eigenvalues and eigenfunctions, we apply in an iterative form the method described in the following proposition. □

Proposition 3.3. *Under the hypotheses of Theorem* 3.2, *if λ is an eigenvalue of multiplicity h of P and $\phi_1, \ldots, \phi_h$ are orthonormal eigenfunctions associated to λ, there exists at least a function $u \longrightarrow (\lambda(u), \phi(u)) \in \mathbb{R} \times E$ which is analytic in a neighborhood of $u = 0$ in Λ such that*

1. *$\lambda(0) = \lambda$, and*
2. *$\phi(u)$ is an eigenfunction of $P(u)$, associated to the eigenvalue $\lambda(u)$.* □

Proof of Proposition 3.3. Let λ be an eigenvalue of multiplicity h of P, and let $\phi_1, \ldots, \phi_h$ be the orthonormal eigenfunctions associated to λ.

Suppose that the maps $u \rightarrow \lambda(u)$, $u \rightarrow \phi(u)$, such that

$$(P(u) + \lambda(u))\phi(u) = 0, \tag{3.6}$$

do exist. Then

$$\begin{aligned} (P+\lambda)\phi(u) &= (P + \lambda - P(u) - \lambda(u) + P(u) + \lambda(u))\phi(u) \\ &= (-R(u) + \lambda - \lambda(u))\phi(u) \\ &= -(R(u) + \lambda(u) - \lambda)\phi(u). \end{aligned} \tag{3.7}$$

Since $Q(P+\lambda) = I - \Pi_N$, we obtain that

$$\phi(u) = -[Q(R(u) + \lambda(u) - \lambda)]\phi(u) + \psi(u), \tag{3.8}$$

where $\psi(u) \in N = \operatorname{Ker}(P+\lambda)$.

Thus

$$\psi(u) = [I + Q(R(u) + \lambda(u) - \lambda)]\phi(u), \tag{3.9}$$

and therefore

$$\phi(u) = [I + Q(R(u) + \lambda(u) - \lambda)]^{-1}\psi(u). \tag{3.10}$$

Moreover, the map $[I + Q\,(R\,(u) + \lambda\,(u) - \lambda)]$ has an inverse in a neighborhood of $u = 0$ in Λ.

Thus, if we know the functions $u \to \lambda\,(u)$ and $u \to \psi\,(u)$, we can obtain the map $u \to \phi\,(u)$.

Let $\phi_1, \ldots, \phi_h$ be an orthonormal basis of $N = \operatorname{Ker}(P + \lambda)$. We must find constants $c_j(u)$ such that

$$\psi\,(u) = \sum_{j=1}^{h} c_j\,(u)\,\phi_j. \tag{3.11}$$

We can see that

$$[R\,(u) + \lambda\,(u) - \lambda]\,\phi\,(u) \in N^{\perp}, \tag{3.12}$$

because, according to (3.6), we have that $[R\,(u) + \lambda\,(u) - \lambda]\,\phi\,(u) \in R\,(P + \lambda)$.

Thus

$$\begin{aligned} 0 &= \langle [R\,(u) + \lambda\,(u) - \lambda]\,\phi\,(u)\,, \phi_j \rangle \\ &= \langle [R\,(u) + \lambda\,(u) - \lambda]\,\{I + Q\,[R\,(u) + \lambda\,(u) - \lambda]^{-1}\}\psi\,(u)\,, \phi_j \rangle \\ &= \sum_{i=1}^{h} c_j\,(u)\,\langle [R\,(u) + \lambda\,(u) - \lambda]\,\{I + Q\,[R\,(u) + \lambda\,(u) - \lambda]^{-1}\}\phi_i, \phi_j \rangle, \end{aligned} \tag{3.13}$$

which is a linear system of equations on the unknowns $c_j(u)$.

This system has a nontrivial solution iff

$$\det(\langle [R\,(u) + \lambda\,(u) - \lambda]\,\{I + Q\,[R\,(u) + \lambda\,(u) - \lambda]^{-1}\}\phi_i, \phi_j \rangle) = 0. \tag{3.14}$$

Now we show the existence of the constants $c_1\,(u)\,, \ldots, c_h\,(u)$, not all of them being zero simultaneously.

We replace $\lambda(u) - \lambda$ by α, and we define

$$f_{ij}\,(\alpha, u) = \langle [R\,(u) + \alpha]\{I + Q[R\,(u) + \alpha]^{-1}\}\phi_i, \phi_j \rangle \tag{3.15}$$

and

$$F\,(\alpha, u) = \det\,(f_{ij}\,(\alpha, u))\,. \tag{3.16}$$

For u small enough, the map $u \longrightarrow [I + Q\,[R(u) + \alpha])]^{-1}$ is well defined. Indeed, for $\alpha = 0$ and $u = 0$ we have that $[I + QR(0)] = I$, and the map is analytic in a neighborhood of $u = 0$ in Λ. On the other hand, as we mentioned above, if $F(\alpha, u) = 0$, system (3.13) has a nontrivial solution $c_1(u), \ldots, c_h(u)$, and then

$$\lambda(u) = \lambda + \alpha \tag{3.17}$$

is an eigenvalue of $P(u)$.

Moreover, from (3.8) and (3.11) we deduce that

$$\phi(u) = \sum_{j=1}^{h} c_j(u)\,[I + Q\,(R(u) + \lambda(u) - \lambda)]^{-1}\,(v_j, p_j) \tag{3.18}$$

is an eigenfunction of $P(u)$ associated to the eigenvalue $\lambda(u)$.

According to our previous discussion, for these values of $\alpha(u)$ and setting $\lambda(u) = \lambda + \alpha(u)$, system (3.13) admits a solution $c_1(u), \dots, c_h(u)$, not all the components being zero simultaneously. We have that

$$\begin{aligned} f_{ij}(\alpha, 0) &= \langle [R(0)+\alpha]\{I + Q[R(0)+\alpha]\}^{-1}\phi_i, \phi_j\rangle \\ &= \langle \alpha \{I + \alpha Q\}^{-1}\phi_i, \phi_j\rangle \\ &= \alpha\langle \{I + \alpha Q\}^{-1}\phi_i, \phi_j\rangle \\ &= \alpha\delta_{ij}, \end{aligned} \tag{3.19}$$

because $[I + \alpha Q]\phi_i = \phi_i$.

Therefore, we have that $F(\alpha, 0) = \det(\alpha I) = \alpha^h$.

Applying the Weierstrass preparation theorem, we deduce that

$$F(\alpha, u) = \left(\alpha^h + a_1(u)\alpha^{h-1} + \cdots + a_h(u)\right) E(\alpha, u)$$

with $E(\alpha, u) \neq 0$ in a neighborhood of $(0, 0)$. Then for (α, u) small enough we have that $E(\alpha, u) \neq 0$, and functions $a_j(u)$ are analytic in a neighborhood of $u = 0$.

Then, $F(\alpha, u) = 0$ iff

$$\alpha^h + a_1(u)\alpha^{h-1} + \cdots + a_h(u) = 0. \tag{3.20}$$

Let $\alpha_j(u), j = 1, \dots, h$ be the complex roots of (3.20). Then there exist constants $c_1(u), \dots, c_h(u)$, not all vanishing simultaneously, which are the solution of system (3.13).

Thus, from (3.18) we obtain that

$$\phi(u) = \sum_{j=1}^{h} c_j(u)\left[I + Q\left(R(u) + (\lambda(u) - \lambda)\right)\right]^{-1}\phi_j$$

and $\lambda(u) = \lambda + \alpha_1(u)$ constitute an eigenpair.

Notice that if $c_j(u)$ is complex, it is enough to consider the real part $\mathcal{R}c_j(u)$ to get a real eigenfunction. Since the operator $P(u)$ is self-adjoint, we have that $\alpha_j(u)$ is real, which completes the proof of Proposition 3.3. □

Remark 3.2. Proposition 3.3 provides the existence of one branch of eigenpairs associated to the root $\alpha(u)$ of (3.20). We do not use the eigenpairs associated to the other roots α_j by now since, so far, we do not know whether they coincide or not with the eigenpair associated to $\alpha_1(u)$. □

Now we prove Theorem 3.2.

Proof of Theorem 3.2. Using induction on h, we prove the existence of the h analytic functions $u \to (\lambda_i(u), \phi_i(u))$, such that

$$(P(u) + \lambda_i(u))\phi_i(u) = 0. \tag{3.21}$$

From Proposition 3.3, there exists an analytic function $u \to (\lambda_1(u), \phi_1(u))$ defined in a neighborhood of $u = 0$ in Λ with values in $\mathbb{R} \times E$, which verifies (3.21).

Therefore, Theorem 3.2 holds for $h = 1$. We must prove it for $h \geq 2$.

Let $\Pi_1(u) : E \longrightarrow E$ be the orthogonal projection on the eigenspace generated by $\phi_1(u)$. Then we define the map

$$B(u) = P(u) - \Pi_1(u). \tag{3.22}$$

Then

$$B(0)\phi_j = \left(P(0) - \Pi_1(0)\right)\phi_j = \lambda\phi_j - \delta_{1j}\phi_j; \tag{3.23}$$

that is,

$$B(0)\phi_j = \lambda\phi_j, \quad j = 2, \dots, h,$$

and

$$B(0)\phi_1 = (\lambda - 1)\phi_1.$$

Then, λ is an eigenvalue of multiplicity $h-1$ of the operator $B = B(0)$, with eigenfunctions $v_2, \dots, v_h$.

Note that other linearly independent eigenfunctions of B associated to λ do not exist. Indeed, if ϕ is another eigenfunction of B associated to the eigenvalue λ such that $\langle \phi, \phi_j \rangle = 0$, $j = 2, \dots, h$, then $\langle \phi, \phi_1 \rangle = 0$ (since ϕ_1 is an eigenfunction associated to the eigenvalue $\lambda - 1$) and $B\phi = \lambda\phi$. Then

$$P\phi = B\phi + \Pi_1\phi = B\phi + \langle \phi, \phi_1 \rangle v_1 = \lambda\phi;$$

that is, ϕ is an eigenfunction of P associated to λ, and thus λ is an eigenvalue of multiplicity $h+1$, which is impossible because the multiplicity of λ is h.

We can see that $B(u)$ satisfies the hypotheses of Proposition 3.3. This allows us to apply Proposition 3.3 in an iterative form and to obtain $h-1$ analytic functions in a neighborhood of $u = 0$ in Λ, $u \longrightarrow \lambda_i(u)$, and $u \longrightarrow \phi_i(u)$, with $i = 1, \dots, h$ such that

$$B(u)\phi_i(u) = \lambda_i(u)\phi_i(u).$$

Moreover, the functions $\phi_2(u), \dots, \phi_h(u)$ form an orthonormal set in E.

This shows us the existence of the h branches of eigenpairs.

Now we prove the last part of the theorem.

Since the eigenvalues $u \to \lambda_i(u)$ are analytic in a neighborhood of $u = 0$, there exist constants c_i such that

$$\left|\lambda_i(u) - \lambda_i(v)\right| \le c_i \left\|u - v\right\|.$$

Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n \dots$ be the eigenvalues of the P, and assume that

$$\cdots \le \lambda_{n-1} < \lambda = \lambda_n = \cdots = \lambda_{n+h-1} < \lambda_{n+h} \le \cdots.$$

Let $I \subset \mathbb{R}$ be an interval such that λ is the unique eigenvalue contained in I.

Then there exists $\delta > 0$ such that $I \subset (\lambda_{n-1} + \delta, \lambda_{n+h} - \delta)$. Let $u \in B(0, \delta/c)$, with $c = \max\{c_i : i = 1, \dots, n+h\}$. Then

$$|\lambda_{n-1}(u) - \lambda_{n-1}| \le c_{n-1}\|u\| < c_{n-1}\frac{\delta}{c} \le \delta,$$

and

$$|\lambda_{n+h}(u) - \lambda_{n+h}| \le c_{n+h}\|u\| < c_{n+h}\frac{\delta}{c} \le \delta.$$

Therefore, $\lambda_{n-1}(u) \notin \overline{I}$ and $\lambda_{n+h}(u) \notin \overline{I}$; that is, $P(u)$ has at most h eigenvalues contained in $\overline{I}$ counting multiplicity. This completes the proof of Theorem 3.2. $\square$

3.2. Equivalent formulation for the plate system. Now, for each $u \in W^{5,\infty}(\Omega, \mathbb{R}^d)$, we consider the eigenvalue problem for the plate system

$$\left\{ \begin{array}{rll} \triangle^2 y & = \lambda y & \text{in } \Omega + u, \\ y & = 0 & \text{on } \partial\left(\Omega + u\right), \\ \dfrac{\partial y}{\partial n} & = 0 & \text{on } \partial\left(\Omega + u\right). \end{array} \right. \tag{3.24}$$

Let $\{\lambda\left(u\right), y\left(u\right)\}$ be a solution of (3.24), and define the function $Y\left(u\right) = y\left(u\right) \circ \left(I + u\right).$

Thus, our problem is to find $\lambda\left(u\right) \in \mathbb{R}$ and $Y\left(u\right)$ such that

$$\left\{ \begin{array}{rll} D_j^2\left(u\right)\left(Jac\left(I+u\right) D_i^2\left(u\right) Y\left(u\right)\right) & = \lambda\left(u\right) Y\left(u\right) Jac\left(I+u\right) & \text{in } \Omega, \\ Y\left(u\right) & \in H_0^2\left(\Omega\right), & \end{array} \right. \tag{3.25}$$

where

$$D_i\left(u\right) g = \sum_j M_{ij}\left(u\right) \partial_j f,$$

with $M_{ij}\left(u\right)$ defined as in (2.8) and $g = f \circ \left(I + u\right).$

3.3. Regularity of the eigenvalues and eigenfunctions. Now, we will see the existence of the branches of eigenvalues $u \rightarrow \lambda(u)$ and eigenfunctions $u \rightarrow y(u)$ of the plate system.

LEMMA 3.4. *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set of class C^4.*

Then the map

$$P : W^{5,\infty}(\Omega, \mathbb{R}^d) \longrightarrow \mathcal{L}\left(H_0^2\left(\Omega\right); H^{-2}\left(\Omega\right)\right),$$

such that

$$P(u)\phi = \frac{1}{Jac\left(I+u\right)} D_j^2\left(u\right)\left(Jac\left(I+u\right) D_i^2\left(u\right)\phi\right) \tag{3.26}$$

is analytic in a neighborhood of $u = 0$ in $W^{5,\infty}(\Omega, \mathbb{R}^d)$. □

Proof. We can see that $Jac(I + u)$ is a polynomial on the first partial derivatives of u. Then the map $u \longrightarrow Jac(I + u)$ is analytic in a neighborhood of $u = 0$ in $W^{5,\infty}(\Omega, \mathbb{R}^d)$. On the other hand,

$$M_{ij}(u) = \frac{1}{Jac(I+u)}(\delta_{ij} + a_{ij}),$$

where $a_{ij}(u)$ is the minor of the matrix $M^{-1}(u)$ associated to its ijth element, which is also a polynomial of the first partial derivatives of u. Moreover, for u small enough $Jac(I + u) > 0$. Therefore, $u \longrightarrow M(u)$ is analytic in a neighborhood of $u = 0$ in $W^{5,\infty}(\Omega, \mathbb{R}^d)$ as well.

Since $D(u)\varphi = M(u)\nabla\varphi$, from the analyticity of the functions $u \longrightarrow Jac(I + u)$ and $u \longrightarrow M(u)$ we obtain that the map $P(u)$ is analytic in a neighborhood of $u = 0$ on $W^{5,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ with values in $\mathcal{L}\left(H_0^2\left(\Omega\right); H^{-2}\left(\Omega\right)\right)$, and the proof is complete. □

Now, we can apply Theorem 3.2 to the operator $P(u)$ to obtain the existence of the h analytic branches $u \to (\lambda_i(u), y_i(u))$, with $i = 1, \ldots, h$.

From Lemma 3.1 we have that the map $A = P - \lambda I$ has a right inverse operator Q which satisfies the hypotheses of Theorem 3.2. Furthermore, we have that if $u \in W^{5,\infty}(\Omega, \mathbb{R}^d)$, the new domain $\Omega + u$ has a boundary of class C^4, and then, the eigenfunctions satisfy that $y_i \in H^4(\Omega) \cap H_0^2(\Omega)$. Thus we have the following result.

Theorem 3.5. *Let $\Omega \subset \mathbb{R}^d$ be an open bounded domain of class C^4. Let λ be an eigenvalue of multiplicity h of the plate system* (1.1) *for $u = 0$ with associated eigenfunctions $y_1, \ldots, y_h$.*

Then, there exist h analytic functions with values in $\mathbb{R}$, $u \to \lambda_i(u)$, and h analytic functions $u \to y_i(u)$, with values in $H^4(\Omega + u) \cap H_0^2(\Omega + u)$, $i = 1, \ldots, h$, defined in a neighborhood of $u = 0$ in $W^{5,\infty}(\Omega, \mathbb{R}^d)$, such that the following hold.

1. $\lambda_j(0) = \lambda, \qquad j = 1, \ldots, h.$
2. *For all u small enough, $(\lambda_j(u), y_j(u))$ is a solution of the plate system defined in the new domain $\Omega + u$.*
3. *For all u small enough, the set $\{y(u), \ldots, y(u)\}$ is orthonormal in $L^2(\Omega{+}u)$.*
4. *For each interval $I \subset \mathbb{R}$ such that $\overline{I}$ contains only the eigenvalue λ of* (1.1), *there exists a neighborhood U of $u = 0$ such that there are exactly h eigenvalues (counting the multiplicity) $\lambda_1(u), \ldots, \lambda_h(u)$ of (P_u) contained on I.* □

4. Local variations of the eigenvalues and the eigenfunctions. Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with boundary of class C^4. Let λ be an eigenvalue of (1.1) of multiplicity h, and let y_i, $i = 1, \ldots, h$ be the associated eigenfunctions, normalized in $L^2(\Omega)$.

Let $y_i(u) \in H^4(\Omega + u) \cap H_0^2(\Omega + u)$, $i = 1, \ldots, h$, be the eigenfunctions of (3.24) associated to the eigenvalue $\lambda_i(u)$, where $\lambda = \lambda_i(0)$, $y_i(0) = y_i$, $i = 1, \ldots, h$.

According to the results of the previous section, the branches of the eigenvalues $u \longrightarrow \lambda_i(u) \in \mathbb{R}$ and the eigenfunctions $u \longrightarrow y_i(u) \in H^4(\Omega + u) \cap H_0^2(\Omega + u)$ are analytic with respect to the perturbation parameter u in a neighborhood of $u = 0$ in $W^{5,\infty}(\Omega, \mathbb{R}^d)$. The first local variation of the branches solves the system

$$\left\{ \begin{array}{lll} \triangle^2 y_i'(u) & = \lambda_i'(u) y_i + \lambda y_i'(u) & \text{in } \Omega, \\ y_i'(u) & = 0 & \text{on } \partial\Omega, \\ \dfrac{\partial y_i'(u)}{\partial n} & = -(u \cdot n) \dfrac{\partial^2 y}{\partial n^2} & \text{on } \partial\Omega. \end{array} \right. \tag{4.1}$$

The following result provides an identity for the local derivative of the eigenvalues.

Lemma 4.1. *Under the above conditions, the first local derivatives of the eigenvalues verify*

$$\delta_{ij} \lambda_i'(u) = -\int_{\partial\Omega} (u \cdot n) \frac{\partial y_i}{\partial n} \cdot \frac{\partial^2 y_j}{\partial n^2} \quad \forall i, j = 1, \ldots, h. \qquad \square \tag{4.2}$$

Note that here and in what follows, $\lambda_i'(u)$ denotes the derivative of λ_i at $u = 0$ in the direction u.

Proof. Multiplying (4.1) by $w \in H_0^2(\Omega)$, we obtain that

$$\int_\Omega \triangle y_i'(u) \triangle w = \lambda_i'(u) \int_\Omega y_i w + \lambda \int_\Omega y_i'(u) w. \tag{4.3}$$

Taking $w = y_j$ in (4.3), we have that

$$\int_\Omega \triangle y_i'(u) \triangle y_j = \lambda_i'(u) \int_\Omega y_i y_j + \lambda \int_\Omega y_i'(u) y_j.$$

Since $y_i \in H^4(\Omega), i = 1, \dots, h$ (see [10, Theorem 7.1.2]), integrating by parts, we deduce that

$$\int_\Omega \triangle^2 y_j \, y_i'(u) + \int_{\partial\Omega} \frac{\partial y_i'(u)}{\partial n} \triangle y_j(u) = \lambda_i'(u)\,\delta_{ij} + \lambda \int_\Omega y_i'(u)\, y_j.$$

Therefore,

$$\begin{aligned}\delta_{ij}\lambda_i'(u) &= \int_{\partial\Omega} \frac{\partial y_i'(u)}{\partial n} \triangle y_j \\ &= -\int_{\partial\Omega} (u\cdot n)\, \frac{\partial^2 y_i}{\partial n^2} \triangle y_j = -\int_{\partial\Omega} (u \cdot n)\, \frac{\partial^2 y_i}{\partial n^2} \frac{\partial^2 y_j}{\partial n^2}.\end{aligned} \tag{4.4}$$

The proof is complete. □

5. Proof of Theorem 1.1. In this section, we prove the generic unique continuation property stated in Theorem 1.1.

First, we state a unique continuation result for the evolution plate system, which is a consequence of the classical Holmgrem uniqueness theorem and which is needed in the proof. Being more precise, in the proof of Theorem 1.1 we use this result only for the eigenfunctions of the plate system or, more precisely, for the corresponding separated variables solutions of (5.1). However, we state the result for general solutions of the evolution plate system for the sake of completeness.

Lemma 5.1 (see [18, Lemma 3.6, p. 276]). *Let $\Omega \subset \mathbb{R}^d$ be an open bounded domain with boundary of class C^4. Let $\Gamma_0 \subset \partial\Omega$ be a nonempty open set, and let $T > 0$.*

Then, if y solves

$$\begin{cases} y'' + \triangle^2 y &= 0 \quad \text{in } \Omega \times (0,T), \\ y &= 0 \quad \text{on } \partial\Omega \times (0,T), \\ \dfrac{\partial y}{\partial n} &= 0 \quad \text{on } \partial\Omega \times (0,T), \\ \dfrac{\partial^2 y}{\partial n^2} &= 0 \quad \text{on } \Gamma_0 \times (0,T), \\ \dfrac{\partial \triangle y}{\partial n} &= 0 \quad \text{on } \Gamma_0 \times (0,T), \end{cases} \tag{5.1}$$

then, necessarily, $y \equiv 0$. □

Proof of Theorem 1.1. Let $\Gamma_0 \subseteq \partial\Omega$ be an open nonempty set.

Assume that λ is an eigenvalue of (1.1) of multiplicity h, and let $y_i,\ i = 1, \dots, h$, be the associated eigenfunctions.

We define the set

$$A_0 = W_0 = \left\{u \in W^{5,\infty}(\Omega, \mathbb{R}^d) \,:\, u = 0 \text{ on } \partial\Omega \setminus \Gamma_0\right\},$$

and for each $n \in \mathbb{N}$ we consider the set

$$\begin{aligned}A_n = \{u \in W^{5,\infty}(\Omega, \mathbb{R}^d) \,:\, u = 0 \text{ on } \partial\Omega \setminus \Gamma_0 &\text{ and the unique continuation} \\ &\text{property of Theorem 1.1 holds for the first } n \text{ branches of eigenvalues}\}.\end{aligned}$$

Note that $A_{n+1} \subset A_n$ for all $n \geq 0$. We will prove that A_n is an open subset of W_0 and A_{n+1} is a dense subset of A_n for all $n \geq 0$. Thus, by using Baire's lemma (see Lemmas 2.1 and 2.2), we will prove that $\bigcap_{n\in\mathbb{N}} A_n$ is residual in W_0.

Obviously, this completes the proof since $\bigcap_{n\in\mathbb{N}} A_n$ coincides with the set of admissible perturbations of the domain Ω such that the unique continuation property of Theorem 1.1 holds simultaneously for all the branches of eigenvalues.

To apply Baire's lemma, the following properties of the sets A_n are needed.

(i) $\boldsymbol{A_n}$ is open in $\boldsymbol{W_0}$.

It is clear that A_0 is an open set in W_0. To see that each set A_n is open in $W^{5,\infty}(\Omega, \mathbb{R}^d) \cap W_0$ for $n \geq 1$, we argue by contradiction. Suppose that A_n^c is not closed. Then, there exists a sequence of deformations $\{u_k\}_k \subset A_n^c$ such that u_k converges to $\widetilde{u} \in A_n$ as $k \to \infty$.

Since $\{u_k\}_k \subset A_n^c$, there exist $\{\lambda(u_k), y(u_k)\}_k$ such that

$$\begin{cases} \triangle^2 y(u_k) & = \lambda(u_k)\, y(u_k) & \text{in } \Omega + u_k, \\ y(u_k) & = 0 & \text{on } \partial\Omega + u_k, \\ \dfrac{\partial y(u_k)}{\partial n(\Omega + u_k)} & = 0 & \text{on } \partial\Omega + u_k, \\ \displaystyle\int_{\Omega+u_k} |y(u_k)|^2 & = 1, & \end{cases}$$

and

$$\frac{\partial^2 y(u_k)}{\partial n^2(\Omega + u_k)} = 0 \quad \text{on } \Gamma_0 + u_k,$$

$\lambda(u_k)$ belonging to one of the branches $\lambda_1(u), \dots, \lambda_n(u)$ and $y(u_k)$ being the corresponding eigenfunction.

From Theorem 3.5 we have that the branches $u \to \lambda(u)$, $u \to y(u)$ are analytic functions in a neighborhood of $u = 0$ in $W^{5,\infty}(\Omega, \mathbb{R}^d)$ with values in $\mathbb{R}$ and $H^4(\Omega + u) \cap H_0^2(\Omega + u)$, respectively. Thus the eigenpair $(\lambda(\widetilde{u}), y(\widetilde{u}))$ is a solution of the problem

$$\begin{cases} \triangle^2 y(\widetilde{u}) & = \lambda(\widetilde{u})\, y(\widetilde{u}) & \text{in } \Omega + \widetilde{u}, \\ y(\widetilde{u}) & = 0 & \text{on } \partial\Omega + \widetilde{u}, \\ \dfrac{\partial y(\widetilde{u})}{\partial n(\Omega + \widetilde{u})} & = 0 & \text{on } \partial\Omega + \widetilde{u}, \\ \dfrac{\partial^2 y(\widetilde{u})}{\partial n^2(\Omega + \widetilde{u})} & = 0 & \text{on } \Gamma_0 + \widetilde{u}, \\ \displaystyle\int_{\Omega+\widetilde{u}} |y(\widetilde{u})|^2 & = 1 & \end{cases}$$

for some $\lambda(\widetilde{u})$ belonging to one of the first n branches $\lambda_1(\widetilde{u}), \dots, \lambda_h(\widetilde{u})$. But this is impossible, since $\widetilde{u} \in A_n$. This shows that A_n is an open set.

(ii) $\boldsymbol{A_{n+1}}$ is dense in $\boldsymbol{A_n}$.

Now we will see that A_{n+1} is dense in A_n for all $n \geq 0$; in particular, A_1 is dense in $A_0 = W_0$.

Suppose that A_{n+1} is not dense in A_n. Then, there exists $u \in A_n \setminus A_{n+1}$ and a neighborhood $\mathcal{V}$ of u such that $\mathcal{V} \subset A_n \setminus A_{n+1}$. Without lost of generality, we may assume that $u = 0$.

For any $u \in \mathcal{V} \subset A_n \setminus A_{n+1}$ there exists a nontrivial eigenfunction $y\,(u) \in H_0^2\,(\Omega + u)$ associated to the $(n+1)$th eigenvalue $\lambda\,(u) = \lambda_{n+1}\,(u)$ such that

$$\frac{\partial^2 y\,(u)}{\partial^2 n\,(\Omega + u)} = 0 \text{ on } \Gamma_0 + u. \tag{5.2}$$

On the other hand, from Theorem 2.6, the local variations satisfy

$$\left\{\begin{array}{rcll} \triangle^2 y'\,(u) & = & \lambda'\,(u)\, y + \lambda y'\,(u) & \text{in } \Omega, \\ y'\,(u) & = & 0 & \text{on } \partial\Omega, \\ \dfrac{\partial y'\,(u)}{\partial n} & = & 0 & \text{on } \partial\Omega, \\ \dfrac{\partial^2 y'\,(u)}{\partial n^2} & = & -\,(u \cdot n)\, \dfrac{\partial^3 y}{\partial n^3} & \text{on } \Gamma_0 \end{array}\right. \tag{5.3}$$

for all $u \in \mathcal{V}$, where

$$y\,(0) = y, \quad \lambda\,(0) = \lambda.$$

Since $u = 0$ on $\partial\Omega \setminus \Gamma_0$, we have that

$$\lambda'\,(u) = -\int_{\partial\Omega} (u \cdot n) \left| \frac{\partial^2 y}{\partial n^2} \right|^2 = 0,$$

and therefore

$$\left\{\begin{array}{rcll} \triangle^2 y'\,(u) & = & \lambda y'\,(u) & \text{in } \Omega, \\ y'\,(u) & = & 0 & \text{on } \partial\Omega, \\ \dfrac{\partial y'\,(u)}{\partial n} & = & 0 & \text{on } \partial\Omega, \\ \dfrac{\partial^2 y'\,(u)}{\partial^2 n} & = & -\,(u \cdot n)\, \dfrac{\partial^3 y}{\partial n^3} & \text{on } \Gamma_0. \end{array}\right. \tag{5.4}$$

That is, for each $u \in \mathcal{V}$, $y'\,(u)$ is an eigenfunction of the plate system associated to the eigenvalue λ.

We now distinguish two cases.

Case 1. λ is a simple eigenvalue. Since $y'\,(u)$ is an eigenfunction of the plate system associated to the eigenvalue λ, there exists a constant c_u such that $y'\,(u) = c_u y$, and therefore

$$\frac{\partial^2 y'\,(u)}{\partial n^2} = -\,(u \cdot n)\, \frac{\partial^3 y}{\partial n^3} = c_u \frac{\partial^2 y}{\partial n^2} = 0 \text{ on } \Gamma_0.$$

Thus, we obtain that $\frac{\partial^3 y}{\partial n^3} = 0$ on Γ_0 as well.

From Lemma 5.1 we conclude that $y \equiv 0$ in Ω, which is impossible since y is an eigenfunction of the plate system.

Case 2. λ is a multiple eigenvalue. First, we assume that the eigenvalue λ has multiplicity two. Let y_1, y_2 be the associated eigenfunctions.

We will show that there exists a perturbation u, as small as we want, such that the eigenvalue $\lambda(u)$ is simple or the unique continuation property holds for the branch $u \rightarrow y_1(u)$. Notice that the case where $\lambda(u)$ is simple has been addressed in Case 1 above.

We can proceed in an analogous form considering the branch $u \to y_2(u)$.

We argue by contradiction.

We suppose that the eigenvalue $\lambda(u)$ remains of multiplicity two and the corresponding eigenfunction $y_1(u)$ satisfies (5.2) in a neighborhood $\mathcal{V}$ of $u = 0$. If $\lambda(u)$ is simple, we can proceed as in Case 1. In other words,

$$\frac{\partial^2 y_1(u)}{\partial n^2(\Omega + u)} = 0 \quad \text{on } \Gamma_0 + u \ \forall u \in \mathcal{V}.$$

Then, as in the case of simple eigenvalues, the local variation $y_1'(u)$ is an eigenfunction associated to the eigenvalue λ. Then, there exist constants $c_1(u)$, $c_2(u)$ such that $y_1'(u) = c_1(u)\, y_1 + c_2(u)\, y_2$. Thus, from (5.4) we have that

$$\frac{\partial^2 y_1'(u)}{\partial n^2} = -\left(u \cdot n\right) \frac{\partial^3 y_1}{\partial n^3} \text{ on } \Gamma_0$$

and

$$\frac{\partial^2 y_1'(u)}{\partial n^2} = c_1(u) \frac{\partial^2 y_1}{\partial n^2} + c_2(u) \frac{\partial^2 y_2}{\partial n^2} = c_2(u) \frac{\partial^2 y_2}{\partial n^2} \quad \text{on } \Gamma_0.$$

Hence, for every pair of deformations u_1 and u_2 we have that

$$c_2(u_1) \frac{\partial^2 y_2}{\partial n^2} = -\left(u_1 \cdot n\right) \frac{\partial^3 y_1}{\partial n^3} \text{ on } \Gamma_0,$$

$$c_2(u_2) \frac{\partial^2 y_2}{\partial n^2} = -\left(u_2 \cdot n\right) \frac{\partial^3 y_1}{\partial n^3} \text{ on } \Gamma_0.$$

If $\frac{\partial^3 y_1}{\partial n^3} = 0$ on Γ_0, in view of Lemma 5.1 we immediately deduce that $y_1 \equiv 0$, which is a contradiction. Thus, we can assume that $\frac{\partial^3 y_1}{\partial n^3} \not\equiv 0$ on Γ_0 and consequently that $\frac{\partial^2 y_2}{\partial n^2} \not\equiv 0$ on Γ_0 as well.

Then

$$\frac{(u_1 \cdot n)}{(u_2 \cdot n)} = \frac{c_2(u_1) \dfrac{\partial^2 y_2}{\partial n^2}}{c_2(u_2) \dfrac{\partial^2 y_2}{\partial n^2}} = \frac{c_2(u_1)}{c_2(u_2)} = \text{constant}.$$

That is, necessarily $(u_1 \cdot n) = c\,(u_2 \cdot n)$ for a suitable constant c. This is impossible since the functions u_1 and u_2 may be chosen arbitrarily.

Thus we reach a contradiction.

Assume now that λ is an eigenvalue of multiplicity $h > 2$, and let $y_1, \ldots, y_h$ be the associated eigenfunctions normalized in $L^2(\Omega)$.

We claim that there exists a deformation u, arbitrarily small, such that the eigenvalue $\lambda(u)$ has multiplicity at most $h - 1$ or the unique continuation property holds for the branch $y_1(u)$.

To prove this we argue by contradiction. Suppose that $\lambda(u)$ is an eigenvalue of multiplicity h for all $u \in \mathcal{V}$, with associated eigenfunctions $y_1(u), \ldots, y_h(u)$. Moreover, assume that

$$\frac{\partial^2 y_1(u)}{\partial n\,(\Omega + u)^2} = 0 \quad \text{on } \Gamma_0 + u \tag{5.5}$$

for each $u \in \mathcal{V}$.

Then, from (5.3) we have that $y_1'(u)$ is an eigenfunction of the plate system associated to the eigenvalue λ for each $u \in \mathcal{V}$.

Thus, there exist constants $c_1(u), \ldots, c_h(u)$ such that

$$y_1'(u) = c_1(u)\, y_1 + \cdots + c_h(u)\, y_h. \tag{5.6}$$

Therefore,

$$\begin{aligned} \frac{\partial^2 y_1'(u)}{\partial n^2} &= -(u \cdot n)\frac{\partial^3 y_1}{\partial n^3} \\ &= c_1(u)\frac{\partial^2 y_1}{\partial n^2} + \cdots + c_h(u)\frac{\partial^2 y_h}{\partial n^2} \\ &= c_2(u)\frac{\partial^2 y_2}{\partial n^2} + \cdots + c_h(u)\frac{\partial^2 y_h}{\partial n^2} \quad \text{on } \Gamma_0. \end{aligned} \tag{5.7}$$

On the other hand,

$$\lambda_i'(u) = -\int_{\Gamma_0} (u \cdot n)\left|\frac{\partial^2 y_i}{\partial n^2}\right|^2 \quad \forall i = 1, \ldots, h. \tag{5.8}$$

From (5.5) we have that

$$\frac{\partial^2 y_1}{\partial n^2} = 0 \quad \text{on } \Gamma_0. \tag{5.9}$$

Since the multiplicity does not decrease in a neighborhood U of $u = 0$, we obtain that

$$\lambda_1'(u) = \cdots = \lambda_h'(u) \tag{5.10}$$

for all $u \in U$ such that $u = 0$ on $\partial\Omega \setminus \Gamma_0$. Moreover, according to (5.7)–(5.9), we also have that $\lambda_1'(u) = 0$. Therefore,

$$\lambda_1'(u) = \cdots = \lambda_h'(u) = 0.$$

Thus

$$\frac{\partial^2 y_i}{\partial n^2} = 0 \quad \text{on } \Gamma_0 \;\; \forall i = 1, \ldots, h. \tag{5.11}$$

From (5.7) and (5.11), we deduce that

$$\frac{\partial^3 y_1}{\partial n^3} = 0 \quad \text{on } \Gamma_0. \tag{5.12}$$

From Lemma 5.1 we have that $y_1 \equiv 0$ on Ω, which is impossible because y_1 is an eigenfunction.

Therefore, there exists a deformation $\widetilde{u} \in \mathcal{V}$ such that the eigenvalue $\lambda_1(\widetilde{u})$ has multiplicity at most $h - 1$ or the unique continuation property holds for $y_1(\widetilde{u})$.

If the unique continuation property does not hold in the new domain $\Omega + \widetilde{u}$, we can apply the same argument in an iterative way and obtain a deformation $\widehat{u}$, arbitrarily small, such that the eigenvalue $\lambda(\widehat{u})$ is of multiplicity two or the unique continuation property holds. Since the case where the multiplicity is two was solved before, this completes the proof of the density of A_{n+1} on A_n for all $n \geq 0$.

Applying Baire's lemma (see Lemmas 2.1 and 2.2), we complete the proof of Theorem 1.1. □

Remark 5.1. Note that in the proof of Theorem 1.1 above, we consider only deformations u such that $u = 0$ on $\partial\Omega \setminus \Gamma_0$. Consequently, the deformations u we use do deform the subset Γ_0 of the boundary.

If we consider deformations which do not deform the set Γ_0 (that is, such that $u = 0$ on Γ_0), the argument of the proof does not apply. That is because we cannot guarantee anymore that the local variations $y_i'(u)$ are eigenfunctions of (1.1). □

6. Proof of Theorem 1.2. The proof of Theorem 1.2 is similar to the proof of the generic simplicity of the eigenvalues of the Stokes system (see [25]). We apply Baire's lemma for a suitable sequence of sets $\{A_n\}_{n\geq 0}$.

Let $n = 1, 2, \dots$, and define the sets

$$A_0 = W_0 = \left\{u \in W^{5,\infty}(\Omega, \mathbb{R}^d) \ : \ u = 0 \ \text{ on } \ \partial\Omega \setminus \Gamma_0\right\}, \tag{6.1}$$

and

(6.2)
$$A_n = \big\{u \in W^{5,\infty}(\Omega, \mathbb{R}^d) \colon u = 0 \ \text{ on } \partial\Omega \setminus \Gamma_0, \ \text{ and the first } n \text{ branches}$$
$$\lambda_1(u), \dots, \lambda_n(u) \text{ of eigenvalues of (3.24) are simple}\big\}.$$

As in the proof of Theorem 1.1, we need to check that A_n is an open subset of W_0 and $A_{n+1}$is a dense subset of A_n for all $n \geq 0$. Then, applying Lemma 2.2, we conclude the proof.

(a) $\boldsymbol{A_n}$ is an open set in $\boldsymbol{W^{5,\infty}(\Omega, \mathbb{R}^d) \cap W_0}$.

It is clear that A_0 is an open set. Let $u \in A_n$ for $n \geq 1$, and then

$$\lambda_i(u) \neq \lambda_j(u) \quad \forall i, j = 1, \dots, n+1, i \neq j.$$

Let

$$\delta = \min\{|\lambda_i(u) - \lambda_j(u)| : i, j = 1, \dots, n, i \neq j\}.$$

Let $u' \in U$, where

$$U = \left\{w \in \mathcal{W} : \|u - w\|_{5,\infty} < \frac{\delta}{2c}\right\}$$

and c is the maximum of the Lipschitz constants for the functions $u' \to \lambda_i(u+u')$, $i = 1, \dots, n+1$.

Then

$$\begin{array}{rcl} \delta & \leq & |\lambda_i(u) - \lambda_j(u)| \\ & \leq & |\lambda_i(u+u') - \lambda_i(u)| + |\lambda_i(u+u') - \lambda_j(u+u')| \\ & + & |\lambda_j(u+u') - \lambda_j(u)| \\ & < & \delta + |\lambda_i(u+u') - \lambda_j(u+u')|. \end{array}$$

Thus

$$|\lambda_i(u') - \lambda_j(u')| > 0, \quad i, j = 1, \dots, n+1,$$

which proves that $u + u' \in A_n$, and the set A_n is open.

(b) A_{n+1} is dense in A_n for all $n \geq 0$.

Let $w \in A_n \setminus A_{n+1}$. Without loss of generality we can assume that $u = 0$. Then, either λ_{n+1} remains to be of multiplicity h in a neighborhood of $u = 0$, or there exists $u \neq 0$ arbitrarily small such that the multiplicity is at most $h - 1$. (In the case of $n = 0$, we have that the first eigenvalue has multiplicity $h \geq 2$; that is, $\lambda = \lambda_1(u) = \cdots = \lambda_h(u)$.) Iterating this argument, it can be shown that the $(n+1)$th eigenvalue becomes simple for suitable arbitrarily small perturbations u or it remains of constant multiplicity $h \geq 2$ in a neighborhood of $u = 0$. If the eigenvalue becomes simple, the proof of the density of A_{n+1} is concluded.

Thus, we can assume that λ_{n+1} is of constant multiplicity $h \geq 2$; that is,

$$\lambda = \lambda_{n+1}(u) = \cdots = \lambda_{n+h}(u)$$

for any u in a neighborhood of $u = 0$.

Let $y_1(u), \ldots, y_h(u)$ be the eigenfunctions associated to λ normalized in $L^2(\Omega + u)$.

In view of the generic unique continuation result of Theorem 1.1, we can also assume that the spectral uniqueness holds in Ω.

Then, from (4.2) we have that

$$\lambda'_{n+i}(\Omega; u)\,\delta_{ij} = -\int_{\partial\Omega} (u \cdot n) \frac{\partial^2 y_i}{\partial n^2} \frac{\partial^2 y_j}{\partial n^2}. \tag{6.3}$$

We will prove that there exists a deformation u such that

$$\lambda'_{n+i}(\Omega; u) \neq \lambda'_{n+j}(\Omega; u) \quad \forall\, i \neq j. \tag{6.4}$$

Assuming for the moment that this holds, we deduce that

$$\lambda_{n+i}(\varepsilon u) \neq \lambda_{n+j}(\varepsilon u) \quad \forall i, j = 1, \ldots, n+h+1, i \neq j,$$

for $\varepsilon > 0$ small enough and then $\varepsilon u \in A_{n+1}$.

We proceed by contradiction. Suppose that (6.4) does not hold. Then there exist $i \neq j$, $i, j \in \{1, \ldots, h\}$, such that

$$\lambda'_{n+i}(\Omega; u) = \lambda'_{n+j}(\Omega; u) \tag{6.5}$$

for all u in a neighborhood of $u = 0$.

Thus, from (4.2) we have that

$$\int_{\partial\Omega} (u \cdot n) \left[\frac{\partial^2 y_i}{\partial n^2} \frac{\partial^2 y_j}{\partial n^2} \right] = 0,$$

and from (6.3) and (6.5) we have that

$$\int_{\partial\Omega} (u \cdot n) \left[\left| \frac{\partial^2 y_i}{\partial n^2} \right|^2 - \left| \frac{\partial^2 y_j}{\partial n^2} \right|^2 \right] = 0$$

for all u in a neighborhood of $u = 0$.

Since $u = 0$ on $\partial\Omega \setminus \Gamma_0$, we deduce that

$$\frac{\partial^2 y_i}{\partial n^2} \frac{\partial^2 y_j}{\partial n^2} = 0 \quad \text{on } \Gamma_0 \tag{6.6}$$

and

$$\left|\frac{\partial^2 y_i}{\partial n^2}\right| = \left|\frac{\partial^2 y_j}{\partial n^2}\right| \quad \text{on } \Gamma_0. \tag{6.7}$$

Therefore, we obtain that

$$\frac{\partial^2 y_i}{\partial n^2} = \frac{\partial^2 y_j}{\partial n^2} = 0 \quad \text{on } \Gamma_0. \tag{6.8}$$

Since the spectral uniqueness property holds in the domain Ω, we obtain that

$$y_i = y_j = 0 \quad \text{in } \Omega.$$

But this is impossible because y_i and y_j are eigenfunctions of the plate system.

Applying Baire's lemma (see Lemmas 2.1 and 2.2) to the sets A_n on the space W_0, we complete the proof.

7. Proof of Theorem 1.3. In this section we analyze the evolution dissipative plate equation (1.3). First, we prove the existence and uniqueness of solutions. Then we derive a generic unique continuation result in an infinite time interval. Finally, we complete the proof of Theorem 1.3.

7.1. Existence and uniqueness of solutions for the evolution plate system. First, we analyze the variational formulation of system (1.3).

Recall that $X_1 = \{\varphi \in H^2(\Omega) \cap H_0^1(\Omega) \;:\; \frac{\partial\varphi}{\partial n} = 0 \text{ on } \partial\Omega \setminus \Gamma_0,\}$, $X_2 = L^2(\Omega)$, and $X = X_1 \times X_2$. Then $X = X_1 \times X_2$ is a Hilbert space endowed with the norm of $H^2(\Omega) \times L^2(\Omega)$.

We also introduce the space $X_3 = \{\varphi \in X_1 \;:\; \triangle^2\varphi \in L^2(\Omega)\}$.

Let $B : X_1 \to X_1'$ be the map

$$\langle Bz, v\rangle_{X_1', X_1} = \int_{\Gamma_0} \frac{\partial z}{\partial n}\frac{\partial v}{\partial n}. \tag{7.1}$$

We can see that the linear map B is continuous and accretive.

Let $z \in X_1$. Then, multiplying (1.3) by z, we have that

$$\begin{aligned} 0 &= \left\langle y'' + \triangle^2 y, z\right\rangle_{X_1 \times X_1'} \\ &= \int_\Omega (y''z + \triangle y \triangle z) + \int_{\partial\Omega}\left(z\frac{\partial(\triangle y)}{\partial n} - \triangle y\frac{\partial z}{\partial n}\right) \\ &= \int_\Omega (y''z + \triangle y\triangle z) + \int_{\Gamma_0}\frac{\partial y'}{\partial n}\frac{\partial z}{\partial n}. \end{aligned} \tag{7.2}$$

We also define the map $A : X_1 \to X_1'$ as

$$\langle Az, v\rangle_{X_1', X_1} = \int_\Omega \triangle z \triangle v. \tag{7.3}$$

Then, (7.2) is equivalent to

$$\langle y'' + Ay + By', v\rangle_{X_1', X_1} = 0 \; \forall v \in X_1. \tag{7.4}$$

We define the operator $\mathcal{A} : D(\mathcal{A}) \subset X \to X$ by

$$\mathcal{A}U = (-U_2, AU_1 + BU_2), \quad \text{where } U = (U_1, U_2),$$

and $D(\mathcal{A}) = \{(y, z) \in X_3 \times X_1 \;:\; \triangle y = -\frac{\partial z}{\partial n} \text{ on } \Gamma_0\}$.

Thus, the problem (1.3) is equivalent to solve

$$\begin{aligned} U' + \mathcal{A}U &= 0 \qquad\qquad \text{in } \mathbb{R}^+, \\ U(0) &= (y_0, y_1), \end{aligned}$$

with $U = (y, y')$.

Proposition 7.1. *The operator $\mathcal{A}$ is m-accretive.* □

Proof. Let $U, V \in D(\mathcal{A})$. Then

$$\begin{aligned} \langle \mathcal{A}U - \mathcal{A}V, U - V \rangle_X &= \langle V_2 - U_2, U_1 - V_1 \rangle_{X_1} \\ &\quad + \langle AU_1 - AV_1 + BU_2 - BV_2, U_2 - V_2 \rangle_{X_1', X_1} \\ &= \int_\Omega \triangle (V_2 - U_2) \triangle (U_1 - V_1) + \int_\Omega \triangle (U_1 - V_1) \triangle (U_2 - V_2) \\ &= + \int_{\Gamma_0} \frac{\partial (U_2 - U_1)}{\partial n} \frac{(\partial (U_2 - V_2)}{\partial n} \\ &= \int_{\Gamma_0} \left| \frac{\partial (U_2 - V_2)}{\partial n} \right|^2 \geq 0, \end{aligned}$$

which proves that $\mathcal{A}$ is an accretive operator.

We must prove that $I + \mathcal{A} : D(\mathcal{A}) \to X$ is onto.

Let $W = (W_1, W_2) \in X$. We must show that there exists $U = (U_1, U_2) \in D(\mathcal{A})$, such that

$$(U_1 - U_2, U_2 + AU_1 + BU_2) = (W_1, W_2). \tag{7.5}$$

System (7.5) is equivalent to

$$U_2 = U_1 - W_1, \qquad U_1 + AU_1 + BU_1 = W_2 + BW_1 + W_1. \tag{7.6}$$

To prove the existence of a solution of (7.6), it is enough to show that for every $f \in X_1'$, there exists $U_1 \in X_1$ such that

$$(I + A + B) U_1 = f. \tag{7.7}$$

We first observe that

$$\langle (I + A + B) v, v \rangle_{X_1', X_1} = \int_\Omega v^2 + |\triangle v|^2 + \int_{\Gamma_0} \left| \frac{\partial v}{\partial n} \right|^2 \geq \alpha \|v\|_{X_1}^2 \quad \forall v \in X_1, \tag{7.8}$$

which shows that the bilinear form associated to $(I + A + B)$ is coercive.

On the other hand, the embedding $H^2(\Omega) \hookrightarrow H^1(\Gamma_0)$ is continuous. Thus, the bilinear form associated to $(I + A + B)$ is continuous. Therefore, from Lax–Milgram's lemma we have that there exists a unique $U_1 \in X_1$ such that (7.7) holds.

Therefore, there exists $(U_1, U_2) = (U_1, U_1 - W_1,) \in X_1 \times X_1$ which satisfies (7.5). Moreover, $U_1 \in X_3$ and $(U_1, U_2) \in D(A)$.

This proves that the map $I + \mathcal{A}$ is onto, and thus the operator $\mathcal{A}$ is an m-accretive operator on $X_3 \times X_1$. □

Remark 7.1. Note that since the operator $\mathcal{A}$ is m-accretive in X, in view of [5, Proposition VII.1, p. 101], we deduce that $D(\mathcal{A})$ is dense on X. □

We have the following result on the existence, uniqueness, and regularity of solutions of (1.3).

PROPOSITION 7.2. *Let* $(y_0, y_1) \in X$. *Then there exists a unique solution* y *of* (1.3) *which verifies*

$$y \in C_b([0,+\infty[; X_1) \cap C_b^1([0,+\infty[; X_2). \tag{7.9}$$

Furthermore, if $(y_0, y_1) \in D(\mathcal{A})$, *the solution* y *verifies*

$$(y, y_t) \in C_b([0,+\infty[; D(A)). \tag{7.10}$$

Moreover, when $(y_0, y_1) \in D(\mathcal{A})$, *the energy* $E : \mathbb{R}^+ \to \mathbb{R}^+$, *defined as*

$$E(t) = \frac{1}{2}\int_\Omega \left[|y_t(x,t)|^2 + |\triangle y(x,t)|^2 \right] dx,$$

is a decreasing and differentiable function on $\mathbb{R}^+$. □

Remark 7.2. Here and in what follows we denote by $C_b^k([0,+\infty[; X)$ the space of functions $C^k([0,+\infty[; X) \cap W^{k,\infty}(0,+\infty; X)$. □

Proof of Proposition 7.2. Since the operator $\mathcal{A}$ m-accretive in X, from the Hille–Yosida theorem we have the following.

1. If $(y_0, y_1) \in X = X_1 \times X_2$, there exists a unique solution $U = (y, y_t)$ of

$$\begin{cases} \dfrac{dU}{dt} + \mathcal{A}U &= 0 \qquad \text{on } (0,+\infty), \\ \qquad U(0) &= (y_0, y_1), \end{cases} \tag{7.11}$$

such that (see, for instance, [5, Theorem VII.5, p. 111])

$$U \in C_b([0,+\infty[; X). \tag{7.12}$$

2. If $(y_0, y_1) \in D(\mathcal{A})$, there exists a unique solution $U = (y, y_t)$ of (7.11) such that (see, for instance, [5, Theorem VII.4, p. 105])

$$U \in C_b^1([0,+\infty[; X) \cap C_b([0,+\infty[; D(\mathcal{A})). \tag{7.13}$$

On the other hand,

$$\frac{dE}{dt}(t) = \int_\Omega [y_{tt}(x,t)\, y_t(x,t) + \triangle y_t(x,t)\, \triangle y(x,t)]\, dx. \tag{7.14}$$

Taking $z = y_t$ in (7.2) and integrating by parts, we have that

$$\frac{dE}{dt}(t) = -\int_{\Gamma_0} \left| \frac{\partial y_t}{\partial n} \right|^2 \leq 0, \tag{7.15}$$

which shows us that the energy E is decreasing. □

7.2. A generic unique continuation result. In order to prove Theorem 1.3, we need a nonstandard unique continuation result for the evolution plate system. This is the object of the following proposition.

PROPOSITION 7.3. *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set of class C^4, and let $\Gamma_0 \subset \partial\Omega$ be a nonempty open set such that the spectral uniqueness property holds.*

If $(y_0, y_1) \in D(\mathcal{A})$ and the corresponding solution of (1.3) is such that

$$\frac{\partial y_t}{\partial n} = 0 \text{ on } \Gamma_0 \times (0,\infty), \tag{7.16}$$

then necessarily $y \equiv 0$. □

Proof. We set $v = y_t$. Then v is a weak solution of the system

$$\left\{\begin{array}{lll} v_{tt} + \triangle^2 v & = 0 & \text{in } \Omega \times (0,\infty), \\ v & = 0 & \text{on } \partial\Omega \times (0,\infty), \\ \dfrac{\partial v}{\partial n} & = 0 & \text{on } \partial\Omega \times (0,\infty), \\ v(x,0) & = v_0 = y_1 & \text{in } \Omega, \\ v_t(x,0) & = v_1 = -\triangle^2 y_0 & \text{in } \Omega, \end{array}\right. \tag{7.17}$$

with initial datum $(v_0, v_1) \in H_0^2(\Omega) \times L^2(\Omega)$, and, furthermore, satisfies the condition

$$\triangle v = 0 \quad \text{on } \Gamma_0 \times (0,\infty). \tag{7.18}$$

We will see that $v = 0$. Thus $y(x,t) \equiv y(x)$. Therefore, $y(x) = y_0$ solves

$$\left\{\begin{array}{lll} \triangle^2 y & = 0 & \text{in } \Omega, \\ y & = 0 & \text{on } \partial\Omega, \\ \dfrac{\partial y}{\partial n} & = 0 & \text{on } \partial\Omega \setminus \Gamma_0, \\ \triangle y & = 0 & \text{on } \Gamma_0 \times (0,\infty). \end{array}\right. \tag{7.19}$$

Multiplying (7.19) by y and integrating by parts, we deduce that $\int_\Omega |\triangle y|^2 \, dx = 0$; that is, $y \equiv 0$, since $y = 0$ on $\partial\Omega$.

Thus, the problem is reduced to show that

$$v \equiv 0. \tag{7.20}$$

To prove (7.20), first we observe that $(v, v_t) \in C_b([0,+\infty); H_0^2(\Omega) \times L^2(\Omega))$.

This can be proved easily by analyzing the well posedness of the conservative plate system (7.17).

In what follows we shall use the notation $X = H_0^2(\Omega) \times L^2(\Omega)$.

The solution v of (7.17) may be developed in a Fourier series. Indeed, let $\{\lambda_n\}_n$ be the eigenvalues of the bilaplacian operator with clamped boundary conditions, and let w_n be the associated eigenfunctions normalized in $L^2(\Omega)$:

$$\left\{\begin{array}{lll} \triangle^2 w_n & = \lambda_n w_n & \text{in } \Omega, \\ w_n = \dfrac{\partial w_n}{\partial n} & = 0 & \text{on } \partial\Omega. \end{array}\right. \tag{7.21}$$

We decompose the initial data in Fourier series:

$$v_0(x) = \sum_{n=1}^{\infty} a_n w_n, \qquad v_1(x) = \sum_{n=1}^{\infty} b_n w_n. \tag{7.22}$$

It is easy to check that

$$\|v_0\|^2_{H_0^2(\Omega)} = \sum_{n=1}^{\infty} \lambda_n |a_n|^2, \qquad \|v_1\|^2_{L^2(\Omega)} = \sum_{n=1}^{\infty} |b_n|^2. \tag{7.23}$$

The solution v of (7.17) may be written as

$$v(x,t) = \sum_{n=1}^{\infty} \left[a_n \cos\left(\sqrt{\lambda_n} t\right) + \frac{b_n}{\lambda_n} \sin\left(\sqrt{\lambda_n} t\right) \right] w_n(x). \tag{7.24}$$

Let us define

$$\mu_k = \sqrt{\lambda_k}, \quad \mu_{-k} = -\sqrt{\lambda_k}, \quad w_{-k}(x) = w_k(x). \tag{7.25}$$

If we define the complex coefficients

$$c_k = \frac{1}{2}\left(a_k - i\frac{b_k}{\sqrt{\lambda_k}} \right), \quad c_{-k} = \frac{1}{2}\left(a_k + i\frac{b_k}{\sqrt{\lambda_k}} \right), \tag{7.26}$$

we can write

$$v(x,t) = \sum_{k \in \mathbb{Z}} c_k e^{i\mu_k t} w_k(x). \tag{7.27}$$

Taking into account that $\triangle v = 0$ on $\Gamma_0 \times (0, \infty)$ and applying Bohr's transform to $\triangle v(x,t)$ on Γ_0 (see [3]), we deduce that

$$c_k \triangle w_k = 0 \quad \text{ on } \Gamma_0 \ \forall k \geq 1. \tag{7.28}$$

Therefore, for every $k \in \mathbb{Z}$, we have that $c_k = 0$ or w_k is a solution of the problem

$$\begin{cases} \triangle^2 w_k & = \lambda_k^2 w_k & \text{in } \Omega, \\ w_k & = 0 & \text{on } \partial\Omega, \\ \dfrac{\partial w_k}{\partial n} & = 0 & \text{on } \partial\Omega, \\ \triangle w_k & = 0 & \text{on } \Gamma_0. \end{cases} \tag{7.29}$$

Since, by assumption, the spectral uniqueness holds in Ω, (7.29) implies that $w_k = 0$ in Ω, which is a contradiction. Thus $c_k = 0$ for all $k \in \mathbb{Z}$ and consequently $v \equiv 0$.

This completes the proof. □

7.3. Proof of Theorem 1.3. Now we prove the stabilization result for system (1.3).

Proof. We distinguish two cases.

1. $(y_0, y_1) \in D(\mathcal{A})$.
2. $(y_0, y_1) \in X$.

Step 1. Regular initial data.

Let $\{S_t\}_{t\geq 0}$ be the contraction semigroup associated to (1.3). Then, given $(y_0, y_1) \in D(\mathcal{A})$, we have that $(y, y_t) = S_t\,(y_0, y_1)$ and $\|y, y_t\|_{D(\mathcal{A})} \leq \|y_0, y_1\|_{D(\mathcal{A})}$.

Therefore, the trajectory $\{y\,(t)\,, y_t\,(t)\}_{t\geq 0}$ is bounded in $D(\mathcal{A})$, and, according to the compactness of the imbedding $D(\mathcal{A}) \hookrightarrow X$, $\{y\,(t)\,, y_t\,(t)\}_{t\geq 0}$ is relatively compact in X.

Moreover, the energy E is a strict Lyapunov functional for the semigroup $\{S_t\}_{t\geq 0}$. Indeed, suppose that $(y_0, y_1) \in D\,(\mathcal{A})$ is such that $E\,(t)$ is constant for all $t \geq 0$. Then y solves (1.3) with

$$\frac{\partial y_t}{\partial n} = 0 \quad \text{on } \Gamma_0 \times (0, \infty)\,. \tag{7.30}$$

From Proposition 7.3, we have that $y \equiv 0$. Therefore, $(y_0, y_1) = (0, 0)\,,$ which is the unique equilibrium point for system (1.3).

Thus, from the La Salle invariance principle (see [6, Theorem 9.2.3, p. 122]), we deduce that the $\omega-$limit of the trajectory $\{y, y_t\}_{t\geq 0}$ in X has a unique point $(y_0, y_1) = (0, 0)\,.$

Therefore,

$$\lim_{t\to+\infty} E\,(t) = 0. \tag{7.31}$$

Step 2. Initial data in *X*.

Let $(y_0, y_1) \in X$. Since $D\,(\mathcal{A})$ is dense in X, there exists a sequence of initial data $(y_0^n, y_1^n) \in D\,(\mathcal{A})$ which converges to (y_0, y_1) in X as $n \to \infty$.

Let y be the solution of (1.3) with initial datum (y_0, y_1), and let y^n be the solution of (1.3) with initial datum $(y_0^n, y_1^n)\,.$

Let

$$E_y\,(t) = \int_\Omega \left[|y_t|^2 + |\triangle y|^2\right].$$

Then

$$0 \leq E_y\,(t) \leq 2E_{y_n}\,(t) + 2E_{y-y_n}\,(t) \leq 2E_{y_n}\,(t) + 2E_{y-y_n}\,(0)\,.$$

Since $(y_0^n, y_1^n) \to (y_0, y_1)$ in X as $n \to \infty$, for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that

$$E_{y-y_n}\,(0) < \varepsilon \quad \text{for } n \geq n_0.$$

Moreover, since $(y_0^{n_0}, y_1^{n_0}) \in D\,(\mathcal{A})\,,$ according to the result of the first step, there exists $t_0 \geq 0$ such that

$$E_{y_{n_0}}\,(t) < \varepsilon, \quad t \geq t_0.$$

Therefore, if $t \geq t_0$, we have that

$$E_y\,(t) \leq 2E_{y_{n_0}}\,(t) + 2E_{y-y_{n_0}}\,(t) < 4\varepsilon.$$

Thus

$$\lim_{t\to+\infty} E_y\,(t) = 0.$$

This completes the proof of Theorem 1.3. □

REFERENCES

[1] J. H. Albert, *Genericity of simple eigenvalues for elliptic pde's*, Proc. Amer. Math. Soc., 48 (1975), pp. 413–418.

[2] J. H. Albert, *Topology of the Nodal and Critical Points Sets for Eigenfunctions of Elliptic Operators*, Ph.D. thesis, M. I. T., Cambridge, MA, 1971.

[3] L. Amerio and G. Prouse, *Almost-Periodic Functions and Functional Equations*, Van Nostrand Reinhold Company, New York, 1971.

[4] J. A. Bello, E. Fernández-Cara, J. Lemoine, and J. Simon, *The differentiability of the drag with respect to the variations of a Lipschitz domain in a Navier–Stokes flow*, SIAM J. Control Optim., 35 (1997), pp. 626–640.

[5] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Masson, Paris, 1983.

[6] T. Cazenave and A. Haraux, *Introduction aux problèmes d'évolution semi-linéaires*, Math. Appl. 1, Ellipses, Paris, France, 1990.

[7] S. Chow and J. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[8] C. V. Coffman, R. J. Duffin, and D. H. Shaffer, *The fundamental mode of vibration of a clamped annular plate is not of one sign*, in Constructive Approaches to Mathematical Models, Academic Press, New York, 1979, pp. 267–277.

[9] R. Courant and D. Hilbert, *Methods of Mathematical Physics* I, Interscience Publishers, New York, 1953.

[10] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 1983.

[11] A. Haraux, *Semi-linear hyperbolic problems in bounded domains*, Math. Rep., 3 (1987), pp. i–xxiv and 1–281.

[12] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Heidelberg, 1980.

[13] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, RAM: Research in Applied Mathematics, Masson, Paris, Wiley, Chichester, UK, 1994.

[14] J. E. Lagnese and J. L. Lions, *Modelling Analysis and Control of Thin Plates*, RAM: Research in Applied Mathematics, Masson, Paris, 1988.

[15] J. E. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Stud. Appl. Math. 10, Philadelphia, PA, 1989.

[16] J. Lemoine, Ph.D. thesis, Blaise Pascal University, Clermont-Ferrand, France, 1995.

[17] B. M. Levitan and V. V. Zhikov, *Almost Periodic Functions and Differential Equations*, Cambridge University Press, Cambridge, UK, 1982.

[18] J. L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*, 1, Rech. Math. Appl. 8, Masson, Paris, 1988.

[19] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[20] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, I, Springer-Verlag, Berlin, Heidelberg, 1972.

[21] J. L. Lions and E. Zuazua, *A generic uniqueness result for the Stokes system and its control theoretical consequences*, in Partial Differential Equations, Lecture Notes in Pure and Appl. Math. 177, P. Marcellini, G. Talenti, and E. Visentini, eds., Marcel Dekker, New York, 1996, pp. 221–235.

[22] A. M. Micheletti, *Perturbazione dello spettro di un operatore ellittico di tipo variazionale, in relazione ad una variazione del campo*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1972), pp. 151–169.

[23] J. Nečas, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Masson, Paris, 1967.

[24] J. Ortega, *Comportamiento asintótico, control y estabilización de algunos sistemas parabólicos y de placas*, Ph.D. thesis, Universidad Complutense de Madrid, Madrid, Spain, 1997.

[25] J. Ortega and E. Zuazua, *Generic Simplicity of the eigenvalues of the Stokes system in two space dimensions*, Adv. Differential Equations, to appear.

[26] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach Science Publishers, New York, 1969.

[27] J. Simon, *Diferenciación con respecto al dominio*, Lecture notes, Universidad de Sevilla, Seville, Spain, 1989.

[28] J. Simon, *Differentiation with respect to the domain in boundary value problems*, Numer. Func. Anal. Optim., 2 (1980), pp. 649–687.

[29] K. Uhlenbeck, *Generic properties of eigenfunctions*, Amer. J. Math., 98 (1976), pp. 1059–1078.

[30] M. M. Vainberg and V. A. Trenogin, *Theory of Branching of Solutions of Nonlinear Equations*, Noordhoff International Publishing, Leyden, The Netherlands, 1974.

[31] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, UK, 1944.

[32] H. Weinberger, *A First Course in Partial Differential Equations with Complex Variables and Transform Methods*, Blaisdell, New York, 1965.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 5, pp. 1615–1632

PURSUIT DIFFERENTIAL GAMES WITH STATE CONSTRAINTS[*]

PIERRE CARDALIAGUET[†], MARC QUINCAMPOIX[†], AND PATRICK SAINT-PIERRE[‡]

Abstract. We prove the existence of a value for pursuit games with state constraints. We also prove that this value is lower semicontinuous.

Key words. differential games, pursuit games, state constraints

AMS subject classification. 90D25

PII. S0363012998349327

Introduction. In this paper, we intend to prove that two-player differential games with state constraints have a value.

We investigate a differential game where the first player, called Ursula, playing with u, controls a first system

$$y'(t) = g(y(t), u(t)), \quad u(t) \in U, \tag{0.1}$$

and has to ensure that the state constraint $y(t) \in K_U$ is fulfilled, while the second player, called Victor, playing with v, controls a second system

$$z'(t) = h(z(t), v(t)), \quad v(t) \in V, \tag{0.2}$$

and has to ensure the state constraint $z(t) \in K_V$. The first player aims the state of the full system $(y(\cdot), z(\cdot))$ at avoiding a target as long as possible, while the second player aims the state of the system at reaching this target in minimal time.

This game is known as the pursuit game. Most of the examples and results of the early theory for differential games are concerned with this problem. (For several examples and for methods of explicit resolution, see Isaacs [26], Flynn [20], Breakwell [10], and Bernhard [9].)

As usually in differential game theory, one can define two value functions for the game: the upper one and the lower one. The purpose of this paper is to give some conditions on the system under which the pursuit game has a value, i.e., that the upper value function is equal to the lower value function. We have to face two difficulties: the presence of state constraints and the fact that the value functions can be discontinuous. Let us point out that in most examples studied in Isaacs's book [26], one has to face at least one of the difficulties and often both.

In the early '70s, the question of the existence of a value for pursuit games was the aim of several papers (Varaiya [37], Varaiya and Lin [38], Osipov [29], Friedman [23], [24], [25], Elliot and Kalton [16], [17], Fleming [19], Krasovskii and Subbotin [27]). These works are mainly dealing with unconstrained problems (except for [25]), and the value function is always continuous. Related problems, such as notions of

[*] Received by the editors December 17, 1998; accepted for publication (in revised form) August 23, 2000; published electronically February 23, 2001.
http://www.siam.org/journals/sicon/39-5/34932.html

[†] Département de Mathématiques, Université de Bretagne Occidentale, 6 Avenue Victor Le Gorgeu, F-29200 Brest, France (Pierre.Cardaliaguet@univ-brest.fr, Marc.Quincampoix@univ-brest.fr).

[‡] Centre de recherche Viabilité, Jeux, Contrôle, Université Paris IX Dauphine, Place du Maréchal de Lattre de Tassigny, F-75775 Paris Cedex, France (stpierre@viab.dauphine.fr).

local asymptotics and local and global stabilizations are studied in the article of Yong [39].

More recently, the techniques of viscosity solutions allow Evans and Sougandinis [18] to simplify the proofs of existence of a value and to characterize this value as a unique solution of some Hamilton–Jacobi equation. Pursuit games without state constraints and for continuous value functions are studied by Soravia in [36]. Alziary de Roquefort [1], [2] uses these methods for a particular continuous pursuit game with state constraints (the lion and man game).

In [33], Rozyev and Subbotin prove the existence of a value for a differential game without continuity and with some state constraints for one player (Victor). However, these results couldn't directly be extended to differential games with separate dynamics and with state constraints on both dynamics. Actually, the basic idea of this method—the so-called "extremal aiming" (which gives the strategy)—is not applicable to situations where there are state constraints for both players. Let us also point out that their approach is devoted to games in a context of strategies slightly different from the one used in this paper. This method is adapted to the kind of strategies we use, the nonanticipative strategies, and to the viscosity solution approach by Bardi, Bottacin, and Falcone in [6]. Our definition of value function is (partially) borrowed from this paper.

When this paper was complete, we received a preprint of Bardi, Koike, and Soravia [7] establishing the existence of a value for pursuit games with state constraints when this value is continuous. Let us point out that the constraints in [7] are more general than ours. However, the method developed in [7] heavily relies on the fact that the value function is continuous. For getting this continuity, the authors assume that one of the value functions vanishes at the boundary of the target and is continuous on this boundary, and that each control system (0.1) and (0.2) is locally controllable.

In this paper, we do not make any controllability condition at the boundary of the target, so that the value of the game is, in general, not continuous. However, following Soner [34], [35] we make some restriction on the dynamics at the boundary of the state constraints. Under this restriction, we prove that the pursuit-evasion game has a value. Let us point out that our result can be used in most examples given in the "classical theory" of differential games.

We follow here the method described in [14] for characterizing the value functions of differential game. Namely, we reduce the study of the pursuit game to a qualitative differential game called an "approach-evasion game." The idea of reducing a quantitative game to a qualitative one comes back to Isaacs [26]. Here we use an idea of Frankowska for control problems, which amounts to characterizing the epigraph of the value function (see [21], for instance, or [14] for further references on the subject). So, in a first step, we study this qualitative game (or game of kind, in Isaacs terminology), we give an "alternative result" for that game, and we characterize the victory domains in a geometric way. Then, in a second step, we interpret the pursuit game as a qualitative pursuit-evasion game in order to prove that there is a value to the problem.

1. Existence of a value for the pursuit game.

In this section, we state the main result of this paper, namely, that pursuit games have a value. For doing so, we first introduce some notations and assumptions.

1.1. Notations and assumptions. The dynamics of the system are

$$\begin{cases} y'(t) = g(y(t), u(t)),\ u(t) \in U,\ y(t) \in K_U, \\ z'(t) = h(z(t), v(t)),\ v(t) \in V,\ z(t) \in K_V, \end{cases}$$

with $y \in \mathbb{R}^l$, $z \in \mathbb{R}^m$. We set $x(t) := (y(t), z(t))$, $N := l + m$, and

$$f(x, u, v) := f(y, z, u, v) := \{g(y, u)\} \times \{h(z, v)\}.$$

The first player (Ursula), controlling u, has to ensure that $y(t) \in K_U$, while the second player (Victor), playing with v, has to ensure that $z(t) \in K_V$.

The sets K_U and K_V have smooth boundaries. Moreover, we assume some transversality conditions of the vector fields at the boundary of the state constraints similar to those of Soner [34], [35]. Namely,

$$(1.1) \quad \begin{cases} \text{(i)} & U \text{ and } V \text{ are compact subsets of some finite} \\ & \text{dimensional spaces,} \\ \text{(ii)} & f : \mathbb{R}^N \times U \times V \to \mathbb{R}^N \text{ is continuous and} \\ & \text{Lipschitz continuous (with Lipschitz constant } \ell) \\ & \text{with respect to } x, \\ \text{(iii)} & \bigcup_u f(x, u, v) \text{ and } \bigcup_v f(x, u, v) \text{ are convex for any } x, \\ \text{(iv)} & K_U = \{y \in \mathbb{R}^l,\ \phi_U(y) \le 0\} \text{ with } \phi_U \in \mathcal{C}^2(\mathbb{R}^l; \mathbb{R}), \\ & \nabla\phi_U(y) \ne 0 \text{ if } \phi_U(y) = 0, \\ \text{(v)} & K_V = \{z \in \mathbb{R}^m,\ \phi_V(z) \le 0\} \text{ with } \phi_V \in \mathcal{C}^2(\mathbb{R}^m; \mathbb{R}), \\ & \nabla\phi_V(z) \ne 0 \text{ if } \phi_V(z) = 0, \\ \text{(vi)} & \forall y \in \partial K_U,\ \exists u \in U \text{ with } \langle \nabla\phi_U(y), g(y, u)\rangle \ < \ 0, \\ \text{(vii)} & \forall z \in \partial K_V,\ \exists v \in V \text{ with } \langle \nabla\phi_V(z), h(z, v)\rangle \ < \ 0. \end{cases}$$

For any time-measurable controls $u(\cdot)$ and $v(\cdot)$, we denote by $y[y_0, u(\cdot)]$, $z[z_0, v(\cdot)]$, and $x[x_0, u(\cdot), v(\cdot)]$ the solutions (in the Caratheodory sense) to (0.1), (0.2), and to

$$(1.2) \quad x'(t) = f(x(t), u(t), v(t)),$$

starting, respectively, from y_0, z_0, and $x_0 := (y_0, z_0)$.

The sets of time-measurable controls $u(\cdot) : \mathbb{R}^+ \to U$ and $v(\cdot) : \mathbb{R}^+ \to V$ are denoted, respectively, by $\mathcal{U}$ and $\mathcal{V}$, while the sets of admissible controls are denoted by $\mathcal{U}(y_0)$ and $\mathcal{V}(z_0)$:

$$\begin{cases} \mathcal{U}(y_0) := \{u(\cdot) \in \mathcal{U} \mid y[y_0, u(\cdot)](t) \in K_U\ \forall t \ge 0\}, \\ \mathcal{V}(z_0) := \{v(\cdot) \in \mathcal{V} \mid z[z_0, v(\cdot)](t) \in K_V\ \forall t \ge 0\}. \end{cases}$$

Under condition (1.1), it is well known (see [4]) that the sets $\mathcal{U}(y_0)$ and $\mathcal{V}(z_0)$ are not empty for any $y_0 \in K_U$ and $z_0 \in K_V$. Moreover, according to Arisawa and Lions [3] (see also [28]), the sets $\mathcal{U}(y_0)$ and $\mathcal{V}(z_0)$ are Lipschitz continuous with respect to y_0 and z_0. Namely, (for instance, for y), we have the following lemma.

LEMMA 1.1. *Under assumption* (1.1)*, for any positive constants Q and T, there is some positive $\lambda = \lambda(Q, T, \ell)$ such that, for any y_0, y_1 belonging to K_U, with $\|y_0\| \le Q$ and $\|y_1\| \le Q$, and for any admissible control $u_0(\cdot) \in \mathcal{U}(y_0)$, there is some admissible control $u_1(\cdot) \in \mathcal{U}(y_1)$ such that*

$$\forall t \in [0, T],\ \|y[y_0, u_0(\cdot)](t) - y[y_1, u_1(\cdot)](t)\| \le \|y_0 - y_1\| e^{\lambda t}\ .$$

Remarks on assumptions (1.1).

1. The regularity assumptions on the domains K_U and K_V can also be weakened by using an extension of Lemma 1.1 recently obtained by Frankowska and Rampazzo in [22] for nonsmooth domains. The transversality condition has to be extended in a suitable way. The results of the present paper also hold true (with the exactly the same proof) under the Frankowska–Rampazzo assumption provided that the sets K_U and K_V are sleek in the sense of [4].
2. Following also [22], the convexity assumption (iii) can be avoided. We have preferred not to do so for simplicity. However, if one omits this assumption, one has to modify the transversality assumption in a suitable way as well as the definition of $\vartheta_C^\flat$ (see [22]).

The players play *nonanticipative strategies.* A map $\alpha : \mathcal{V}(z_0) \to \mathcal{U}(y_0)$ is a nonanticipative strategy (for the first player Ursula and for the point $x_0 := (y_0, z_0) \in K_U \times K_V$) if, for any $\tau > 0$ and for any control $v_1(\cdot)$ and $v_2(\cdot)$ belonging to $\mathcal{V}(z_0)$, which coincide almost everywhere (a.e.) on $[0, \tau]$, $\alpha(v_1(\cdot))$ and $\alpha(v_2(\cdot))$ coincide a.e. on $[0, \tau]$. We denote by $S_U(x_0)$ the set of such nonanticipative strategies for Ursula.

The nonanticipative strategies β for the second player Victor are defined symmetrically, and we denote by $S_V(x_0)$ the set of such strategies.

Throughout this paper, B denotes the closed unit ball of $\mathbb{R}^N$ (endowed with the Euclidean norm), and $d_S(x)$ denotes the distance from a point x to a set S. Moreover, if S is a subset of $\mathbb{R}^N$ and ε is positive, we denote by $S + \varepsilon B$ the set

$$S + \varepsilon B = \{x \in \mathbb{R}^N \mid d_S(x) \leq \varepsilon\} \, .$$

1.2. The main theorem. Let $C \subset K_U \times K_V$ be a closed target. The hitting-time of C for a trajectory $x(\cdot) := (y(\cdot), z(\cdot))$ is

$$\theta_C(x(\cdot)) := \min\{t \geq 0 \mid x(t) \in C\}.$$

If $x(t) \notin C$ for every $t \geq 0$, then we set $\theta_C(x(\cdot)) := +\infty$. In the pursuit game, Ursula wants to maximize θ_C, while Victor wants to minimize it.

Definition 1.2 (value functions). *The lower optimal hitting-time function is the map $\vartheta_C^\flat : K_U \times K_V \to \mathbb{R}^+ \cup \{+\infty\}$ defined, for any $x_0 := (y_0, z_0)$, by*

$$\vartheta_C^\flat(x_0) := \inf_{\beta(\cdot) \in S_V(x_0)} \sup_{u(\cdot) \in \mathcal{U}(y_0)} \theta_C\left(x[x_0, u(\cdot), \beta(u(\cdot))]\right).$$

The upper optimal hitting-time function is the map $\vartheta_C^\sharp : K_U \times K_V \to \mathbb{R}^+ \cup \{+\infty\}$ defined, for any $x_0 := (y_0, z_0)$, by

$$\vartheta_C^\sharp(x_0) := \lim_{\varepsilon \to 0^+} \sup_{\alpha(\cdot) \in S_U(x_0)} \inf_{v(\cdot) \in \mathcal{V}(z_0)} \theta_{C + \varepsilon B}\left(x[x_0, \alpha(v(\cdot)), v(\cdot)]\right).$$

By convention, we set $\vartheta_C^\flat(x) = \vartheta_C^\sharp(x) = 0$ on C.

Remarks.

1. Let us point out that the limit in the definition of $\vartheta_C^\sharp$ exists because the quantity

$$\sup_{\alpha(\cdot) \in S_U(x_0)} \inf_{v(\cdot) \in \mathcal{V}(z_0)} \theta_{C + \varepsilon B}\left(x[x_0, \alpha(v(\cdot)), v(\cdot)]\right)$$

is nondecreasing with respect to ε. Such a definition is used, for instance, in [6]. The meaning of such a definition is the following. Whatever strategy α

is played, the second player can ensure the state of the system to go as close as he wants to the target before $\vartheta^\sharp_C(x_0)$ (but the state of the system need not reach the target).

2. The following definition of the upper value function has been used in several papers:

$$\widehat{\vartheta}^\sharp_C(x_0) := \sup_{\alpha(\cdot)\in S_U(x_0)} \inf_{v(\cdot)\in\mathcal{V}(z_0)} \theta_C\left(x[x_0,\alpha(v(\cdot)),v(\cdot)]\right).$$

However, without controllability assumptions on the boundary of the target, we cannot hope to have a value with this definition of upper value function. For instance, if one considers the unconstrained pursuit game where the dynamics is

$$\begin{cases} y'(t) = u(t), \text{ where } u(t) \in U = [-1,1], \\ z_1'(t) = v(t), \text{ where } v(t) \in V = [-1,1], \\ z_2'(t) = 1, \end{cases}$$

and the target is $C = \{(y,z_1,z_2) \in \mathbb{R}^3 \mid y = z_1 \text{ and } z_2 = 1\}$, then

$$\widehat{\vartheta}^\sharp_C(0,0,0) = +\infty,$$

while

$$\vartheta^\flat_C(0,0,0) = \vartheta^\sharp_C(0,0,0) = 1.$$

Proof. We first prove the last equality. Let us notice that $\beta(u(\cdot))(t) = u(t)$ is an optimal strategy for the second player. Hence $\vartheta^\flat_C(0,0,0) = 1$. Equality $\vartheta^\flat_C(0,0,0) = \vartheta^\sharp_C(0,0,0)$ comes from Theorem 1.3 below.

We now prove the first equality. We define the nonanticipative strategy α in the following way. For any control $v(\cdot)$, let $z_1(\cdot)$ be the solution to

$$\begin{cases} z_1'(t) = v(t), \\ z_1(0) = 0. \end{cases}$$

We set $c = \liminf_{h\to 0^+} z_1(h)/h$. Since $v(t) \in [-1,1]$, we have $c \in [-1,1]$. If $c > -1$, we set $\alpha(v(\cdot))(t) = -1$, while, if $c = -1$, we set $\alpha(v(\cdot))(t) = 1$. Let us point out that such a map α is a nonanticipative strategy. We claim that, for any control $v(\cdot)$, the solution $x(\cdot) = x[0,\alpha(v(\cdot)),v(\cdot)]$ never touches the target.

We consider two cases. If, on the one hand, $c > -1$, then there is some $\tau \in (0,1)$ such that

$$\forall t \in (0,\tau],\ z_1(t) \geq \frac{(c-1)}{2}t \ > \ -t = y(t)\ .$$

Hence

$$\forall t \geq \tau,\ z_1(t) \geq \frac{(c-1)}{2}\tau - (t-\tau) \ > \ -t = y(t)\ .$$

Therefore, for any t, $z_1(t) > y(t)$, so that $x(t) \notin C$.

If, on another hand, $c = -1$, there is a sequence $t_n \to 0^+$ such that $\lim_n z_1(t_n)/t_n = -1$. Hence there is some n such that $t_n \in (0, 1)$ and $z_1(t_n) \leq t_n/2$. Then

$$\forall t \geq t_n,\ z_1(t) \leq t_n/2 + (t - t_n) \ < \ t = y(t)\ .$$

Therefore, for any $t \geq t_n$, $z_1(t) < y(t)$, so that $x(t) \notin C$.

3. A natural question is "What happens if one modifies the lower value function in the same way?" We show below that

$$\vartheta^\flat_C(x_0) := \lim_{\varepsilon \to 0^+} \inf_{\beta(\cdot) \in S_V(x_0)} \sup_{u(\cdot) \in \mathcal{U}(y_0)} \theta_{C+\epsilon B}\left(x[x_0, u(\cdot), \beta(u(\cdot))]\right)$$

(see Proposition 3.2).

4. A last remark, which is not really interesting for differential games, might be interesting from a PDE point of view. In the theory of viscosity solutions, one often characterizes the solution through its lower semicontinuous and upper semicontinuous envelope. In the above example we can notice that the lower semicontinuous envelope of $\widehat{\vartheta^\sharp_C}$ *is not equal* to $\vartheta^\sharp_C$.

THEOREM 1.3. *Assume that conditions* (1.1) *are fulfilled. Then the game has a value:*

$$\forall x_0 \in K_U \times K_V,\ \vartheta^\flat_C(x_0) = \vartheta^\sharp_C(x_0).$$

This theorem is proved in section 4 by reducing the pursuit game to a qualitative differential game called the pursuit-evasion game. We deduce from the "alternative theorem" for this qualitative game (see Theorem 2.6, below) the existence of a value for the pursuit game.

Moreover, the proof gives a geometric characterization of the value function. This characterization can be formulated as a Hamilton–Jacobi–Isaacs equation (see [14]). We shall not do so for the sake of brevity. As indicated in [14], we can also derive from this characterization numerical schemes for computing the value function.

2. An alternative theorem for a qualitative differential game with state constraints.

In this section, we study the differential game in which the first player, Ursula, controlling system (0.1), aims the state of the full system at reaching an open target $\mathcal{O}$ while the other player, Victor, controlling system (0.2), aims the state of the system at avoiding $\mathcal{O}$ and—if possible—at reaching some given evasion set $\mathcal{E}$. This game is very close to the approach-evasion game of Krasovskii and Subbotin [27].

2.1. Statement of the qualitative problem.

DEFINITION 2.1. *Let $\mathcal{O} \subset K_U \times K_V$ be an open target, and let $\mathcal{E} \subset K_U \times K_V$ be a closed evasion set. The victory domains of the players are defined as follows.*

- *Victor's victory domain is the set of initial positions $x_0 := (y_0, z_0)$ belonging to $K_U \times K_V$ for which there is an admissible nonanticipative strategy $\beta : \mathcal{U}(y_0) \to \mathcal{V}(z_0)$ such that, for any admissible control $u(\cdot) \in \mathcal{U}(y_0)$, the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ avoids $\mathcal{O}$ as long as it does not reach $\mathcal{E}$ (or avoids $\mathcal{O}$ on $[0, +\infty)$ if it never reaches $\mathcal{E}$).*
- *Ursula's victory domain is the set of initial positions $x_0 := (y_0, z_0)$ belonging to $K_U \times K_V$ for which there are $T \geq 0$, $\varepsilon > 0$, and an admissible nonanticipative strategy $\alpha : \mathcal{V}(z_0) \to \mathcal{U}(y_0)$ such that, for any admissible control $v(\cdot) \in \mathcal{V}(z_0)$, the solution $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ reaches $\mathcal{O}_\varepsilon := \{x \in \mathcal{O} \mid d_{\partial\mathcal{O}}(x) > \varepsilon\}$ at some time $\tau \leq T$ and does not reach $\mathcal{E} + \varepsilon B$ on $[0, \tau]$.*

In this section, we prove the following alternative result: if x belongs to $K_U \times K_V$, then x belongs to one and only one victory domain. Moreover, we give a geometric characterization of the victory domains.

2.2. The discriminating domains. For $x := (y, z) \in \mathbb{R}^N$, we set

$$U(y) := \{u \in U \mid g(y, u) \in T_{K_U}(y)\},$$

where $T_{K_U}(y)$ is the usual tangent half-space to the set with smooth boundary K_U at y. Let us notice that, under assumptions (1.1), the set-valued map $y \to f(y, U(y))$ is lower semicontinuous with convex compact values (see [5]).

Let us introduce the Hamiltonian of our system:

$$(2.1) \qquad H(x,p) := \begin{cases} \inf_{v\in V} \sup_{u\in U(y)} \langle f(x,u,v), p\rangle & \text{if } x \notin \mathcal{E}, \\ \min\{0\ ,\ \inf_{v\in V} \sup_{u\in U(y)} \langle f(x,u,v), p\rangle\} & \text{otherwise,} \end{cases}$$

where $x := (y, z)$ and where

$$\inf_{v\in V} \sup_{u\in U(y)} \langle f(x,u,v), p\rangle = \sup_{u\in U(y)} \langle g(y,u), p_y\rangle + \inf_{v\in V} \langle h(z,v), p_z\rangle.$$

DEFINITION 2.2. *A closed subset D of $K_U \times K_V$ is a discriminating domain for H if and only if*

$$\forall x \in D,\ \forall p \in NP_D(x),\ H(x,p) \le 0,$$

where $NP_D(x)$ denotes the set of proximal normal to D at x, i.e., the set of $p \in \mathbb{R}^N$ such that the distance of $x + p$ to D is equal to $\|p\|$.

For the original definition of discriminating domains, see Aubin [4].

Discriminating domains can be characterized in two different ways.

THEOREM 2.3. *Suppose that assumptions* (1.1) *are fulfilled. A closed subset D of $K_U \times K_V$ is a discriminating domain for H if and only if, for any initial position $x_0 = (y_0, z_0) \in D$, there is a nonanticipative strategy $\beta \in S_V(x_0)$ such that, for any $u(\cdot) \in \mathcal{U}(y_0)$, the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ remains in D until it reaches $\mathcal{E}$ (or remains in D on $[0, +\infty)$ if it never reaches $\mathcal{E}$).*

Remark. This result was proved independently and in the same time by Plaskacz [30] and by the first author [11] when $K_U = \mathbb{R}^l$ and $\mathcal{E} = \emptyset$. For time-measurable dynamics, see also [15]. Theorem 2.3 (in a more general form) was announced in [13].

THEOREM 2.4. *Suppose that assumptions* (1.1) *are fulfilled. A closed set $D \subset K_U \times K_V$ is a discriminating domain for H if and only if, for any initial position $x_0 := (y_0, z_0) \in D$, for any admissible nonanticipative strategy $\alpha : \mathcal{V}(z_0) \to \mathcal{U}(y_0)$, for any $T \ge 0$, and for any $\varepsilon > 0$, there is an admissible control $v(\cdot) \in \mathcal{V}(z_0)$ such that the solution $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ remains in $D + \varepsilon B$ on $[0, T]$ as long as it does not reach $\mathcal{E} + \varepsilon B$. Namely,*

- *either there is some $\tau \le T$ such that $x[x_0, \alpha(v(\cdot)), v(\cdot)](\tau)$ belongs to $\mathcal{E} + \varepsilon B$ and $x[x_0, \alpha(v(\cdot)), v(\cdot)](t) \in D + \varepsilon B$ for $t \in [0, \tau]$,*
- *or $x[x_0, \alpha(v(\cdot)), v(\cdot)](t) \in D + \varepsilon B$ for $t \in [0, T]$.*

Remark. Note that, if $D \subset (K_U \times K_V)\backslash\mathcal{O}$ is a discriminating domain, then D is a subset of Victor's victory domain (according to Theorem 2.3) and has an empty intersection with Ursula's victory domain (according to Theorem 2.4).

The proof of Theorems 2.3 and 2.4, being rather technical, are given in the appendix.

2.3. The alternative theorem. We now characterize Victor's and Ursula's victory domains. For that purpose, we first recall the definition of the discriminating kernel.

PROPOSITION 2.5 (see [12]). *Let $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be a lower semicontinuous map. If K is a subset of $\mathbb{R}^N$, then K contains a largest (for the inclusion)* closed *discriminating domain for H. This set is called the discriminating kernel of K for H and is denoted by $Disc_H(K)$.*

THEOREM 2.6 (alternative theorem). *Let H be defined by* (2.1) *and assume that* (1.1) *is fulfilled. Then,*

- *Victor's victory domain is equal to $Disc_H(\mathcal{K})$, and*
- *Ursula's victory domain is equal to $(K_U \times K_V)\backslash Disc_H(\mathcal{K})$,*

where $\mathcal{K} := (K_U \times K_V)\backslash\mathcal{O}$.

In particular, any point of $(K_U \times K_V)$ belongs either to Victor's victory domain or to Ursula's. When $K_U = K_V = \mathbb{R}^N$, this characterization can be found in [11].

The proof is given in the appendix.

3. Proof of the existence of a value for the pursuit game. We come back to the problem of the existence of a value (see section 1), i.e., to the equality between $\vartheta_C^\flat$ and $\vartheta_C^\sharp$. We are going to prove that their epigraphs are equal. Let us recall that the epigraph of $\vartheta_C^\flat$ (for instance) is a subset of $\mathbb{R}^{N+1}$ defined by

$$\mathcal{E}pi(\vartheta_C^\flat) = \{(x,\rho) \in \mathbb{R}^{N+1} \mid \vartheta_C^\flat(x) \le \rho\}.$$

In what follows, we always denote by (x,ρ) any point of $\mathbb{R}^{N+1}$, where $x \in \mathbb{R}^N$ and $\rho \in \mathbb{R}$.

THEOREM 3.1. *Assume that conditions* (1.1) *are fulfilled. Then we have*

$$\mathcal{E}pi(\vartheta_C^\flat) = Disc_H(\mathcal{K}) = \mathcal{E}pi(\vartheta_C^\sharp), \tag{3.1}$$

where $\mathcal{K} := K_U \times K_V \times \mathbb{R}^+$ and where the Hamiltonian $H : \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} \to \mathbb{R}$ is defined by

$$\begin{aligned}
&\forall (x,\rho) \in \mathbb{R}^{N+1},\ (p_x, p_\rho) \in \mathbb{R}^{N+1},\\
&H(x,\rho,p_x,p_\rho)\\
&\quad := \sup_{u \in U(y)} \inf_{v \in V} \langle f(x,u,v), p_x\rangle - p_\rho \qquad\qquad \textit{if } x \notin C,\\
&\quad := \min\left\{ 0\,;\ \sup_{u \in U(y)} \inf_{v \in V} \langle f(x,u,v), p_x\rangle - p_\rho \right\} \quad \textit{otherwise.}
\end{aligned}$$

Remarks.

- This result proves Theorem 1.3 since the functions $\vartheta_C^\flat$ and $\vartheta_C^\sharp$, having the same hypograph, are equal.
- We can deduce from this result that the map $\vartheta_C^\flat = \vartheta_C^\sharp$ is lower semicontinuous since its epigraph is closed (the set $Disc_H(\mathcal{K})$ being closed from Proposition 2.5).
- We can also derive from the proof of Theorem 3.1 the existence of an optimal strategy for the pursuer (Victor).

Proof of Theorem 3.1. *Proof of the first equality of* (3.1).

Let us introduce the dynamic $\widetilde{f} : \mathbb{R}^N \times \mathbb{R} \times U \times V \to \mathbb{R}^N \times \mathbb{R}$ given by

$$\widetilde{f}(x,\rho,u,v) := \{f(x,u,v)\} \times \{-1\}.$$

Note that the Hamiltonian H defined above is actually of the form of the Hamiltonian H defined by (2.1) for the dynamics $\widetilde{f}$ and for the closed evasion set defined by $\mathcal{E} := C \times \mathbb{R}$.

From Theorem 2.6, the discriminating kernel of $\mathcal{K}$ for $\widetilde{H}$ is the set of initial positions $(x_0, \rho_0) \in \mathcal{K}$ for which there is some nonanticipative strategy $\beta \in S_V(x_0)$ for Victor such that, for any $u(\cdot) \in \mathcal{U}(y_0)$, the solution to

$$\text{(3.2)} \qquad \begin{cases} x'(t) = f(x(t), u(t), \beta(u(\cdot))(t)), \\ \rho'(t) = -1, \\ x(0) = x_0, \qquad \text{and} \qquad \rho(0) = \rho_0 \end{cases}$$

remains in $\mathcal{K}$ until it reaches the set $\mathcal{E}$.

Let (x_0, ρ_0) belong to $Disc_{\widetilde{H}}(\mathcal{K})$, let β be an associated nonanticipative strategy, and let $(x(\cdot), \rho(\cdot))$ be the solution to (3.2). Since the solution $(x(\cdot), \rho(\cdot))$ remains in $\mathcal{K}$ as long as it does not reach $\mathcal{E}$, $\rho(t) = \rho_0 - t \geq 0$ as long as $x(t) \notin C$. Thus the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ reaches C before ρ_0. In particular,

$$\text{(3.3)} \qquad \vartheta_C^\flat(x_0) \leq \sup_{u(\cdot) \in \mathcal{U}} \theta_C(x[x_0, u(\cdot), \beta(u(\cdot))]) \leq \rho_0 \ .$$

So (x_0, ρ_0) belongs to $\mathcal{E}pi(\vartheta_C^\flat(\cdot))$ and $Disc_{\widetilde{H}}(\mathcal{K}) \subset \mathcal{E}pi(\vartheta_C^\flat(\cdot))$.

For proving the converse inclusion, let x_0 belong to the domain of $\vartheta_C^\flat(\cdot)$, and let $\rho_0 > \vartheta_C^\flat(x_0)$. There is some nonanticipative strategy β for Victor such that, for any $u(\cdot) \in \mathcal{U}(y_0)$, the solution $x[x_0, u(\cdot), \beta(u(\cdot))]$ reaches C before ρ_0. In particular, the solution $(x(\cdot), \rho(\cdot))$ to (3.2) remains in $\mathcal{K}$ until it reaches $\mathcal{E}$, so that (x_0, ρ_0) belongs to $Disc_{\widetilde{H}}(\mathcal{K})$ from Theorem 2.6. This holds true for any $\rho_0 > \vartheta_C^\flat(x_0)$.

Since the discriminating kernel is a closed set, we have proved that

$$\mathcal{E}pi(\vartheta_C^\flat(\cdot)) \subset Disc_{\widetilde{H}}(\mathcal{K}).$$

Proof of the second equality of (3.1). Let (y_0, z_0, w_0) belong to $Disc_{\widetilde{H}}(\mathcal{K})$. Fix $\varepsilon > 0$, and let $\alpha : \mathcal{V}(z_0) \to \mathcal{U}(y_0)$ be such that

$$\forall v(\cdot) \in \mathcal{V}(z_0), \ \theta_{C+\varepsilon B}(x[x_0, \alpha(v(\cdot)), v(\cdot)]) \geq \vartheta_C^\sharp(y_0, z_0) - \varepsilon \ .$$

From Theorem 2.6, for this nonanticipative strategy α, for this $\varepsilon > 0$, and for $T := w_0 + \varepsilon$, there is a control $v(\cdot) \in \mathcal{V}(z_0)$ such that the solution to

$$\text{(3.4)} \qquad \begin{cases} x'(t) = f(x(t), \alpha(v(\cdot))(t), v(t)), \\ w'(t) = -1, \\ y(0) = y_0, \ z(0) = z_0, \ w(0) = w_0 \end{cases}$$

remains in $\mathcal{K} + \varepsilon B$ on $[0, T]$ as long as it does not reach $(C \times \mathbb{R}) + \varepsilon B$.

Set $\tau := \inf\{\theta_{C+\varepsilon B}((y(\cdot), z(\cdot))); T\}$. Then $w(t) = w_0 - t$ on $[0, \tau]$ and $w(t) \geq -\varepsilon$ for any $t \in [0, \tau]$. So $\tau \leq w_0 + \varepsilon$.

So finally, $w_0 \geq \vartheta_C^\sharp(y_0, z_0) - 2\varepsilon$. Since this holds true for any $\varepsilon > 0$, we have finally proved that $w_0 \geq \vartheta_C^\sharp(y_0, z_0)$. Thus $Disc_{\widetilde{H}}(\mathcal{K})$ is a subset of $\mathcal{E}pi(\vartheta_C^\sharp)$.

Conversely, let $w_0 > \vartheta_C^\sharp(y_0, z_0)$. Then, for any nonanticipative strategy $\alpha : \mathcal{V}(z_0) \to \mathcal{U}(y_0)$, for any positive ε, there is a control $v(\cdot) \in \mathcal{V}(z_0)$ such that

$$\theta_{C+\varepsilon B}(x[x_0, \alpha(v(\cdot)), v(\cdot)]) \leq w_0.$$

Note that the function $t \to (x[x_0, \alpha(v(\cdot)), v(\cdot))](t), w_0 - t)$ is a solution to (3.4). Moreover, this solution remains in $\mathcal{K}$ as long as the solution does not reach $\mathcal{E} + \varepsilon B$ (i.e., on $[0, \theta_{C+\varepsilon B}(x[x_0, \alpha(v(\cdot)), v(\cdot)])])$. So (y_0, z_0, w_0) belongs to $Disc_{\widetilde{H}}(\mathcal{K})$. Since $Disc_{\widetilde{H}}(\mathcal{K})$ is closed, this holds true for any $w_0 \geq \vartheta^\sharp_C(y_0, z_0)$.

This proves the equality between $Disc_{\widetilde{H}}(\mathcal{K})$ and $\mathcal{E}pi(\vartheta^\sharp_C)$.

We can also use the previous theorem in order to prove some stability results. Namely, we have the following proposition.

Proposition 3.2. *Under the assumptions of Theorem* 3.1*,*

$$\vartheta^\flat_C(x_0) := \lim_{\varepsilon \to 0} \inf_{\beta(\cdot) \in S_V(x_0)} \sup_{u(\cdot) \in \mathcal{U}(y_0)} \theta_{C+\epsilon B}\left(x[x_0, u(\cdot), \beta(u(\cdot))]\right).$$

Proof. Let us denote by $\vartheta^\flat_{C+\epsilon B}$ and H_ϵ the value and the Hamiltonian associated with the target $C + \epsilon B$:

$$\begin{array}{l} \forall (x,\rho) \in \mathbb{R}^{N+1},\ (p_x, p_\rho) \in \mathbb{R}^{N+1}, \\ H(x, \rho, p_x, p_\rho) \\ \quad := \displaystyle\sup_{u \in U(y)} \inf_{v \in V} \langle f(x,u,v), p_x\rangle - p_\rho \qquad\qquad\qquad \text{if } x \notin C + \epsilon B, \\ \quad := \min\left\{ 0\ ;\ \displaystyle\sup_{u \in U(y)} \inf_{v \in V} \langle f(x,u,v), p_x\rangle - p_\rho \right\} \quad \text{otherwise.} \end{array}$$

According to Theorem 3.1, we have

$$\mathcal{E}pi(\vartheta^\flat_{C+\epsilon B}) = Disc_{H_\epsilon}(\mathcal{K}),$$

where $\mathcal{K} := K_U \times K_V \times \mathbb{R}^+$. Since $H_\epsilon \leq H$, the following inequality is obvious:

$$Disc_H(\mathcal{K}) \subset Disc_{H_\epsilon}(\mathcal{K})\ .$$

Conversely, since the lower semicontinuous Hamiltonians H_ϵ converge in the sense of Proposition 1.2 of [12], this proposition states that the decreasing limit of $Disc_{H_\epsilon}(\mathcal{K})$ is a discriminating domain for H. Since this limit is contained in $\mathcal{K}$, because so are $Disc_{H_\epsilon}(\mathcal{K})$, it is contained in $Disc_H(\mathcal{K})$. So we have proved that

$$\bigcap_{\epsilon > 0} Disc_{H_\epsilon}(\mathcal{K}) = Disc_H(\mathcal{K})\ ,$$

which is equivalent to saying that

$$\vartheta^\flat_C(x_0) := \lim_{\epsilon \to 0^+} \inf_{\beta(\cdot) \in S_V(x_0)} \sup_{u(\cdot) \in \mathcal{U}(y_0)} \theta_{C+\epsilon B}\left(x[x_0, u(\cdot), \beta(u(\cdot))]\right)\ .$$

4. Appendix.

We now prove Theorems 2.3, 2.4, and 2.6. The proof of these results has the same framework as the proof of Theorems 2.1, 2.2, 2.3, and 2.4 of [11]. However, the key points of the proofs essentially differ because of the presence of the constraints. Hence, for the sake of brevity, we refer to [11] for the framework of the proofs, and we only give the key points.

4.1. Proof of Theorem 2.3. The condition is sufficient.

Assume that D is a discriminating domain. The crucial point of the proof is the following lemma.

Lemma 4.1. *Suppose that the assumptions of Theorem* 2.3 *are fulfilled, and assume that D is a discriminating domain for H. Then for any $x_0 = (y_0, z_0) \in D$*

and for any control $u(\cdot) \in \mathcal{U}(y_0)$, there is a control $v(\cdot) \in \mathcal{V}(z_0)$ such that the solution $x(\cdot) := x[x_0, u(\cdot), v(\cdot)]$ remains in D until it reaches $\mathcal{E}$ (or remains in D forever if it does not reach $\mathcal{E}$).

The sequel of the proof runs as in Theorem 2.1 of [11] or as in [15].

Proof of Lemma 4.1. Let us assume that D is a discriminating domain for H. Let $x_0 \in D$ and $u_0(\cdot) \in \mathcal{U}(y_0)$. It is enough to prove that there is a measurable control $v(\cdot) \in \mathcal{V}$ and a time $T > 0$ such that $x(\cdot) := x[x_0, u_0(\cdot), v(\cdot)]$ remains in D on $[0,T]$ until it reaches $\mathcal{E}$. Let us notice that, if $x_0 \in \mathcal{E}$, then the proof is obvious. We now assume that $x_0 \notin \mathcal{E}$.

We divide the proof in two steps. In the first step we assume that y_0 belongs to the interior of K_U, and in the second step that y_0 belongs to ∂K_U.

First step. We assume that there is some $T > 0$ such that the solution $y(\cdot) := y[y_0, u_0(\cdot)]$ remains in $\text{Int}(K_U)$ on $[0,T]$. Then we are going to prove that there is a control $v(\cdot) \in \mathcal{V}(z_0)$ such that $x[x_0, u(\cdot), v(\cdot)]$ remains in D on $[0,T]$.

Let $\varepsilon > 0$ be such that

$$\forall t \in [0,T], \quad y(t) + \varepsilon B \subset \text{Int}(K_U).$$

Let us introduce the following open set:

$$W := \{(y,z) \in \mathbb{R}^N \mid \exists t \in [0,T],\ \|y(t) - y\| < \varepsilon\} = (y([0,T]) + \varepsilon \overset{o}{B}) \times \mathbb{R}^m.$$

Let us define the set-valued map:

$$F(t,x) := \begin{cases} \bigcup_{v \in V} f(x, u_0(t), v) & \text{if } x \notin \mathcal{E}, \\ \overline{Co}\{\{0\} \cup \bigcup_{v \in V} f(x, u_0(t), v)\} & \text{if } x \in \mathcal{E}. \end{cases}$$

The main point of the proof is to check the assumptions for applying the measurable viability theorem of [21] for the set $W \cap D$ and F.

Let us notice that F is measurable and upper semicontinuous with respect to x. Moreover, F has convex compact values. We claim that the set $W \cap D$ is a locally compact viability domain for F. Indeed, if $x := (y,z) \in W \cap D$, then $y \in \text{Int}(K_U)$, so that $U(y) = U$. Since D is a discriminating domain, $U(y) = U$ on W, and $f(x,u,V)$ is convex, the proximal normal condition can be replaced by the following one (see, for instance, [12]):

$$\forall x \in D \cap W,\ \forall u \in U,\ \exists v \in V \text{ with } f(x,u,v) \in T_D(x).$$

Thus

$$F(t,x) \cap T_D(x) \neq \emptyset \qquad \text{a.e. } t \in [0,T],$$

where $T_D(x) = \{v \in \mathbb{R}^N\ ,\ \liminf_{h \to 0^+} d_D(x + hv)/h = 0\}$. So we have proved that $W \cap D$ is a locally compact viability domain for F on $[0,T]$.

The measurable viability theorem of [21] states that there is a solution $\bar{x}(\cdot)$ to the differential inclusion

$$\begin{cases} \bar{x}'(t) \in F(t, \bar{x}(t)), \\ \bar{x}(0) = x_0, \end{cases}$$

which remains in D as long as it belongs to W. Note that $\bar{x}(t) = (\bar{y}(t), \bar{z}(t))$, where $\bar{y}(\cdot) = y[y_0, u_0(\cdot)] = y(\cdot)$.

In particular, $\bar{x}(t)$ remains in W on $[0,T]$ and thus also in D on $[0,T]$. As long as $\bar{x}(\cdot)$ does not belong to $\mathcal{E}$, $\bar{x}(\cdot)$ is the solution to

$$\bar{x}'(t) \in \bigcup_{v} f(\bar{x}(t), u_0(t), v)\ .$$

Then the measurable selection theorem (Theorem 8.3.1 of [5]) states that there is some control $v(\cdot)$ such that the solution $x[x_0, u_0(\cdot), v(\cdot)]$ is equal to $\bar{x}(\cdot)$ until the solution reaches $\mathcal{E}$. When the solution has reached $\mathcal{E}$, we can set $v(t) = \bar{v}$, where $\bar{v}$ is any element of V. So we have finally defined a control $v(\cdot)$ such that the solution $x[x_0, u_0(\cdot), v(\cdot)]$ remains in D on $[0,T]$ until it reaches $\mathcal{E}$.

Second step. Let us now assume that y_0 belongs to ∂K_U. From assumption (1.1), there is some $\bar{u} \in U(y_0)$ such that $w := g(y_0, \bar{u})$ belongs to $\mathrm{Int}(T_{K_U}(y_0))$.

Fix $\theta > 0$, and define $x_\theta = (y_0 + \theta w, z_0)$ and $y_\theta(\cdot) := y[y_0 + \theta w, u_0(\cdot)]$. Since K_U is smooth, the solution $y_\theta(\cdot)$ remains in $\mathrm{Int}(K_U)$ on some interval $[0,T]$ (with $T > 0$ independent of θ) for any θ sufficiently small.

Thanks to the first step, we know that there is some control $v_\theta(\cdot)$ such that the solution $x_\theta(\cdot) := x[x_\theta, u_0(\cdot), v_\theta(\cdot)]$ remains in D on $[0,T]$ until it reaches $\mathcal{E}$. Since the set-valued map $\bigcup_{v\in V} f(x, u_0(t), v)$ is measurable, upper semicontinuous with respect to x, and has convex compact values, and since $\mathcal{E}$ is closed, there are a sequence $\theta_n \to 0^+$ and a sequence $x_{\theta_n}(\cdot)$ which converge to some $x(\cdot)$ starting from x_0 and remaining in D on $[0,T]$ until it reaches $\mathcal{E}$. Since, as long as the $x(\cdot)$ has not reached $\mathcal{E}$, $x(\cdot)$ is a solution to

$$x'(t) \in \bigcup_{v} f(x(t), u_0(t), v),$$

the measurable selection theorem states that there exists a control $v(\cdot) \in \mathcal{V}(z_0)$ such that $x(\cdot) = x[x_0, u_0(\cdot), v(\cdot)]$ as long as this solution has not reached $\mathcal{E}$. After the solution has reached $\mathcal{E}$, we can set $v(t) = \bar{v}$, where $\bar{v}$ is any element of V.

The condition is necessary. Assume that the closed set D satisfies the property given in Theorem 2.3. Let $\bar{x} = (\bar{y}, \bar{z}) \in D\backslash\mathcal{E}$, $p \in NP_D(\bar{x})$, and $\bar{u} \in U(\bar{y})$. We have to prove that

$$\sup_{u\in U(\bar{y})} \inf_{v\in V} \langle f(\bar{x}, u, v), p\rangle \le 0.$$

Since the set-valued map $y \to g(y, U(y))$ is lower semicontinuous with compact convex values, the Michael selection theorem (see [5]) yields the existence of a continuous selection $\widetilde{w} : K_U \to U$ of this set-valued map, i.e., $\widetilde{w}(y) \in g(y, U(y))$ such that $\widetilde{w}(\bar{y}) = g(\bar{y}, \bar{u})$. Let $y(\cdot)$ be any solution to the differential equation

$$\begin{cases} y'(t) = \widetilde{w}(y(t)),\ y(t) \in K_U, \\ y(0) = \bar{y}. \end{cases}$$

Then from the measurable selection theorem, there is some measurable control $u(\cdot) \in \mathcal{U}(\bar{y})$ with $y(t) = y[\bar{y}, u(\cdot)](t)$. Note that $y'(0) = g(\bar{y}, \bar{u})$.

Let β be a nonanticipative strategy as in Theorem 2.3. Then $x(\cdot) := x[\bar{x}, u(\cdot), \beta(u(\cdot))]$ remains in D and thus in $\overline{\mathbb{R}^N\backslash(\bar{x} + p + \|p\|B)}$. Then from standard arguments, there is a sequence $t_k \to 0^+$ such that $(x(t_k) - \bar{x})/t_k$ converges to some $v = (v_y, v_z)$ with $\langle v, p\rangle \le 0$. From the very construction of $u(\cdot)$, $v_y = g(\bar{y}, \bar{u})$. Since $h(z, V)$ is convex, $v_z \in h(\bar{z}, V)$. Thus

$$\sup_{u\in U(\bar{y})} \inf_{v\in V} \langle f(\bar{x}, u, v), p\rangle \le 0.$$

4.2. Proof of Theorem 2.4. The condition is sufficient. The proof of the sufficient condition follows the proof of Theorem 2.3 in [11] with Lemma 4.2 below instead of Lemma 4.4 of [11]. In Lemma 4.4 of [11], we use a kind of "extremal aiming" method. Extremal aiming [27] amounts to associating with any point $x \notin D$ some projection $\bar{x}$ of x onto D, and to playing for Ursula some $u \in U$ such that $\inf_v \langle f(\bar{x}, u, v), x - \bar{x}\rangle$ is maximum, and for Victor some $v \in V$ such that $\sup_u \langle f(\bar{x}, u, v), x - \bar{x}\rangle$ is minimum. Unfortunately, this method fails here since the players have to play admissible strategies, and the strategies given by the extremal aiming method have no reason to be admissible.

For stating Lemma 4.2, let us fix some notations. Since we are working only on the bounded interval $[0, T]$ and with solutions starting from initial position x_0, we denote by Q a radius such that any solution starting from x_0 remains in QB on $[0, T]$. We denote by M a upper bound of $\|f\|$ on QB.

LEMMA 4.2. *Under the assumptions of Theorem* 2.4*, there are positive constants (depending on Q, T, M, and ℓ) a and b such that, for any $x \in (K_U \times K_V) \cap QB$, with $x \notin D$ but $d_D(x) \leq Q$, for any admissible nonanticipative strategy $\alpha(\cdot) \in S_U(x)$, and for any $\tau \in [0, T]$, there is some admissible control $v(\cdot) \in \mathcal{V}(z)$ such that*

$$d_D^2(x[x, \alpha(v(\cdot)), v(\cdot)](\tau)) \leq d_D^2(x)[1 + a\tau] + b\tau^2.$$

Proof of Lemma 4.2. Let $\bar{x} := (\bar{y}, \bar{z})$ belong to the projection of x onto D. We also set $p = (p_y, p_z) = x - \bar{x}$. Recall that p belongs to $NP_D(\bar{x})$ and $\|p\| \leq Q$. Let us fix $\tau \in (0, T]$.

Let us consider $\widetilde{u}(\cdot) \in \mathcal{U}(y)$ an admissible control such that

$$\langle y[y, \widetilde{u}(\cdot)](\tau), p_y\rangle = \max_{u(\cdot) \in \mathcal{U}(y)} \langle y[y, u(\cdot)](\tau), p_y\rangle.$$

Thanks to Lemma 1.1, there is some control $\bar{u}(\cdot) \in \mathcal{U}(\bar{y})$ such that

$$\|y[\bar{y}, \bar{u}(\cdot)](\tau) - y[y, \widetilde{u}(\cdot)](\tau)\|^2 \leq \|p_y\|^2 e^{2\lambda\tau} .$$

From Lemma 4.1, there is some admissible control $\bar{v}(\cdot) \in \mathcal{V}(\bar{z})$ such that the solution $x[\bar{x}, \bar{u}(\cdot), \bar{v}(\cdot)]$ remains in D. Thanks to Lemma 1.1, there is some control $v(\cdot) \in \mathcal{V}(z)$ such that

$$\|z[\bar{z}, \bar{v}(\cdot)](\tau) - z[z, v(\cdot)](\tau)\|^2 \leq \|p_z\|^2 e^{2\lambda\tau}.$$

Set $u(\cdot) := \alpha(v(\cdot))$. Then

$$\begin{array}{ll} \multicolumn{2}{l}{d_D^2(x[x, u(\cdot), v(\cdot)](\tau))} \\ \leq & \|x[\bar{x}, \bar{u}(\cdot), \bar{v}(\cdot)](\tau) - x[x, u(\cdot), v(\cdot)](\tau)\|^2 \\ \leq & \|p_z\|^2 e^{2\lambda\tau} + \|y[\bar{y}, \bar{u}(\cdot)](\tau) - y[y, u(\cdot)](\tau)\|^2 \\ \leq & (1 + a\tau)\|p_z\|^2 + \|y[\bar{y}, \bar{u}(\cdot)](\tau) - y[y, u(\cdot)](\tau)\|^2. \end{array}$$

If we set $\bar{y}(\cdot) := y[\bar{y}, \bar{u}(\cdot)]$, $y(\cdot) := y[y, u(\cdot)]$, and $\widetilde{y}(\cdot) := y[y, \widetilde{u}(\cdot)]$, we have

$$\begin{aligned} &\|\bar{y}(\tau) - y(\tau)\|^2 \\ &\quad = \|\bar{y}(\tau) - \widetilde{y}(\tau)\|^2 + \|\widetilde{y}(\tau) - y(\tau)\|^2 + 2\langle \bar{y}(\tau) - \widetilde{y}(\tau); \widetilde{y}(\tau) - y(\tau)\rangle \\ &\quad \leq \|p_y\|^2 e^{2\lambda\tau} + 4M^2\tau^2 + 2\langle -p_y; \widetilde{y}(\tau) - y(\tau)\rangle + 4M^2\tau^2 \\ &\quad \leq (1 + a\tau)\|p_y\|^2 + b\tau^2 \end{aligned}$$

for some constants a and b, since

$$\langle p_y ; y[y, \widetilde{u}(\cdot)](\tau)\rangle \;\geq\; \langle p_y ; y[y, u(\cdot)](\tau)\rangle$$

from the very construction of $\widetilde{u}(\cdot)$. Therefore,

$$d_D^2(x[x, u(\cdot), v(\cdot)](\tau)) \leq (1 + a\tau)\|p\|^2 + b\tau^2,$$

which is the desired result since $\|p\| = d_D(x)$.

Necessary condition. Let us assume that D is not a discriminating domain for H. We are going to prove the existence of some point x_0 and of a nonanticipative strategy $\alpha : \mathcal{V}(z_0) \to \mathcal{U}(y_0)$ of positive ε and T such that, for any $v(\cdot) \in \mathcal{V}(z_0)$, the solution $x[x_0, \alpha(v(\cdot)), v(\cdot)]$ leaves $D + \varepsilon B$ before T and avoids $\mathcal{E} + \varepsilon B$ on $[0, T]$.

Let $x_0 = (y_0, z_0) \in \partial D$ be such that the normal condition is not satisfied. Clearly $x_0 \notin \mathcal{E}$, and there is some $p = (p_y, p_z) \in NP_D(x_0)$ and some $\gamma > 0$ such that

$$\inf_{v \in V} \langle h(z_0, v), p_z\rangle + \sup_{u \in U(y_0)} \langle g(y_0, u), p_y\rangle \;\geq\; \gamma\;.$$

Let $\bar{u} \in U(y_0)$, which achieves the maximum of $\langle g(y_0, u), p_y\rangle$. Then, since $y \to g(y, U(y))$ is lower semicontinuous with convex compact values, there is a continuous selection $\widetilde{g}(y)$ of $g(y, U(y))$ such that $\widetilde{g}(y_0) = g(y_0, \bar{u})$ (by the Michael selection theorem in [5] for instance).

Let $y(\cdot)$ be any solution to $y'(t) = \widetilde{g}(y(t))$, $y(0) = y_0$, $y(t) \in K_U$ for any $t \geq 0$ (the Nagumo theorem [4] states that such a solution exists). There is some control $u(\cdot)$ such that $y(\cdot) = y[y_0, u(\cdot)]$ and $u(\cdot) \in \mathcal{U}(y_0)$ since $y(t) \in K_U$ for $t \geq 0$. Moreover, $y'(0) = g(y_0, \bar{u})$. In particular, there is some $\tau > 0$ such that,

$$\forall t \in (0, \tau),\; \langle y(t) - y_0, p_y\rangle \;\geq\; t \max_{u \in U(y_0)} \langle g(y_0, u), p_y\rangle - t\gamma/3.$$

For $\tau > 0$ sufficiently small and for any admissible control $v(\cdot) \in \mathcal{V}(z_0)$, we have

$$\forall t \in (0, \tau),\; \langle z[z_0, v(\cdot)] - z_0, p_z\rangle \;\geq\; t \inf_{v \in V(z_0)} \langle h(z_0, v), p_z\rangle - t\gamma/3.$$

Let $v(\cdot) \in \mathcal{V}(z_0)$ be any admissible control. Then, since p is a proximal normal to D at x_0, for any $t \in (0, \tau)$,

$$d_D(x[x_0, u(\cdot), v(\cdot)](t)) \geq \|p\| - \|x[x_0, u(\cdot), v(\cdot)](t) - x_0 - p\|.$$

Note that

$$\begin{aligned}
&\|x[x_0, u(\cdot), v(\cdot)](t) - x_0 - p\|^2 \\
&\leq \|p\|^2 - 2\langle y[y_0, u(\cdot)](t) - y_0, p_y\rangle - 2\langle z[z_0, v(\cdot)](t) - z_0, p_z\rangle + Ct^2 \\
&\leq \|p\|^2 - 2t(\gamma - 2\gamma/3) + Ct^2
\end{aligned}$$

for some constant C. So

$$d_D(x[x_0, u(\cdot), v(\cdot)](t)) \geq \|p\| - [\|p\|^2 - 2t\gamma/3 + Ct^2]^{1/2},$$

which is positive for any $t \in (0, \tau)$ if $\tau > 0$ is sufficiently small.

In conclusion, the desired nonanticipative strategy $\alpha : \mathcal{V}(z_0) \to \mathcal{U}(y_0)$ is the constant map $\alpha(v(\cdot)) = u(\cdot)$.

4.3. Proof of Theorem 2.6. We first prove that $Disc_H(\mathcal{K})$ is equal to Victor's victory domain. This runs as in the proof of Theorem 2.2 of [11], where we use the following sequence K_n instead of the sequence K_n of [11]:

$$\begin{cases} K_0 = \mathcal{K}, \\ K_{n+1} := \left\{ x = (y,z) \in K_n \mid \begin{array}{c} \forall u(\cdot) \in \mathcal{U}(y),\ \exists v(\cdot) \in \mathcal{V}(z),\ \exists \tau \in [0,+\infty], \\ \text{such that } x[x,u(\cdot),v(\cdot)](\tau) \in \mathcal{E} \text{ if } \tau < +\infty, \text{ and} \\ \forall t \in [0,\tau),\ x[x,u(\cdot),v(\cdot)](t) \in K_n \end{array} \right\}. \end{cases}$$

It is easy to check that $\bigcap_n K_n$ is a discriminating domain. Hence $\bigcap_n K_n \subset Disc_H(\mathcal{K})$. On the other hand, since the K_n contain $Disc_H(\mathcal{K})$, one has $\bigcap_n K_n = Disc_H(\mathcal{K})$. Hence $\bigcap_n K_n = Disc_H(\mathcal{K})$.

For the characterization of Ursula's victory domain, we introduce the complement of Ursula's victory domain:

$$\mathcal{L} := \left\{ x_0 \in \mathcal{K} \mid \begin{array}{c} \forall \alpha(\cdot),\ T \geq 0,\ \varepsilon > 0,\ \exists v(\cdot) \in \mathcal{V}(z_0), \\ \text{such that, if } x(\cdot) := x[x_0, \alpha(v(\cdot)), v(\cdot)], \\ \text{then, either } \forall t \in [0,T], x(t) \in \mathcal{K} + \varepsilon B \\ \text{or } \exists \tau \leq T,\ x(\tau) \in \mathcal{E} + \varepsilon B, \text{ and} \\ x(t) \in \mathcal{K} + \varepsilon B \text{ for } t \in [0,\tau] \end{array} \right\}.$$

Thanks to Theorem 2.4, $Disc_H(\mathcal{K}) \subset \mathcal{L}$. So we have to prove the converse inclusion.

For doing this, the key argument of [11, Lemma 5.1] is replaced by the following lemma.

LEMMA 4.3. *If x_0 belongs to $\mathcal{K}$ but not to $\mathcal{L}$, there are positive η, ε, and T and, for any $x := (y,z) \in (x_0 + \eta B) \cap \mathcal{K}$, a nonanticipative strategy $\alpha : \mathcal{V}(z) \to \mathcal{U}(y)$ such that, for any $v(\cdot) \in \mathcal{V}(z)$, the solution $x[x, \alpha(v(\cdot)), v(\cdot)]$ does not reach $\mathcal{E} + \varepsilon B$ and leaves $\mathcal{K} + \varepsilon B$ before T. Namely, if we set $x(\cdot) := x[x, \alpha(v(\cdot)), v(\cdot)]$,*

$$\exists \tau \in [0,T],\ x(\tau) \notin \mathcal{K} + \varepsilon B,\ \textit{and}\ x(t) \notin \mathcal{E} + \varepsilon B\ \textit{for}\ t \in [0,\tau]\ .$$

Lemma 4.3 states that the ε and T appearing in the definition of Ursula's victory domain are locally uniform.

Proof of Lemma 4.3. From the very definition of $\mathcal{L}$, if x_0 does not belong to $\mathcal{L}$, there is a nonanticipative strategy $\alpha_0(\cdot)$, T and $\varepsilon_0 > 0$ such that for any $v(\cdot) \in \mathcal{V}(z_0)$, if we set $x_0(\cdot) := x[x_0, \alpha_0(v(\cdot)), v(\cdot)]$,

$$\exists \tau \in [0,T],\ x_0(\tau) \notin \mathcal{K} + \varepsilon_0 B \text{ and } x_0(t) \notin \mathcal{E} + \varepsilon_0 B \text{ for } t \in [0,\tau]\ . \tag{4.1}$$

In the proof of Lemma 5.1 of [11], the same strategy $\alpha_0(\cdot)$ gave the desired result for any $x \in (x_0 + \eta B)$—provided that η was sufficiently small. The situation is more complicated here because α_0 is not necessarily an admissible strategy for $x \in (x_0 + \eta B) \cap \mathcal{K}$ because of the constraints. For solving this difficulty, we use the following lemma proved below. Although we shall apply this lemma indifferently to y and to z, we only formulate it for y.

LEMMA 4.4. *Let Q and T be fixed positive constants. There is some $\lambda > 0$ (depending on the constants of the problem on Q and on T) such that, for any y and y_0 belonging to K_U, with $\|y\| \leq Q$ and $\|y_0\| \leq Q$, there is a nonanticipative strategy $\sigma : \mathcal{U}(y) \to \mathcal{U}(y_0)$ with, for any $u(\cdot) \in \mathcal{U}(y)$ and for $t \in [0,T]$,*

$$\|y[y_0, \sigma(u(\cdot))](t) - y[y, u(\cdot)](t)\| \leq \|y - y_0\| e^{\lambda t}\ .$$

Let us complete the proof of Lemma 4.3. We fix a constant Q sufficiently large in such a way that any solution starting from a point of $x_0 + B$ remains in the ball QB on $[0, T]$. Fix $\eta := \varepsilon_0 e^{-\lambda T}/4$, where λ is defined by Lemma 4.4. Let $x := (y, z) \in (x_0 + \eta B)$.

From Lemma 4.4, there is a nonanticipative map $\sigma_1 : \mathcal{V}(z) \to \mathcal{V}(z_0)$ such that, for any $v(\cdot) \in \mathcal{V}(z)$ and for any $t \in [0, T]$,

$$\|z[z, v(\cdot)](t) - z[z_0, \sigma_1(v(\cdot))](t)\| \leq \eta e^{\lambda t} .$$

Then $\alpha_0 \circ \sigma_1 : \mathcal{V}(z) \to \mathcal{U}(y_0)$ is a nonanticipative map. From Lemma 4.4 again, there is a nonanticipative strategy $\sigma_2 : \mathcal{U}(y_0) \to \mathcal{U}(y)$ such that, for $t \in [0, T]$,

$$\|y[y_0, u(\cdot)](t) - y[y, \sigma_2(u(\cdot))](t)\| \leq \eta e^{\lambda t}$$

for any $u(\cdot) \in \mathcal{U}(y)$. Then $\alpha := \sigma_2 \circ \alpha_0 \circ \sigma_1$ is a nonanticipative strategy from $\mathcal{V}(z)$ to $\mathcal{U}(y)$.

Fix some control $v(\cdot) \in \mathcal{V}(z)$. Set $x_0(\cdot) := x[x_0, \alpha_0(\sigma_1(v(\cdot))), \sigma_1(v(\cdot))]$ and $x(\cdot) := x[x, \alpha(v(\cdot)), v(\cdot)]$.

Then, for $\tau \in [0, T]$ defined in formula (4.1),

$$d_{\mathcal{K}}(x(\tau)) \geq d_{\mathcal{K}}(x_0(\tau)) - \|x(\tau) - x_0(\tau)\| \geq \varepsilon_0 - 2\eta e^{\lambda T} \geq \varepsilon_0/2 .$$

In the same way, $d_{\mathcal{E}}(x(t)) \geq \varepsilon_0/2$ on $[0, \tau]$, which completes the proof of Lemma 4.3.

Proof of Lemma 4.4. From Lemma 1.1, there is some $\lambda > 0$ such that, for any $u_0(\cdot) \in \mathcal{U}(y_0)$ and for any $y \in K_U$, there is some control $u(\cdot) \in \mathcal{U}(y)$ with

$$\|y[y, u(\cdot)](t) - y[y_0, u_0(\cdot)](t)\| \leq \|y - y_0\| e^{\lambda t} \tag{4.2}$$

for any $t \in [0, T]$. Let us consider the set-valued map $\Sigma : \mathcal{U}(y_0) \to \mathcal{U}(y)$ defined by

$$\Sigma(u_0(\cdot)) := \{u(\cdot) \in \mathcal{U}(y) \mid (4.2) \text{ is satisfied}\}.$$

It is easy to check that Σ is nonanticipative in the sense of [15], so that, from the Plaskacz lemma (Lemma 2.7 of [15]), it enjoys a nonanticipative selection σ, i.e., σ is nonanticipative and $\sigma(u_0(\cdot)) \in \Sigma(u_0(\cdot))$ for any $u_0 \in \mathcal{U}(y_0)$.

REFERENCES

[1] B. Alziary de Roquefort, *Jeux différentiels et approximation numérique de fonctions valeur* I: *Étude theorique*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 517–533.

[2] B. Alziary de Roquefort, *Jeux différentiels et approximation numérique de fonctions valeur* II: *Étude numérique*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 535–560.

[3] M. Arisawa and P.L. Lions, *Continuity of admissible trajectories for state constraints control problems*, Discrete Contin. Dynam. Systems, 2 (1996), pp. 297–305.

[4] J.-P. Aubin, *Viability Theory*, Birkhäuser Boston, Boston, MA, 1991.

[5] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, MA, 1990.

[6] M. Bardi, S. Bottacin, and M. Falcone, *Convergence of discrete schemes for discontinuous value functions of pursuit-evasion games*, in New Trends in Dynamic Games and Applications, Ann. Internat. Soc. Dynam. Games 3, Birkhäuser Boston, Boston, MA, 1995, pp. 273–304.

[7] M. Bardi, S. Koike, and P. Soravia, *Pursuit-Evasion Game with State Constraints: Dynamic Programming and Discrete-Time Approximations*, preprint, 1998.

[8] L. Berkovitz, *Differential games of generalized pursuit and evasion*, SIAM J. Control Optim., 24 (1986), pp. 361–373.

[9] P. Bernhard, *Contribution à l'étude des Jeux Différentiels à deux joueurs, somme nulle, et information parfaite*, Thèse de Doctorat d'Etat, Université Pierre et Marie Curie—Paris 6, Paris, France, 1979.

[10] J.V. Breakwell, *Zero-sum differential games with terminal payoff*, in Differential Games and Applications, Lecture Notes in Control and Inform. Sci. 3, P. Hagedorn, H.W. Knobloch, and G.H. Olsder, eds., Springer-Verlag, New York, 1977, pp. 70–95.

[11] P. Cardaliaguet, *A differential game with two players and one target*, SIAM J. Control Optim., 34 (1996), pp. 1441–1460.

[12] P. Cardaliaguet, *Non smooth semi-permeable barriers, Isaacs equation and application to a differential game with one target and two players*, Appl. Math. Optim., 36 (1997), pp. 125–146.

[13] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Contribution à l'étude des jeux différentiels quantitatifs et qualitatifs avec contraintes sur l'état*, C.R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 1543–1548.

[14] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Set-valued numerical analysis for optimal control and differential games*, in Stochastic and Differential Games: Theory and Numerical Methods, Ann. Internat. Soc. Dynam. Games 4, M. Bardi, T.E.S. Raghavan, and T. Parthasarathy, eds., Birkhäuser Boston, Boston, MA, 1999, pp. 177–247.

[15] P. Cardaliaguet and S. Plaskacz, *Invariant solutions of differential games and Hamilton–Jacobi–Isaacs equations for time-measurable Hamiltonians*, SIAM J. Control Optim., 38 (2000), pp. 1501–1520.

[16] R.J. Elliot and N.J. Kalton, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126 (1972).

[17] R.J. Elliot and N.J. Kalton, *The existence of value in differential games of pursuit and evasion*, J. Differential Equations, 12 (1972), pp. 504–523.

[18] L.C. Evans and P.E. Sougandinis, *Differential games and representation formulas for solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 282 (1984), pp. 487–502.

[19] W. H. Fleming, *The convergence problem for differential games*, J. Math. Anal. Appl., 3 (1967), pp. 102–116.

[20] J. Flynn, *Lion and man: The boundary constraint*, SIAM J. Control Optim., 11 (1973), pp. 397–411.

[21] H. Frankowska, M. Plascasz, and T. Rzezuchowski, *Measurable viability theorem and Hamilton–Jacobi–Bellman equations*, J. Differential Equations, 116 (1995), pp. 265–305.

[22] H. Frankowska and F. Rampazzo, *Filippov's and Filippov–Wazewski's theorems on closed domains*, J. Differential Equations, 161 (2000), pp. 449–478.

[23] A. Friedman, *On the definition of differential games and the existence of value and saddle points*, J. Differential Equations, 7 (1970), pp. 69–91.

[24] A. Friedman, *Existence of value and saddle points for differential games of pursuit and evasion*, J. Differential Equations, 7 (1970), pp. 92–125.

[25] A. Friedman, *Differential games with restricted phase coordinates*, J. Differential Equations, 8 (1970), pp. 135–162.

[26] R. Isaacs, *Differential Games*, Wiley, New York, 1965.

[27] N.N. Krasovskii and A.I. Subbotin, *Game-Theorical Control Problems*, Springer-Verlag, New-York, 1988.

[28] P. Loreti and M.E. Tessitore, *Approximation and regularity results on constrained viscosity solutions of Hamilton–Jacobi–Bellman equations*, J. Math. Systems Estim. Control, 4 (1994), pp. 467–483.

[29] Ju. S. Osipov, *An alternative in a differential-difference game*, Dokl. Akad. Nauk SSSR, 197 (1971), pp. 1022–1025 (in Russian).

[30] S. Plaskacz, *private communication*, 1994.

[31] N.N. Petrov, *On the existence of a value for pursuit games*, Soviet Mat. Dokl., 11 (1970), pp. 292–294.

[32] E. Roxin, *The axiomatic approach in differential games*, J. Optim. Theory Appl., 3 (1969), pp. 153–163.

[33] I. Rozyev and A.I. Subbotin, *Semicontinuous solutions of Hamilton–Jacobi equations*, Prikl. Math. Mekh., 52 (1988), pp. 141–146 (in Russian).

[34] H.M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[35] H.M. Soner, *Optimal control with state-space constraint* II, SIAM J. Control Optim., 24 (1986), pp. 1110–1112.

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

[37] P.P. Varaiya, *On the existence of solutions to a differential game*, SIAM J. Control, 5 (1967), pp. 153–162.

[38] P. Varaiya and J. Lin, *Existence of saddle points in differential games*, SIAM J. Control, 7 (1969), pp. 141–157.

[39] J.M. Yong, *On differential pursuit games*, SIAM J. Control Optim., 26 (1988), pp. 478–495.

SIAM J. Control Optim.
Vol. 39, No. 5, pp. 1633–1649

A PROXIMAL POINT METHOD FOR THE VARIATIONAL INEQUALITY PROBLEM IN BANACH SPACES*

REGINA SANDRA BURACHIK[†] AND SUSANA SCHEIMBERG[‡]

Abstract. In this paper we prove well-definedness and weak convergence of the generalized proximal point method when applied to the variational inequality problem in reflexive Banach spaces. The proximal version we consider makes use of Bregman functions, whose original definition for finite dimensional spaces has here been properly extended to our more general framework.

Key words. maximal monotone operators, proximal point algorithm, Banach spaces, convergence, algorithmic scheme

AMS subject classifications. Primary, 90C25; Secondary, 49D45, 49D37

PII. S0363012998339745

1. Introduction.

Let B be a reflexive Banach space and $\Omega \subset B$ a nonempty closed and convex set. Given $T\colon B \to \mathcal{P}(B^*)$ a maximal monotone operator, we consider the *classical variational inequality problem* for T and Ω, $VIP(T,\Omega)$, defined by

Find $x^* \in \Omega$ such that there exists $y^* \in T(x^*)$ with

$$\langle y^*, x - x^* \rangle \geq 0 \tag{1.1}$$

for all $x \in \Omega$, where $\langle \cdot, \cdot \rangle$ stands for the dual product in B.

In the particular case in which T is a subdifferential, i.e., $T = \partial\varphi$, where $\varphi : B \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex, and lower semicontinuous functional, (1.1) reduces to the nonsmooth constrained optimization problem

$$\min_{x\in\Omega} \varphi(x). \tag{1.2}$$

The following natural hypotheses will be assumed in our study:

H_1: $D(T)^0 \cap \Omega \neq \emptyset$ or $D(T) \cap \Omega^0 \neq \emptyset$.

Our aim is to study the algorithm given by:

1. Take $x^0 \in \Omega \cap D(T)$.
2. Given x^k, define x^{k+1} by the inclusion

$$0 \in (T + N_\Omega + \lambda_k \nabla f)x^{k+1} - \lambda_k \nabla f(x^k), \tag{1.3}$$

where f is a Bregman function (see Definition 2.1 in section 2.1), $\lambda_k > 0$, and $N_\Omega(\cdot)$ is the normality operator associated to Ω.

3. If $x^{k+1} = x^k$, STOP.

When $B = \Omega = H$, for H a Hilbert space, and $f = \frac{1}{2}\|\cdot\|_2^2$, algorithm (1.3) reduces

*Received by the editors July 9, 1998; accepted for publication (in revised form) August 29, 2000; published electronically February 23, 2001.

http://www.siam.org/journals/sicon/39-5/33974.html

[†] Engenharia de Sistemas e Computação, COPPE–UFRJ, CP 68511, Rio de Janeiro-RJ, CEP 21945-970, Brazil (regi@cos.ufrj.br). The research of this author was partially supported by PRONEX–Optimization.

[‡] Universidade Federal do Rio de Janeiro, UFRJ, COPPE–IM, CP 68511, Rio de Janeiro-RJ, CEP 21945-970, Brazil (susana@cos.ufrj.br). The research of this author was partially supported by CNPq grant 302393/85-4(RN).

to the *classical* proximal point method in a Hilbert space [43]. Thus, scheme (1.3) can be seen as an extension to a Banach space of the proximal point method. The intention of this paper is to study existence and convergence of the sequence $\{x^k\}$ given by (1.3).

An important motivation for analyzing the convergence properties of algorithm (1.3) is related to the so-called *mesh independence principle* [2, 1, 28, 32, 35]. The mesh independence principle relies on infinite-dimensional convergence results for predicting the convergence properties of the discretized finite-dimensional method. Furthermore, it provides a theoretical foundation for the justification of refinement strategies and helps to design this refinement process. Since the focus is the infinite-dimensional solution, a fine discretization scheme has to be chosen so that the discrete solution approximates the infinite-dimensional solution appropriately. Many real-world problems in economics and engineering are modeled in infinite-dimensional spaces. These include optimal control problems, shape optimization problems, and the problem of minimal area surface with obstacles, among many others. In many shape optimization problems [35], the function space is only a Banach and not a Hilbert space, motivating an analysis in this more general framework. For solving constrained optimization problems in Banach spaces, the application of (1.3) to the saddle point operator associated with the problem has been proposed in [37]. The connection of these methods with our results will be discussed in section 4.

The family of Bregman functions we consider in this work includes, for instance, $f(x) = \|x\|_p^p$ when $B = L^p$ or $B = l_p$, with $p \in (1, +\infty)$. Hence, the case $f = \frac{1}{2}\|\cdot\|_2^2$ considered in the classical proximal point method can be seen as a particular case of (1.3). A natural question is, Why should we use a generic Bregman function in (1.3) instead of $f = \frac{1}{2}\|\cdot\|^2$? A first answer is that inclusion (1.3) can be substantially simplified by an appropriated choice of f. For instance, if $B = L^p$ or $B = l_p$, with $p \in (1, +\infty)$, then by choosing $f(x) = \|x\|_p^p$, (1.3) is much simpler than its version for $f(x) = \frac{1}{2}\|x\|_p^2$ (see [13]). Second, consider the case in which we have an optimization problem, i.e., $T = \partial\varphi$ in a Banach space contained in a Hilbert space H. Assume that φ can be extended to H. The proximal point method can then be applied in the whole space H, in which case, the generated sequence may (weakly) converge to an element not in B. In this situation, the use of a "proper" Bregman function defined on B makes sense.

Other extensions of the classical proximal point method to Banach spaces can be found in [17] and [36].

The notion of Bregman function has its origin in [5] and this name was first used by Censor and Lent in [19]. Bregman functions have been extensively used for convex optimization algorithms (e.g., [4], [19], [25], [20]) in finite-dimensional spaces. It has also been used for defining "generalized" versions of the proximal point method (e.g., [21], [23], [26], [30], [34] for finite-dimensional spaces, [10] for Hilbert spaces, and [40], [13], [15] for Banach spaces). A useful tool for comparing the Bregman distance with the distance induced by the norm of the Banach space leads to the notions of *modulus of convexity* and *total convexity* of f (the latter introduced in [8]). More material on these concepts can be found in [22] and [11]. The paper is organized as follows. Section 2 contains theoretical preliminaries. In subsection 2.1 we consider the concept of a Bregman function and examples in a Banach space. In subsection 2.2 we give definitions, properties, and examples related to *modulus of convexity* and *total convexity.* In subsection 2.3 we recall some classical material about maximal monotone operators. We also extend to a reflexive Banach space a result known to hold in Hilbert spaces [6, Lemme 1]. This result will ensure existence of the iterates. Section 3 establishes results concerning the existence and convergence of the iterates. Finally, we present in section 4 an application of algorithm (1.3) to the stochastic convex feasibility problem [9].

2. Theoretical preliminaries.

2.1. Bregman distances in Banach spaces.

Let B be a reflexive Banach space and $f : B \to \mathbb{R} \cup \{+\infty\}$ a strictly convex, proper, and lower semicontinuous function with closed domain $\mathcal{D} := dom(f)$. Assume from now on that $\mathcal{D}^\circ \neq \emptyset$ and that f is Gâteaux differentiable on $\mathcal{D}^\circ$.

The *Bregman distance* with respect to f is the function $D_f : \mathcal{D} \times \mathcal{D}^\circ \to \mathbb{R}$ defined by

$$D_f(z,x) := f(z) - f(x) - \langle \nabla f(x), z - x \rangle, \tag{2.1}$$

where $\nabla f(\cdot)$ is the differential of f defined in $\mathcal{D}^\circ$. The function $D_f(\cdot,\cdot)$ is not a distance in the usual sense of the term (in general, it is not symmetric and does not satisfy the triangular inequality). However, there is a "three point property" which takes the place of this inequality in the proofs.

PROPERTY 2.1. *Given $x \in \mathcal{D}$, $y, z \in \mathcal{D}^\circ$, the following equality is straightforward:*

$$\langle \nabla f(y) - \nabla f(z), z - x \rangle = D_f(x,y) - D_f(x,z) - D_f(z,y). \tag{2.2}$$

Consider the following set of assumptions on f:

B_1 : The *right* level sets of $D_f(y,\cdot)$:

$$S_{y,\alpha} := \{z \in \mathcal{D}^\circ \ : \ D_f(y,z) \leq \alpha\}$$

are bounded for all $\alpha \geq 0$ and for all $y \in \mathcal{D}$.

B_2 : If $\{x^k\} \subset \mathcal{D}^\circ$ and $\{y^k\} \subset \mathcal{D}^\circ$, $w - \lim_{k\to\infty} x^k = x$, $w - \lim_{k\to\infty} y^k = x$, and $\lim_{k\to\infty} D_f(x^k, y^k) = 0$, then

$$\lim_{k\to\infty} \left(D_f(x,x^k) - D_f(x,y^k)\right) = 0.$$

B_3 : If $\{x^k\} \subset \mathcal{D}$ is bounded, $\{y^k\} \subset \mathcal{D}^\circ$ is such that $w - \lim_{k\to\infty} y^k = y$ and $\lim_{k\to\infty} D_f(x^k, y^k) = 0$, then $w - \lim_{k\to\infty} x^k = y$.

B_4 : (zone coerciveness)
For every $y \in B^*$, there exists $x \in \mathcal{D}^\circ$ such that $\nabla f(x) = y$.

The following alternative condition will also be considered in our analysis.

B_2^* : If $\{x^k\} \subset \mathcal{D}^\circ$ and $\{y^k\} \subset \mathcal{D}^\circ$ are bounded sequences such that $\lim_{k\to\infty} \|x^k - y^k\| = 0$, then

$$\lim_{k\to\infty} \left(\nabla f(x^k) - \nabla f(y^k)\right) = 0.$$

Remark 2.1.

(i) Condition B_2^* will be used as an optional assumption in our convergent analysis. Examples of auxiliary functions which satisfy this property are relevant. It has been proved in [16] that for the important cases in which $B = \ell_p$ or $B = L_p$, with $p > 1$, condition B_2^* is satisfied for the family of functions $f(x) = \|x\|^s$, $s > 1$.

(ii) When $B = \mathbb{R}^n$, assumption B_2^* implies B_2.

(iii) Condition B_2 has been considered for the first time in [10] for the case in which $B = H$ a Hilbert space. When $B = R^n$, Censor and Lent [19] considered the condition if $\{x^k\} \subset \mathcal{D}^\circ$, $\lim_{k\to\infty} x^k = x$, then $\lim_{k\to\infty} D_f(x, x^k) = 0$. We point out that this assumption is stronger than B_2. Conditions B_1 and B_3 are natural extensions of the ones required in [19].

Definition 2.1. (i) *We say that f is a* Bregman function *if it satisfies B_1–B_3.*
(ii) *The function f is said to be a* coercive *Bregman function if it satisfies B_1–B_4.*

In our study, the following assumption will be made on the Bregman function f:
$H_2 : \Omega \subset \mathcal{D}^\circ$.

2.2. Total convexity. Let $\mathbb{R}_{++} := \{\alpha \in \mathbb{R} \ : \ \alpha > 0\}$ and $\mathbb{R}_{+} := \{\alpha \in \mathbb{R} \ : \ \alpha \geq 0\}$. Let $\mathcal{D} \subset B$ be a closed and convex set, with $\mathcal{D}^\circ \neq 0$ and $f : B \to \mathbb{R} \cup \{+\infty\}$ a convex function which is Gâteaux differentiable in $\mathcal{D}^\circ$. Following [12], we define the *modulus of convexity* of f, $\nu_f : \mathcal{D}^\circ \times \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\nu_f(z,t) := \inf\{D_f(x,z) \ : \ \|x-z\| = t\}, \tag{2.3}$$

where $D_f(\cdot,\cdot)$ is given by (2.1). The function f is said to be *totally convex* in $\mathcal{D}^\circ$ if and only if $\nu_f(z,t) > 0$ for all $z \in \mathcal{D}^\circ$ and $t > 0$. The result below, which will be useful in what follows, has been proved in [12].

Property 2.2. *Let $z \in \mathcal{D}^\circ$. The function $\nu_f(z,\cdot)$ is* increasing *on $\mathbb{R}_{++}$, i.e., if $0 < \alpha < \beta$, then $\nu_f(z,\alpha) < \nu_f(z,\beta)$.*

Totally convex functions are strictly convex, and in finite dimension total convexity is equivalent to strict convexity, but in infinite dimensional Banach spaces (e.g., in ℓ_p) there exist strictly convex functions which are not totally convex (see [12]). On the other hand, total convexity is a weaker condition than uniform convexity (i.e., $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \gamma \|x-y\|^p$ for some $\gamma > 0$, $p > 1$). Uniformly convex functions are totally convex (see [12]), but in the spaces L^p and ℓ_p with $1 < p < 2$ the function $f(x) = \|x\|_p^p$ is totally convex (see [12], [16]), while it is not uniformly convex. We will see later on that total convexity also turns out to be a key property in our convergence analysis. We are now ready to present our last assumption on f.

B_5 : (uniform total convexity). The function f is said to be *uniformly totally convex* if for any bounded set $K \subset \mathcal{D}^\circ$, and any $t \in \mathbb{R}_{++}$, it holds that

$$\inf_{x\in K} \nu_f(x,t) > 0 \ . \tag{2.4}$$

Remark 2.2.

(i) This condition is called *sequential consistency* in [11].

(ii) Assumption B_5 will be an alternative requirement on f in our convergence analysis. When $B = \ell_p$ or $B = L^p$, with $p > 1$, the family of functions $f(x) = \|x\|_p^s$, $s > 1$ is uniformly totally convex (see [16]). Moreover, the latter result has been extended in [14], where the authors prove that in any uniformly convex Banach space the function $f(x) = \|x\|_p^s$, $s > 1$ is uniformly totally convex.

(iii) Let $f : B \to \mathbb{R} \cup \{+\infty\}$ be a Gâteaux differentiable function on $\mathcal{D}^\circ$ such that it verifies assumptions B_2^* and B_5. Then condition B_2 holds. Indeed, take two sequences $\{x^k\}, \{y^k\} \subset \mathcal{D}^\circ$, weakly converging (and thus bounded) to the same point x, such that

$$\lim_{k\to\infty} D_f(x^k, y^k) = 0. \tag{2.5}$$

We must prove that

$$\lim_{k\to\infty} \big(D_f(x,x^k) - D_f(x,y^k)\big) = 0. \tag{2.6}$$

First we see that under this condition, it holds that $\lim_{k\to\infty} \|x^k - y^k\| = 0$. If this is not true, there exists a subsequence $\{x^{k_j} - y^{k_j}\}_{j\geq 0}$ and a positive number α such that

$$\|x^{k_j} - y^{k_j}\| > \alpha, \text{ for all } j \geq 0. \tag{2.7}$$

Let C be a bounded set which contains the sequence $\{y^k\}$. We have that

$$\begin{array}{rcl} D_f(x^{k_j}, y^{k_j}) & \geq & \inf\{D_f(z, y^{k_j}) \, : \, \|x^{k_j} - y^{k_j}\| = \|z - y^{k_j}\|\} \\ & = & \nu_f(y^{k_j}, \|x^{k_j} - y^{k_j}\|) \\ & > & \nu_f(y^{k_j}, \alpha) \\ & \geq & \inf_{y\in C} \nu_f(y, \alpha) > 0, \end{array}$$

where we used that the function $\nu_f(z,\cdot)$ is increasing and B_5. This contradicts the fact that $\lim_{k\to\infty} D_f(x^k, y^k) = 0$. Thus it holds that $\lim_{k\to\infty} \|x^k - y^k\| = 0$. By B_2^*, this implies that

$$\lim_{k\to\infty} \|\nabla f(x^k) - \nabla f(y^k)\| = 0. \tag{2.8}$$

By Property 2.1,

$$D_f(x,x^k) - D_f(x,y^k) = \langle \nabla f(y^k) - \nabla f(x^k), x - x^k\rangle - D_f(x^k, y^k).$$

Combining now (2.5) and (2.8), we obtain (2.6).

2.3. Maximal monotone operators in a reflexive Banach space. For an arbitrary point to set operator $A : B \to \mathcal{P}(B^*)$, we recall the following definitions:
domain of A:

- $D(A) := \{y \in B \, : \, A(y) \neq \emptyset\}$,

graph of A:

- $G(A) := \{(y,v) \in B \times B^* \, : \, v \in A(y)\}$,

range of A:

- $R(A) := \{v \in B^* \, : \, v \in A(y) \text{ for some } y \in B\}$.

The operator $T : B \to \mathcal{P}(B^*)$ is *monotone* if

$$\langle u - v, x - y\rangle \geq 0$$

for all $x, y \in B$ and all $u \in T(x), v \in T(y)$. A monotone operator is called *maximal* if for any other monotone operator $\tilde{T}$ with $\tilde{T}(x) \supseteq T(x)$ for all $x \in B$, it holds that $\tilde{T} = T$.

Maximal monotone operators have an important closedness property.

PROPOSITION 2.2 (see [38, p. 105]). *Any maximal monotone operator $S : B \to \mathcal{P}(B^*)$ is* demiclosed*, i.e., the conditions below hold:*

If $\{x^k\}$ converges weakly to x and $\{w^k \in Sx^k\}$ converges strongly to w, then $w \in Sx$.

If $\{x^k\}$ converges strongly to x and $\{w^k \in Sx^k\}$ converges weakly to w, then $w \in Sx$.

The following well-known result is very useful for asserting maximality of the sum of maximal monotone operators.

Proposition 2.3 (see [42, Theorem 1]). *Let B be a reflexive Banach space, and let T_1 and T_2 be maximal monotone operators from B to $\mathcal{P}(B^*)$. If $D(T_1) \cap D(T_2)^0 \neq \emptyset$, then $T_1 + T_2$ is a maximal monotone operator.*

Definition 2.4. *An operator A is said to be* coercive (*see* [38]) *if $D(A)$ is bounded or*

$$\lim_{\substack{(z,v) \in G(A) \\ \|z\| \to +\infty}} \frac{\langle v, z \rangle}{\|z\|} = +\infty. \tag{2.9}$$

The following definition has been introduced in [3] and [6]. It also appears in [39].

Definition 2.5. *An operator $A : B \to \mathcal{P}(B^*)$ is said to be* regular *if*

$$\text{for all } u \in R(A) \text{ and for all } y \in D(A), \quad \sup_{(z,v) \in G(A)} \langle v - u, y - z \rangle < \infty. \tag{2.10}$$

Proposition 2.6 (see [6, p. 167]). *The subdifferential of a proper lower semicontinuous convex function φ is a regular operator.*

Recall that for a nonempty convex and closed set $\Omega \subset B$, the normality operator $N_\Omega : B \to \mathcal{P}(B^*)$ is given by

$$N_\Omega(x) = \begin{cases} \{\phi \in B^* \mid \langle \phi, z - x \rangle \leq 0 \text{ for any } z \in \Omega\} & \text{if } x \in \Omega\,, \\ \emptyset & \text{if } x \notin \Omega\,. \end{cases}$$

It is easy to check that $N_\Omega(\cdot)$ is the subdifferential of the indicator function χ_Ω associated to the set Ω, i.e.,

$$\chi_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega\,, \\ +\infty & \text{if } x \notin \Omega\,. \end{cases}$$

The function χ_Ω is proper convex and lower semicontinuous; hence $N_\Omega(\cdot)$ is a maximal monotone operator. For further use we emphasize that $x \in \Omega^0$ if and only if $N_\Omega(x) = 0$.

Recall that the normalized duality mapping $J : B \to \mathcal{P}(B^*)$ is defined by the property $v \in Ju$ if and only if $\|v\|_{B^*}^2 = \|u\|_B^2 = \langle v, u \rangle$ for any $u \in B$, where $\|\cdot\|_X$ means the norm in the space X. The next property, which we introduce here, will allow us to obtain the existence of the iterates in (1.3). It extends to a Banach space a lemma proved in [6] for Hilbert spaces.

Lemma 2.7. *Assume that B is a reflexive Banach space and J its normalized duality mapping. Let $C, T_0 : B \to \mathcal{P}(B^*)$ be maximal monotone operators such that*

(a) *C is regular,*

(b) *$D(T_0) \cap D(C) \neq \emptyset$ and $R(C) = B^*$,*

(c) *$C + T_0$ is maximal monotone.*

Then $R(C + T_0) = B^*$.

Proof. The proof of this lemma consists of two steps.

Step 1.

In this step we prove the following statement: Let T_1 be a maximal monotone operator such that there exists a convex set $F \subset B^*$ with the following condition: for all $u \in F$ there exists an element $y \in B$ such that

$$\sup_{(z,v) \in G(T_1)} \langle v - u, y - z \rangle < \infty, \tag{2.11}$$

then $F^0 \subset R(T_1)$.

Step 2.

In this part of the proof, we apply the statement of Step 1 for the choices $F := R(C) + R(T_0) (= B^*$ by condition (b)) and $T_1 := C + T_0$. Namely, we show here that the set $F = B^*$ satisfies (2.11) for the operator $C + T_0$. Then the conclusion of Step 1 readily gives $R(T_1) = R(T_0 + C) = B^*$.

Step 1.

If $F^0 = \emptyset$, the conclusion trivially holds. Assume now that $F^0 \neq \emptyset$. For any $f \in F$, we can consider the inclusion

$$f \in (T_1 + \varepsilon J)(x_\varepsilon), \;\; \varepsilon > 0, \tag{2.12}$$

where J is the normalized duality mapping in B. We observe that (2.12) has a unique solution thanks to Browder's result (see [7]). Take $v_\varepsilon \in Jx_\varepsilon$ such that $f - \varepsilon v_\varepsilon \in T_1 x_\varepsilon$. Let $a \in B$ be such that (2.11) holds for $y = a$, i.e., there exists $c \in \mathbb{R}$ such that

$$\text{for all } (z, v) \in G(T_1), \; \langle v - f, a - z\rangle \leq c. \tag{2.13}$$

For the choice $(z, v) := (x_\varepsilon, f - \varepsilon v_\varepsilon)$, (2.13) writes

$$\varepsilon\|x_\varepsilon\|^2 \leq \varepsilon\langle v_\varepsilon, a\rangle + c.$$

Using the fact that for any $p, q \in B$, $w \in Jp$ $\|p\|^2 - 2\langle w, q\rangle + \|q\|^2 \geq 0$, the inequality above becomes

$$\frac{\varepsilon}{2}\|x_\varepsilon\|^2 \leq \frac{\varepsilon}{2}\|a\|^2 + c,$$

which gives $\sqrt{\varepsilon}\|x_\varepsilon\| \leq r$ for some $r > 0$ and $\varepsilon > 0$ small enough.

Now let $\tilde{f} \in F^0$ and $\rho > 0$ such that $\tilde{f} + B^*(0, \rho) := \{\tilde{f} + \varphi \in B^* \; : \; \|\varphi\| < \rho\} \subset F$. For any $\varphi \in B^*(0, \rho)$, it holds by (2.11) that there exists $a(\varphi) \in B$ and $c(\varphi) \in \mathbb{R}$ such that

$$\text{for all } (z, v) \in G(T_1), \; \langle v - (\tilde{f} + \varphi), a(\varphi) - z\rangle \leq c(\varphi). \tag{2.14}$$

Take in (2.12) $f = \tilde{f}$ and $x_\varepsilon = \tilde{x}_\varepsilon$. Then choose $(z, v) := (\tilde{x}_\varepsilon, \tilde{f} - \varepsilon\tilde{v}_\varepsilon)$ in (2.14) to get

$$\langle -\varepsilon\tilde{v}_\varepsilon - \varphi, a(\varphi) - \tilde{x}_\varepsilon\rangle \leq c(\varphi).$$

Some rearrangements of the expression above yield

$$\langle \varphi, \tilde{x}_\varepsilon\rangle \leq c(\varphi) + \langle \varphi, a(\varphi)\rangle + \varepsilon\langle \tilde{v}_\varepsilon, a(\varphi)\rangle - \varepsilon\|\tilde{x}_\varepsilon\|^2,$$

which implies

$$\langle \varphi, \tilde{x}_\varepsilon\rangle \leq c(\varphi) + \langle \varphi, a(\varphi)\rangle + \frac{\varepsilon}{2}(\|a(\varphi)\|^2 - \|\tilde{x}_\varepsilon\|^2).$$

Altogether, we conclude

$$\langle \varphi, \tilde{x}_\varepsilon\rangle \leq c(\varphi) + \langle \varphi, a(\varphi)\rangle + \frac{\varepsilon}{2}\|a(\varphi)\|^2,$$

where we are now considering the elements $\tilde{x}_\varepsilon$ as functionals defined in B^*. Taking in the expression above $\varphi = -\psi \in B^*(0, \rho)$, we obtain a bound $K(\varphi)$ such that

$$|\tilde{x}_\varepsilon(\varphi)| \leq K(\varphi), \text{ for all } \varphi \in B^*(0, \rho). \tag{2.15}$$

Take now $\varepsilon = \frac{1}{k}$, and define $x^k := \tilde{x}_{\frac{1}{k}}$ for any k. By (2.15), $|x^k(\varphi)| \leq K(\varphi)$ for any k. Using now Banach–Saks–Steinhauss theorem (see [24]), there exists $\bar{K}$ a constant such that $|x^k(\varphi)| \leq \bar{K}$ for any k and for any $\varphi \in B^*(0,\rho)$. This implies that the sequence $\{x^k\}$ is bounded and hence it has a weakly convergent subsequence. Call $\{x^{k_j}\}$ such a subsequence and $\bar{x}$ its weak limit. Take $v^{k_j} \in Jx^{k_j}$ such that (2.12) holds for $f = \tilde{f}$ and $x_\varepsilon = x^{k_j}$. Observe that J maps bounded sets on bounded sets. Thus, there exists a subsequence of $\{x^{k_j}\}$ (which we still call $\{x^{k_j}\}$ for simplicity) such that

$$\begin{array}{l} w - \lim_{j\to\infty} x^{k_j} = \bar{x}, \\ \tilde{f} - \frac{1}{k_j} v^{k_j} \in T_1(x^{k_j}), \\ w - \lim_{j\to\infty} \tilde{f} - \frac{1}{k_j} v^{k_j} = \tilde{f}. \end{array} \tag{2.16}$$

Using now maximality of T_1, we will show that $\tilde{f} \in T_1(\bar{x})$. Indeed, for any $(z,v) \in G(T_1)$, we have that

$$0 \leq \left\langle \tilde{f} - \frac{1}{k_j} v^{k_j} - v, x^{k_j} - z \right\rangle = \left\langle \tilde{f} - \frac{1}{k_j} v^{k_j}, x^{k_j} - z \right\rangle - \langle v, x^{k_j} - z\rangle.$$

It is clear that $\lim_{j\to\infty}\langle v, x^{k_j} - z\rangle = \langle v, \bar{x} - z\rangle$ and $\lim_{j\to\infty}\langle \tilde{f}, x^{k_j} - z\rangle = \langle \tilde{f}, \bar{x} - z\rangle$. On the other hand,

$$\frac{1}{k_j}|\langle v^{k_j}, x^{k_j} - z\rangle| \leq \frac{1}{k_j}\|v^{k_j}\|\,\|x^{k_j} - z\|,$$

the rightmost term tending to zero because $\{x^{k_j}\}$ and hence $\{v^{k_j}\}$ are bounded, and $k_j \to \infty$. This yields

$$0 \leq \langle \tilde{f} - v, \bar{x} - z\rangle,$$

which implies by maximality that $\tilde{f} \in T_1(\bar{x})$, as we claimed.

Step 2.

We must prove that the set $F = R(C) + R(T_0)$ satisfies (2.11) for the operator $C + T_0$. Let $u \in R(C) + R(T_0)$, $x \in D(C) \cap D(T_0)$, and $w \in T_0(x)$. Then

$$u = w + (u - w).$$

Since $R(C) = B^*$, we can find $y \in B$ such that $(u - w) \in C(y)$. Using now regularity of C, we know that given $(u - w) \in R(C)$ and $x \in D(C)$, there exists some $c \in \mathbb{R}$ with $\sup_{(z,s)\in G(C)}\langle s - (u - w), x - z\rangle \leq c$, which implies that for any $(z,s) \in G(C)$

$$\langle s - (u - w), x - z\rangle \leq c. \tag{2.17}$$

Let $z \in D(T_0) \cap D(C)$ and $v \in T_0(z)$, by monotonicity we get

$$\langle v - w, x - z\rangle \leq 0. \tag{2.18}$$

Adding (2.17) and (2.18) we obtain

$$\langle (s + v) - u, x - z\rangle \leq c$$

for any $s \in C(z)$, $v \in T_0(z)$, i.e., for any $s + v =: t \in (C + T_0)z$. Therefore,

$$\sup_{(z,t)\in G(C+T_0)} \langle t - u, x - z\rangle < \infty.$$

This establishes (2.11) for $F = R(C) + R(T_0)$ and $T_1 = C + T_0$. Now by Step 1 and the fact that $F = B^*$, we obtain that $C + T_0$ is onto. $\square$

Our convergence theorems require two conditions on the operator T, namely para- and pseudomotonicity, which we discuss next. The notion of paramonotonicity was introduced in [18] and further studied in [29]. It is defined as follows. Let $T : B \to \mathcal{P}(B^*)$, and Ω a closed and convex set.

DEFINITION 2.8. *T is paramonotone in Ω if it is monotone and $\langle z-z', w-w'\rangle = 0$ with z, $z' \in \Omega$, $w \in T(z)$, $w' \in T(z')$ implies $w \in T(z'), w' \in T(z)$.*

The next proposition, whose proof can be found in [29], presents the main properties of paramonotone operators.

PROPOSITION 2.9. (i) *If T is the subdifferential ∂f of a convex function $f : B \to \mathbb{R}$, then T is paramonotone in B.*

(ii) *If T is paramonotone in Ω, x^* solves $VIP(T, \Omega)$ and $\bar{x} \in \Omega$ satisfies that there exists an element $\bar{u} \in T(\bar{x})$ such that $\langle \bar{u}, x^* - \bar{x}\rangle \geq 0$, then $\bar{x}$ also solves $VIP(T, \Omega)$.*

(iii) *If T_1 and T_2 are paramonotone in Ω, then $T_1 + T_2$ is paramonotone in Ω.*

Next we recall the definition of pseudomonotonicity, which has been taken from [7], and should not be confused with other uses of the same word, e.g., [31].

DEFINITION 2.10. *Let B be a reflexive Banach space and $T : B \to \mathcal{P}(B^*)$ such that $D(T)$ is closed and convex. T is said to be pseudomonotone if and only if it satisfies the following condition:*

Take any sequence $\{x^k\} \subset D(T)$, converging weakly to an element $x^0 \in D(T)$ and any sequence $\{w^k\} \subset B^*$, with $w^k \in Tx^k$ for all k, such that

$$\limsup_k \langle w^k, x^k - x^0\rangle \leq 0.$$

Then for each $y \in D(T)$ there exists an element $w^0 \in Tx^0$, such that

$$\langle w^0, x^0 - y\rangle \leq \liminf_k \langle w^k, x^k - y\rangle.$$

PROPOSITION 2.11.

(i) *N_C is pseudomonotone for any $C \subset B$, a closed and convex set.*

(ii) *The sum of pseudomonotone operators is pseudomonotone.*

Proof. (i) Consider a sequence $\{(x^k, w^k) \in G(T)\}_k$ in the conditions of Definition 2.10. Call x^0 the weak limit of $\{x^k\}$. As $N_C(\cdot)$ is a closed cone, $0 \in N_C(x^0)$. Take now any $y \in D(N_C) = C$ and $w^0 = 0 \in N_C(x^0)$, then

$$\liminf_k \langle w^k, x^k - y\rangle \geq 0 = \langle w^0, x^0 - y\rangle,$$

which proves our claim.

(ii) See [38, p. 97]. $\square$

3. The variational inequality problem in a Banach space: Existence and convergence analysis. In this section we study first the existence of iterates in algorithm (1.3). It is known that a maximal monotone operator with bounded domain is surjective (see [7]); hence the iterates clearly exist when the domain of $T + N_\Omega + \nabla f$ is bounded. Other classical results due to Browder (see [7]) concerning surjectivity of the sum of maximal monotone operators apply if we assume that $\nabla f(\cdot)$ or T are coercive operators (see Definition 2.4), with no extra assumptions (besides the natural requirement of existence of solutions of the original problem). When $\nabla f(\cdot)$ or T are *not* coercive, what can be said about the existence of the sequence? We answer this question in the case of a coercive Bregman function.

COROLLARY 3.1. *Let f be a coercive Bregman function. Then algorithm* (1.3) *is well defined.*

Proof. Define the operators $T_0 := T + N_\Omega$ and $C := \lambda_k \nabla f$. It holds that

(a) C is regular (since it is the gradient of a proper closed convex function).

(b) $D(T_0) \cap D(C) = D(T) \cap \Omega \cap \mathcal{D}^\circ = D(T) \cap \Omega \neq \emptyset$ (by H_1 and H_2).

(c) $C + T_0$ is maximal monotone (by (b) and Proposition 2.3).

Then by Lemma 2.7, $T_0 + \lambda_k C = T + N_\Omega + \lambda_k \nabla f$ is onto. Hence, given $\lambda_k \nabla f(x^k) \in B^*$ there exists x^{k+1} such that

$$\lambda_k \nabla f(x^k) \in (T + N_\Omega + \lambda_k \nabla f)(x^{k+1}).$$

The uniqueness follows from the strict monotonicity of f. □

Now we analyze the convergence of the sequence given by (1.3) in a reflexive Banach space B. When the solution set is not empty, we get boundedness of the iterates (see Theorem 3.3). Moreover, we establish in Theorem 3.6 that boundedness of the iterates and nonemptyness of X^* are equivalent. Weak convergence of the whole sequence is established when $\nabla f(\cdot)$ is weak-to-weak continuous or when the set X^* of solutions of problem (1.1) is a singleton. Note that the iterative step of algorithm (1.3) can be rewritten in terms of the Bregman distance: given z^k, z^{k+1} solves the inclusion

$$0 \in (T + N_\Omega + \lambda_k \nabla D_f(\cdot, z^k))(z^{k+1}). \tag{3.1}$$

$$\text{If } z^{k+1} = z^k, \textbf{ stop}. \tag{3.2}$$

THEOREM 3.2. *Suppose that the sequence $\{z^k\}$ given by* (1.3) *is well defined and finite. Then the last term is a solution of* (1.1).

Proof. If the sequence is finite, then it must stop at step (3.2). In this case, inclusion (3.1) for $z^{k+1} = z^k$ writes

$$0 \in (T + N_\Omega + \lambda_k \nabla D_f(\cdot, z^k))(z^k) = (T + N_\Omega)(z^k), \tag{3.3}$$

where we used that $\nabla D_f(z^k, z^k) = \nabla f(z^k) - \nabla f(z^k) = 0$. Inclusion (3.3) readily implies that $z^k \in X^*$. □

From now on we assume that the sequence $\{z^k\}$ generated by (1.3) is well defined and infinite.

THEOREM 3.3. *Assume that $X^* \neq \emptyset$. Then it holds that*

(i) *The sequence $\{z^k\}$ generated by* (1.3) *is bounded,*

(ii)$\sum_{k=0}^{\infty} D_f(z^{k+1}, z^k) < \infty$.

Proof. Let $z^* \in X^*$. In order to prove the boundedness of $\{z^k\}$, we will show that the sequence $D_f(z^*, z^k)$ is decreasing, in which case the result will follow from boundedness of the level sets of the function $D_f(z^*, \cdot)$. By the "three point property" we have that

$$\begin{aligned} D_f(z, z^{k+1}) &= D_f(z, z^k) - D_f(z^{k+1}, z^k) \\ &+ \langle \nabla f(z^k) - \nabla f(z^{k+1}), z - z^{k+1} \rangle \end{aligned} \tag{3.4}$$

for any $z \in \Omega$. As we are supposing that the algorithm (1.3) is well defined, there exist $u^{k+1} \in Tz^{k+1}$ and $\varphi_{k+1} \in N_\Omega(z^{k+1})$ such that

$$\frac{1}{\lambda_k}(u^{k+1} + \varphi_{k+1}) = \nabla f(z^k) - \nabla f(z^{k+1}). \tag{3.5}$$

Combining (3.4) with (3.5) and using the definition of normality operator, expression (3.4) for $z = z^*$ becomes

$$D_f(z^*, z^{k+1}) \leq D_f(z^*, z^k) - D_f(z^{k+1}, z^k) - \frac{1}{\lambda_k}\langle u^{k+1}, z^{k+1} - z^*\rangle. \tag{3.6}$$

By definition of X^*, the rightmost term is nonnegative, implying

$$D_f(z^*, z^{k+1}) \leq D_f(z^*, z^k) - D_f(z^{k+1}, z^k). \tag{3.7}$$

Therefore the sequence $D_f(z^*, z^k)$ is decreasing and, as pointed out before, this fact ensures the boundedness of the sequence $\{z^k\}$. Actually, taking $\alpha = D_f(z^*, z^0)$ we have that $\{z^k\} \subset S_{z^*,\alpha}$, which is a bounded set (condition B_1). This completes the proof of (i).

(ii) As the sequence $\{D_f(z^*, z^k)\}$ is also bounded below, it converges to a limit, which we call l^*. By (3.7),

$$D_f(z^{k+1}, z^k) \leq D_f(z^*, z^k) - D_f(z^*, z^{k+1}). \tag{3.8}$$

Summing up inequality (3.8),

$$\sum_{k=0}^{\infty} D_f(z^{k+1}, z^k) \leq \sum_{k=0}^{\infty}(D_f(z^*, z^k) - D_f(z^*, z^{k+1})) = D_f(z^*, z^0) - l^* < \infty,$$

and (ii) is established. □

The theorem above establishes boundedness of the sequence, and thus existence of weak accumulation points. Now we establish below two different hypotheses on the data under which the weak accumulation points are in fact solutions of the original problem (1.1). The first hypothesis is an assumption on the Bregman function f alone, and is the total uniform convexity, together with B_2^* instead of B_2. The second hypothesis is a requirement on the operator T, which we ask to be para- and pseudomonotone.

THEOREM 3.4. *Let X^* be nonempty and $\lambda_k < \lambda$. Suppose further any of the assumptions below:*

A_1: *T is pseudo- and paramonotone with closed domain,*

A_2: *f is a Bregman function with B_2^* instead of B_2, which is also uniformly totally convex.*

Then any weak accumulation point is a solution of the $VIP(T, \Omega)$.

Proof. By (i) of Theorem 3.3, there exists a subsequence $\{z^{k_j}\} \subset \{z^k\}$ which is weakly convergent to a point $\bar{z}$. Using now Theorem 3.3 (ii), $\lim_{j\to\infty} D_f(z^{k_j+1}, z^{k_j}) = 0$. Condition B_3 implies that also $\lim_{j\to\infty} z^{k_j+1} = \bar{z}$.

Case 1: Assume condition A_1. We use B_2 for $x^j = z^{k_j+1}$ and $y^j = z^{k_j}$, which yields

$$\lim_{j\to\infty}(D_f(\bar{z}, z^{k_j+1}) - D_f(\bar{z}, z^{k_j})) = 0. \tag{3.9}$$

Using (3.5), (3.4) for $z =: \bar{z}$, and (3.9) we obtain

$$\begin{aligned} 0 &= \lim\nolimits_{j\to\infty}(D_f(\bar{z}, z^{k_j+1}) - D_f(\bar{z}, z^{k_j}) - D_f(z^{k_j+1}, z^{k_j})) \\ &\leq \liminf\nolimits_{j\to\infty} \tfrac{1}{\lambda_{k_j}}\langle u^{k_j+1}, \bar{z} - z^{k_j+1}\rangle. \end{aligned}$$

By pseudomonotonicity of T, for $z^* \in X^*$, there exists $u \in T(\bar{z})$ with the property

$$\langle u, \bar{z} - z^* \rangle \le \liminf_{j\to\infty} \langle u^{k_j+1}, z^{k_j+1} - z^* \rangle = 0,$$

where the last equality is obtained by (3.6). In fact, for $u^* \in T(z^*)$,

$$\begin{aligned} 0 &\le \langle u^*, z^{k_j+1} - z^* \rangle \le \langle u^{k_j+1}, z^{k_j+1} - z^* \rangle \\ &\le \lambda_{k_j} (D_f(z^*, z^{k_j}) - D_f(z^*, z^{k_j+1}) - D_f(z^{k_j+1}, z^{k_j})), \end{aligned}$$

where we used monotonicity in the second inequality. The claim follows taking limit for $j \to \infty$. By Proposition 2.9(ii), $\bar{z}$ is a solution of $VIP(T, \Omega)$.

Case 2: Consider hypothesis A_2. We follow the same steps as in Remark 2.2(ii) for $\{x^j\} = \{z^{k_j}\}$ and $\{y^j\} = \{z^{k_j+1}\}$. It holds that $\lim_{j\to\infty} \|z^{k_j} - z^{k_j+1}\| = 0$. By B_2^*, we have that $\lim_{j\to\infty}(\nabla f(z^{k_j}) - \nabla f(z^{k_j+1})) = 0$. Now the definition of the algorithm yields $\lim_{j\to\infty} \frac{1}{\lambda_{k_j}}(u^{k_j+1} + \varphi^{k_j+1}) = 0$, for some $u^{k_j+1} \in Tz^{k_j+1}$ and $\varphi^{k_j+1} \in N_\Omega(z^{k_j+1})$. Our hypothesis on λ implies that $\lim_{j\to\infty}(u^{k_j+1} + \varphi^{k_j+1}) = 0$. By demiclosedness of the graph of $T + N_\Omega$ (see Proposition 2.2), we conclude that $0 \in (T + N_\Omega)(\bar{z})$, as we wanted to prove. □

Remark 3.1. The result below establishes conditions under which the whole sequence defined by algorithm (1.3) converges weakly to a solution. A crucial hypothesis is the weak-to-weak continuity of ∇f. This requirement is fulfilled when $X = l^p$ and $f(x) = 1/p\|x\|_p^p$.

THEOREM 3.5. *Consider the same assumptions as in Theorem* 3.4. *If X^* is a singleton or if ∇f is a weak-to-weak continuous mapping, then the whole sequence converges weakly to a solution, i.e., there exists a unique weak accumulation point.*

Proof. The statement is obvious if X^* is a singleton. For the second claim suppose that $\nabla f(\cdot)$ is weak-to-weak continuous. We will show that there exists only one weak accumulation point. Suppose there are two points z_1, z_2, which are weak limits of subsequences of $\{z^k\}$. By Theorems 3.3 and 3.4, we know that z_1 and z_2 belong to X^*. By (3.7), the sequences $\{D_f(z_1, z^k)\}$ and $\{D_f(z_2, z^k)\}$ are convergent. Call l_1 and l_2 their limits. Then

$$\begin{aligned} &\lim_{k\to\infty} \left(D_f(z_1, z^k) - D_f(z_2, z^k)\right) = l_1 - l_2 \\ &= f(z_1) - f(z_2) + \lim_{k\to\infty} \langle \nabla f(z^k), z_2 - z_1 \rangle. \end{aligned} \tag{3.10}$$

Call $l := \lim_{k\to\infty} \langle \nabla f(z^k), z_2 - z_1 \rangle$. If $w - \lim_{j\to\infty} z^{k_j} = z_1$, and $w - \lim_{j\to\infty} z^{l_j} = z_2$, taking $k = k_j$ in (3.10) and using the weak-to-weak continuity of $\nabla f(\cdot)$, we get that $l = \langle \nabla f(z_1), z_2 - z_1 \rangle$. Repeating the same argument with $k = l_j$ in (3.10), we get $l = \langle \nabla f(z_2), z_2 - z_1 \rangle$. Hence, $\langle \nabla f(z_2) - \nabla f(z_1), z_2 - z_1 \rangle = 0$. By strict convexity of f, we conclude that $z_1 = z_2$, which establishes the uniqueness of the weak accumulation point. □

The result below establishes that boundedness of the sequence given by algorithm (1.3) is a sufficient and necessary condition for existence of solutions of problem (1.1).

THEOREM 3.6. *Consider the assumptions A_1 or A_2 of Theorem* 3.4. *If $X^* = \emptyset$, then the sequence is unbounded. Conversely, if $X^* \neq \emptyset$, then the sequence is bounded.*

Proof. The second claim follows from Theorem 3.3(i). We prove the first one. Suppose that $X^* = \emptyset$ and $\{z^k\}$ bounded. Then there are a convex, closed, and bounded set S and a positive number r such that

$$\overline{\{z^k\}}^w \subset B(0, r) \subset S^0, \tag{3.11}$$

where $\overline{X}^w$ stands for the weak closure and $B(0,r) := \{y \in B : \|y\| \leq r\}$ is a (strongly and weakly) closed ball in B. Define the operator $\tilde{T} := T + N_S$, where N_S is the normality operator associated with S. The operator $\tilde{T}$ is maximal monotone by Proposition 2.3 and the fact that $D(T) \cap S^0 \supset \{z^k\} \neq \emptyset$.

Consider the sequence $\{\tilde{z}^k\}$ generated by method (1.3) for $\tilde{T}$. First, we prove that this sequence is well defined. Recall that any maximal monotone operator with bounded domain is onto (see [7]). The operator $\tilde{T}$ has bounded domain, and so the same happens with the operator $T^k(\cdot) := (\tilde{T} + N_\Omega + \lambda_k \nabla f)(\cdot) - \lambda_k \nabla f(z^k)$ used in each iteration. Thus, T^k is onto, and so $\tilde{z}^{k+1}$ exists. The uniqueness of $\tilde{z}^{k+1}$ follows from strict monotonicity of $\nabla f(\cdot)$. Altogether, we have that $\tilde{z}^{k+1}$ is uniquely defined as the solution of the inclusion on y:

$$0 \in (\tilde{T} + N_\Omega + \lambda_k \nabla f)(y) - \lambda_k \nabla f(\tilde{z}^k). \tag{3.12}$$

Thus, we established the well-definedness of the sequence $\{\tilde{z}^k\}$.

Our second step will be to prove by induction that if method (1.3) is applied to $\tilde{T}$ with $\tilde{z}^0 = z^0$, then $\tilde{z}^k = z^k$ for all k. The claim is true for $k = 0$. Suppose that $\tilde{z}^k = z^k$. The point z^{k+1} satisfies

$$0 \in (T + N_\Omega + \lambda_k \nabla f) z^{k+1} - \lambda_k \nabla f(z^k). \tag{3.13}$$

Since z^{k+1} belongs to S^0, we have that $N_S(z^{k+1}) = 0$, which, together with (3.13) and the induction hypothesis, gives

$$0 \in (T + N_S)(z^{k+1}) + N_\Omega(z^{k+1}) + \lambda_k(\nabla f(z^{k+1}) - \nabla f(\tilde{z}^k)).$$

Therefore, z^{k+1} is the unique solution of (3.12), and, as a consequence, $\tilde{z}^{k+1} = z^{k+1}$. The induction step is complete and we conclude that $\tilde{z}^k = z^k$ for all k. Our goal is to prove that any weak accumulation point of $\{z^k\}$ is a solution of $VIP(T, \Omega)$, thus contradicting the hypothesis $X^* = \emptyset$. As $\tilde{z}^k = z^k$ for all k, it will be enough to prove that any weak accumulation point of $\{\tilde{z}^k\}$ is a solution of $VIP(T, \Omega)$. Before we establish this fact, let us observe that if z is a weak accumulation point of $\{\tilde{z}^k\}$ which solves $VIP(\tilde{T}, \Omega)$, then it also solves $VIP(T, \Omega)$. Indeed, let $z \in \overline{\{\tilde{z}^k\}}^w$ be such that there exists $u \in \tilde{T}(z)$ with $\langle u, x - z\rangle \geq 0$ for any $x \in \Omega$. As $z \in \overline{\{\tilde{z}^k\}}^w = \overline{\{z^k\}}^w \subset S^0$, it holds that $\tilde{T}(z) = T(z) + N_S(z) = T(z)$. Hence, there exists $u \in T(z) = \tilde{T}(z)$ such that $\langle u, x - z\rangle \geq 0$ for any $x \in \Omega$. This proves that z solves $VIP(T, \Omega)$. It only remains to prove that any weak accumulation point of $\{\tilde{z}^k\}$ (and hence of $\{z^k\}$) solves $VIP(\tilde{T}, \Omega)$. In order to establish this result, let us check that the requirements of Theorem 3.4 hold for $\tilde{T}$ under conditions A_1 or A_2. In order to do this, let us check first that the set of solutions of $VIP(\tilde{T}, \Omega)$ is not empty. As mentioned above, the elements of this set are the zeroes of $\tilde{T} + N_\Omega = T + N_S + N_\Omega$. Note that $D(\tilde{T} + N_\Omega) = D(T) \cap \Omega \cap S \subset S$. Since S is bounded, $\tilde{T} + N_\Omega$ is onto (see [7]), and therefore it has zeroes. Then $VIP(\tilde{T}, \Omega)$ is not empty.

We prove now that any of the assumptions A_1 or A_2 on the data imply that algorithm (1.3) applied to $\tilde{T}$ generates a sequence whose accumulation points are all solutions of $VIP(\tilde{T}, \Omega)$. This is trivial for hypothesis A_2, which does not depend on the operator.

Assume now hypothesis A_1. Since $\tilde{T} = T + \partial(\chi_S)$, $\tilde{T}$ is paramonotone as a sum of two paramonotone operators (see Proposition 2.9(iii)). It follows from Proposition 2.11(i) that N_S is pseudomonotone. Thus $\tilde{T}$ is the sum of two pseudomonotone operators, and therefore it is pseudomonotone (see Proposition 2.11(ii)). Hence all the requirements of Theorem 3.4 hold for $\tilde{T}$. This implies that any weak accumulation point of $\{\tilde{z}^k\}$ is a solution of $VIP(\tilde{T},\Omega)$.

By boundedness, $\{\tilde{z}^k\}$, and therefore $\{z^k\}$, has weak cluster points and, as proved in Theorem 3.4, all of them are solutions of $VIP(\tilde{T},\Omega)$. As observed above this implies that all this elements are also solutions of $VIP(T,\Omega)$, in contradiction with our hypothesis. We conclude that $\{z^k\}$ is unbounded. □

4. An application: Augmented Lagrangians in Banach spaces. The connection between augmented Lagrangian methods and proximal-type methods in the context of finite dimensional spaces or Hilbert spaces has been the subject of intense investigation since the pioneer work of [44], [43]; see, for example, [27], [26], [34]. In an arbitrary reflexive Banach space, augmented Lagrangians have been studied for the first time in [11], [37], as an application for solving the *stochastic convex feasibility problem* (*SCFP*) (see [9]). This problem is the minimization of a convex function subject to (possibly infinitely many) constraints. In [11], the authors define a Lagrangian and a dual problem for (SCFP). They establish Karush–Khun–Tucker conditions, which, under reasonable circumstances, are necessary and sufficient for a pair to be a primal-dual solution of (SCFP). It is proved in [11] that primal-dual solutions of (SCFP) are the zeroes of a maximal monotone operator $\tilde{T}$, associated to the Lagrangian. Hence, algorithm (1.3) for finding zeroes of monotone operators can be applied to $\tilde{T}$, as long as we find a Bregman function which satisfies conditions B_2^* and B_5 in the primal and dual variables. First we describe briefly (SCFP). Second we present a particular instance of (SCFP) for which a specific Bregman function satisfying conditions B_2^* and B_5 can be given.

Problem (SCFP):

Take X to be a reflexive separable Banach space with dual X^*, $f : X \to \mathbb{R}$ a convex and continuously differentiable function, Λ a nonempty set, and $(\Lambda, \mathcal{A}, \mu)$ a complete probability space. For defining the constraints of the problem, consider $F : X \times \Lambda \to \mathbb{R}$ such that

- $F(\cdot,\lambda) : X \to \mathbb{R}$ is convex and continuously differentiable.
- $F(x,\cdot)$, $\partial_x F(x,\cdot) : \Lambda \to \mathbb{R}$ belong to $\mathcal{L}^p(\Lambda)$, with $p > 1$.

Consider the problem

$$\text{(SCFP)} \qquad \begin{array}{c} \min f(x) \\ \text{subject to } F(x,\lambda) \le 0 \quad \text{a.e. } (\Lambda), \end{array}$$

where the expression "a.e. (Λ)" means that

$$\mu(\{\lambda \in \Lambda \mid F(x,\lambda) \le 0\}) = 1.$$

The Lagrangian $L : X \times \mathcal{L}^q(\Lambda) \to \overline{\mathbb{R}}$ associated with (SCFP) is defined as

$$L(x,y) := \begin{cases} f(x) + \int_\Lambda F(x,\lambda) y(\lambda) d\mu(\lambda) & \text{if } y(\lambda) \ge 0 \ \text{ a.e. } (\Lambda), \\ \\ -\infty & \text{c.c.} \end{cases}$$

Take $q = p/(p-1)$. In a similar way as in [44], define the (maximal monotone) operator $\tilde{T} : X \times \mathcal{L}^q(\Lambda) \to \mathcal{P}(X^* \times \mathcal{L}^p(\Lambda))$ as

$$\tilde{T}(x,y) := (\partial_x L(x,y), -\partial_y L(x,y)).$$

Then the *dual* problem associated with the *primal* problem (SCFP) is find $y \in \mathcal{L}^q(\Lambda)$ such that

$$y \in arg\max\{\varphi(z) \mid z \in \mathcal{L}^q(\Lambda),\ z(\lambda) \ge 0 \quad \text{a.e. } (\Lambda)\},$$

where $\varphi : \mathcal{L}^q(\Lambda) \to [-\infty, \infty)$ is the function

$$\varphi(z) := \inf\{L(x, z) \mid x \in X\}.$$

Karush–Khun–Tucker conditions can be stated in this context, and under mild assumptions a pair (x, y) satisfies these conditions if and only if it is a zero of $\tilde{T}$ [11, Chapter 2]. These are the *primal-dual optimal pairs* associated to (SCFP). In [37], a doubly augmented Lagrangian method is defined. It requires auxiliary Bregman functions g, h in the primal and dual variables, respectively.

Particular instance of (SCFP):

Take $X = \mathcal{L}^s(\Lambda)$, $s > 1$, $g : \mathcal{L}^s(\Lambda) \to \mathbb{R}$, and $h : \mathcal{L}^q(\Lambda) \to \mathbb{R}$ given by

$$g(x) := (1/s)\|x\|_s^s\,, \quad h(y) := (1/q)\|y\|_q^q.$$

The fact that g and h satisfy conditions B_2^* and B_5 has been observed in Remarks 2.1(i) and 2.2(ii), respectively. We can now set in the definition of algorithm (1.3):

$$\begin{aligned} &B := \mathcal{L}^s(\Lambda) \times \mathcal{L}^q(\Lambda),\\ &\Omega := \mathcal{L}^s(\Lambda),\\ &T := \tilde{T},\\ &f(x, y) := g(x) + h(y). \end{aligned} \tag{4.1}$$

As f satisfies A_2 of Theorem 3.4, algorithm (1.3) generates a bounded sequence such that any weak accumulation point is a primal-dual solution of (SCFP), as long as there exist primal-dual solutions of (SCFP). If Λ is denumerable, then by Remark 3.1 and Theorem 3.5, the whole sequence converges weakly to a primal-dual solution.

Acknowledgements. We thank Professor Alfredo N. Iusem and Professor B. F. Svaiter for their constructive remarks. We also thank two anonymous referees, whose comments greatly improved this work.

REFERENCES

[1] E. L. Allgower and K. Böhmer, *Application of the mesh independence principle to mesh refinement strategies*, SIAM J. Numer. Anal., 24 (1987), pp. 1335–1351.

[2] E. L. Allgower, K. Böhmer, F.-A. Potra, and W. C. Rheinboldt, *A mesh independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.

[3] A. Auslender, *Optimisation. Méthodes Numériques*, Masson, Paris, 1976.

[4] H. H. Bauschke and J. M. Borwein, *Legendre Functions and the method of random Bregman functions*, J. Convex Anal., 4 (1997), pp. 27–67.

[5] L. M. Bregman, *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[6] H. Brézis and A. Haraux, *Image d'une somme d'opérateurs monotones et applications*, Israel J. Math., 23 (1976), pp. 165–186.

[7] F. E. Browder, *Nonlinear operators and nonlinear equations of evolution in Banach spaces*, in Proceedings Sympos. Pure Math. 18, AMS, Providence, RI, 1976, pp. 1–308.

[8] D. Butnariu, Y. Censor, and S. Reich, *Iterative averaging of entropic projections for solving stochastic convex feasibility problems*, Comput. Optim. Appl., 8 (1997), pp. 21–39.

[9] D. Butnariu and S. J. Flam, *Strong convergence of expected projection methods in Hilbert spaces*, Numer. Funct. Anal. Optim., 16 (1995), pp. 601–636.

[10] R. S. Burachik and A. N. Iusem, *A generalized proximal point algorithm for the variational inequality problem in a Hilbert space*, SIAM J. Optim., 8 (1998), pp. 197–216.

[11] D. Butnariu and A. N. Iusem, *Totally Convex Functions for Fixed Point Computation and Infinite Dimensional Optimization*, Kluwer, Dordrecht, The Netherlands, 2000.

[12] D. Butnariu and A. N. Iusem, *Local moduli of convexity and their applications to finding almost common points of measurable families of operators*, in Recent Developments in Optimization Theory and Nonlinear Analysis, Y. Censor and S. Reich, eds., Contemp. Math. 204, AMS, Providence, RI, 1997, pp. 61–91.

[13] D. Butnariu and A. N. Iusem, *On a proximal point method for convex optimization in Banach spaces*, Numer. Funct. Anal. Optim., 18 (1997), pp. 723–744.

[14] D. Butnariu, A. N. Iusem, and E. Resmerita, *Total convexity of powers of the norm in uniformly convex Banach spaces*, J. Convex Anal., 7 (2000), pp. 319–324.

[15] R. S. Burachik, D. Butnariu, and A. N. Iusem, *Iterative methods for solving stochastic convex feasibility problems and applications*, Comput. Optim. Appl., 15 (2000), pp. 269–307.

[16] D. Butnariu, C. A. Isnard, and A. N. Iusem, *A mixed Hölder and Minkowski inequality*, Proc. Amer. Math. Soc., 127 (1999), pp. 2405–2415.

[17] R. E. Bruck and S. Reich, *Nonexpansive projections and resolvents of accretive operators in Banach spaces*, Houston J. Math., 3 (1977), pp. 459–470.

[18] Y. Censor, A. N. Iusem, and S. A. Zenios, *An interior point method with Bregman functions for the variational inequality problem with paramonotone operators*, Math. Programming, 81 (1998), pp. 373–400.

[19] Y. Censor and A. Lent, *An iterative row-action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321–353.

[20] Y. Censor, A. De Pierro, T. Elfving, G. T. Herman, and A. N. Iusem, *On iterative methods for linearly constrained entropy maximization*, Numer. Anal. Math. Modelling, A. Waculitz, ed., Banach Center Publication Series, Banach Center, Warsaw, 24 (1990), pp. 145–163.

[21] Y. Censor and S. A. Zenios, *The proximal minimization algorithm with with D-functions*, J. Optim. Theory Appl., 73 (1992), pp. 451–464.

[22] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms and Applications*, Oxford University Press, New York, 1997.

[23] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like optimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.

[24] M. Cotlar and R. Cignoli, *An Introduction to Functional Analysis*, North-Holland, Amsterdam, 1974.

[25] A. De Pierro and A. N. Iusem, *A relaxed version of Bregman's method for convex programming*, J. Optim. Theory Appl., 51 (1986), pp. 421–440.

[26] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.

[27] D. Gabay, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods Mathematics: Application to the Solution of Boundary Value Problems, M. Fortain and R. Glowinski, eds., North-Holland, Amsterdam, 1983.

[28] M. Heinkenschloss, *Mesh independence for nonlinear least squares problems with norm constraints*, SIAM J. Optim., 3 (1993), pp. 81–117.

[29] A. N. Iusem, *On some properties of paramonotone operators*, J. Convex Anal., 5 (1998), pp. 269–278.

[30] A. N.Iusem, *On some properties of generalized proximal point methods for quadratic and linear programming*, J. Optim. Theory Appl., 85 (1995), pp. 593–612.

[31] S. Karamardian, *Complementarity problems over cones with monotone and pseudomonotone maps*, J. Optim. Theory Appl., 18 (1976), pp. 445–455.

[32] C. T. Kelley and E. W. Sachs, *Mesh independence of Newton-like methods for infinite dimensional problems*, J. Integral Equations Appl., 3 (1991), pp. 549–573.

[33] K. Kiwiel, *Free-steering relaxation methods for problems with strictly convex costs and linear constraints*, Math. Oper. Res., 22 (1997), pp. 326–349.

[34] K. C. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.

[35] M. Laumen, *Newton's mesh independence principle for a class of optimal shape design problems*, SIAM J. Control Optim., 37 (1999), pp. 1070–1088.

[36] O. Nevanlinna and S. Reich, *Strong convergence of contraction semigroups and of iterative methods for accretive operators in Banach spaces*, Israel J. Math., 32 (1979), pp. 44–58.

[37] R. Otero and A. N. Iusem, *Inexact Versions of Proximal Point and Augmented Lagrangian Algorithms in Banach Spaces*, unpublished manuscript.

[38] D. Pascali and S. Sburlan, *Nonlinear Mappings of Monotone Type*, Martinus Nijhoff, Dordrecht, The Netherlands, 1978.

[39] S. Reich, *The range of sums of accretive and monotone operators*, J. Math. Anal. Appl., 68

(1979), pp. 310–317.

[40] S. Reich, *A weak convergence theorem for the alternating method with Bregman distances*, in Theory and Applications of Nonlinear Operators of Accretive and Monotone Type, A. Kartsatos, ed., Marcel Dekker, New York, 1996, pp. 313–318.

[41] R. T. Rockafellar, *Local boundedness of nonlinear monotone operators*, Michigan Math. J., 16 (1969), pp. 397–407.

[42] R. T. Rockafellar, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.

[43] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[44] R. T. Rockafellar, *Augmented Lagrangians and applications to the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.

SIAM J. Control Optim.
Vol. 39, No. 5, pp. 1650–1650

ERRATUM: HOFFMAN'S ERROR BOUND, LOCAL CONTROLLABILITY, AND SENSITIVITY ANALYSIS[*]

ABDERRAHIM JOURANI[†]

Abstract. The proofs of Lemma 5.2 and Theorem 5.1 in [*SIAM J. Control Optim.,* 38 (2000), pp. 947–970] have a common error which concerns the existence of (locally) Lipschitz selections. Indeed, Lemma 5.2 is true in the case where the multivalued mapping G is convex-valued, and in our case this assumption is not satisfied. So some modifications are needed. The statement of Theorem 5.1 (which does not use Lemma 5.2) is as follows.

PII. S0363012900370466

Theorem 1. *Suppose that the graph GrF of the multivalued mapping F is closed and that*

$$(0, v^*) \in \partial^F_{(u,v)} h(\overline{u}, 0, \overline{y}) \quad \Longrightarrow \quad v^* = 0, \tag{1}$$

where $h(u,v,y) = d(u,v,GrF_y)$ and $F_y : u \to F(u,y)$. Suppose also that the multivalued mapping $y \to F_y(u)$ is pseudo-Lipschitzian around $\overline{y}$ uniformly in u in some neighborhood of $\overline{u}$; that is, there exist $a > 0$ and $r > 0$ such that for all $y, y' \in \overline{y} + rB_Y$, and $u \in \overline{u} + rB_U$

$$F_y(u) \cap rB_V \subset F_{y'}(u) + a\|y - y'\|B_V.$$

Then the multivalued mapping $(y, v) \to F_y^{-1}(v)$ is pseudo-Lipschitzian around $(\overline{y}, 0, \overline{u})$.

This allows us to modify the statement of Corollary 5.4 in [1].

Corollary 2. *Let f be locally Lipschitzian around the point $(\overline{u}, \overline{y})$ solution of*

$$f(u, y) = 0.$$

Suppose that

$$0 \in \partial^F_u (v^* \circ f)(\overline{u}, \overline{y}) \quad \Longrightarrow \quad v^* = 0.$$

Then the multivalued mapping S defined by

$$S(y) = \{x : \quad f(x, y) = 0\}$$

is pseudo-Lipschitzian around $(\overline{y}, \overline{u})$.

Consequently, I am withdrawing Lemma 5.2 and the third remark after it in [1].

REFERENCES

[1] A. Jourani, *Hoffman's error bound, local controllability, and sensitivity analysis,* SIAM J. Control Optim., 38 (2000), pp. 947-970.

[*] Received by the editors April 10, 2000; accepted for publication (in revised form) June 19, 2000; published electronically February 23, 2001.

http://www.siam.org/journals/sicon/39-5/37046.html

[†] Département de Mathématiques, Analyse Appliquée et Optimisation, Université de Bourgogne, BP 47 870, 21078 Dijon Cedex, France (jourani@u-bourgogne.fr).

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1651–1676

VARIATIONAL ANALYSIS OF THE ABSCISSA MAPPING FOR POLYNOMIALS[*]

JAMES V. BURKE[†] AND MICHAEL L. OVERTON[‡]

Abstract. The abscissa mapping on the affine variety $\mathcal{M}_n$ of monic polynomials of degree n is the mapping that takes a monic polynomial to the maximum of the real parts of its roots. This mapping plays a central role in the stability theory of matrices and dynamical systems. It is well known that the abscissa mapping is continuous on $\mathcal{M}_n$, but not Lipschitz continuous. Furthermore, its natural extension to the linear space $\mathcal{P}_n$ of polynomials of degree n or less is not continuous. In our analysis of the abscissa mapping, we use techniques of modern nonsmooth analysis described extensively in Variational Analysis (R. T. Rockafellar and R. J.-B. Wets, Springer-Verlag, Berlin, 1998). Using these tools, we completely characterize the subderivative and the subgradients of the abscissa mapping, and establish that the abscissa mapping is everywhere subdifferentially regular. This regularity permits the application of our results in a broad context through the use of standard chain rules for nonsmooth functions. Our approach is epigraphical, and our key result is that the epigraph of the abscissa map is everywhere Clarke regular.

Key words. nonsmooth analysis, polynomials, stability, subgradient, Clarke regular

AMS subject classifications. 34D05, 34D20, 49J52

PII. S0363012900367655

Let $\mathcal{P}_n$ denote the linear space of complex polynomials of degree n or less, and let $\mathcal{M}_n$ denote the affine variety in $\mathcal{P}_n$ consisting of the monic polynomials of degree n. In this article we study variational properties of the *abscissa mapping*

$$a : \mathcal{M}_n \to \mathbb{R}$$

given by

$$a(p) = \max \{ \operatorname{Re} \zeta \mid p(\zeta) = 0 \} .$$

Our study is partly motivated by the need to provide tools for understanding the variational behavior of the *spectral abscissa mapping* on the n by n complex matrices defined by

$$\alpha(M) = a(\det(\lambda I - M)).$$

Properties of the spectral abscissa are closely tied to stability theory for matrices and dynamical systems. Thus, the variational behavior of the spectral abscissa has important consequences for the sensitivity of the stability properties of such systems under perturbation. In [BO], we apply the variational results obtained in this paper to study the variational behavior of the spectral abscissa map.

The abscissa mapping has a number of characteristics that make it difficult to analyze. It is well known that a is continuous, but not Lipschitz continuous, on

[*] Received by the editors February 7, 2000; accepted for publication September 25, 2000; published electronically February 28, 2001.
http://www.siam.org/journals/sicon/39-6/36765.html

[†] Department of Mathematics, University of Washington, Seattle, WA 98195 (burke@math.washington.edu). The research of this author was supported in part by National Science Foundation grant DMS-9971852.

[‡] Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 (overton@cs.nyu.edu). The research of this author was supported in part by National Science Foundation grant CCR-9731777 and the U.S. Department of Energy contract DE-FG02-98ER25352.

$\mathcal{M}_n$. In addition, the natural extension of a to all of $\mathcal{P}_n$ is not continuous at any point of the subspace $\mathcal{P}_{n-1}$. In this paper, we show that the techniques of modern nonsmooth analysis described in the recent book [RW98] are ideally suited to the study of mappings of this type. Thus, a secondary purpose of this paper is to illustrate the usefulness of the nonsmooth analysis techniques developed by many authors over the last 30 years by applying them to a classical function of great practical importance. Using techniques from nonsmooth analysis, we are able to establish that the abscissa mapping is everywhere *subdifferentially regular*. This remarkable result has major consequences for the development of a calculus for the variational behavior of the abscissa mapping under composition.

It needs to be stated that our analysis owes a great debt to earlier work of Levantovskii [Lev80]. Levantovskii studied the set of stable polynomials, i.e., the set of polynomials whose abscissa is nonpositive, and provided an outline for the derivation of the tangent cone to this set. We generalize this proof technique to establish the key result of section 1 (Theorem 1.2).

The paper is organized as follows; we assume that the reader is familiar with [RW98]. Section 1 is devoted to the derivation of the *subderivative* of a. This is done via an *epigraphical* approach, where we derive the formula for the subderivative from a description of the *tangent cone* to the epigraph of the abscissa mapping a. In addition, we develop some basic tools that relate the prime factorization of a polynomial to a factorization of the tangent cone. The key to this result is the local factorization Lemma 1.4. In section 2, we use the representation of the tangent cone obtained in section 1 to derive a representation for the set of *regular normals* to the epigraph of a. This in turn yields a representation for the set of *regular subgradients* for a at any point in $\mathcal{M}_n$. In section 3, we establish that the abscissa mapping is everywhere subdifferentially regular. The key result is that the epigraph of the abscissa map is Clarke regular.

Most of the notation that we use is introduced as it is required. However, it is useful to briefly describe our conventions for discussing polynomials in their distinct roles as points in the linear space $\mathcal{P}_n$ and as functions over the complex field. One could identify $\mathcal{P}_n$ with $\mathbb{C}^{n+1}$ and attempt to derive the variational properties of a as a mapping on $\mathbb{C}^{n+1}$, but this would completely ignore the very rich underlying algebraic structure of polynomials. Since it is the roots of polynomials that lie at the heart of the mapping a, it is the polynomial perspective that drives our analysis. Given a polynomial $p \in \mathcal{P}_n$, we will always use the Greek letter λ to denote the indeterminant associated with representing the polynomial as a function. Thus we write $p(\lambda)$ as the associated polynomial function. Monomials and shifted monomials play a central role in our analysis. For this reason we give them a special notation so that we can discuss them as points in $\mathcal{P}_n$. We write

$$e_{(\ell,\lambda_0)}(\lambda) = (\lambda - \lambda_0)^\ell.$$

1. The subderivative and the tangent cone.

To apply the tools developed in [RW98], we first extend the definition of a to the entire linear space $\mathcal{P}_n$:

$$a : \mathcal{P}_n \to \mathbb{R}$$

is given by

$$a(p) = \begin{cases} \max\{\operatorname{Re}\zeta \mid p(\zeta) = 0\} & \text{if } p \in \mathcal{M}_n, \\ +\infty & \text{otherwise.} \end{cases}$$

This extension allows us to focus our attention on the set of monic polynomials. In particular, we have $\mathrm{dom}\,(a) = \{p \mid a(p) < +\infty\} = \mathcal{M}_n$. Given $p \in \mathcal{M}_n$, our goal is to derive a formula for $da(p)$, the *subderivative* of the mapping a. Following [RW98, Definition 8.1], the subderivative of a at a point $p \in \mathcal{M}_n$ is the mapping $da(p) : \mathcal{P}_n \to \mathbb{R} \cup \{\pm\infty\}$ given by

$$da(p)(\hat{q}) = \liminf_{\substack{\tau \searrow 0 \\ q \to \hat{q}}} \frac{a(p + \tau q) - a(p)}{\tau},$$

where the parameter τ is understood to be real. Since a is $+\infty$ on $\mathcal{P}_n \backslash \mathcal{M}_n$, we have

$$\mathrm{dom}\,(da(p)) = \{p \mid da(p) < +\infty\} \subset \mathcal{P}_{n-1}.$$

Hence, we restrict our attention to the behavior of $da(p)$ on the subspace $\mathcal{P}_{n-1}$.

We approach the problem of computing $da(p)$ from an epigraphical perspective. The epigraph of a is the set

$$\mathrm{epi}\,(a) = \{(p, \mu) \mid a(p) \leq \mu < +\infty\}.$$

Using this set, we can construct $da(p)$ from the formula

$$\mathrm{epi}\,(da(p)) = T_{\mathrm{epi}\,(a)}(p, a(p)) \tag{1.1}$$

[RW98, Theorem 8.2]. Here $T_{\mathrm{epi}\,(a)}(p, a(p))$ is the *tangent cone* to the set $\mathrm{epi}\,(a)$ at the point $(p, a(p))$. For a subset C of a finite dimensional linear space X, we have

$$T_C(x) = \left\{ d \;\middle|\; \begin{array}{c} \exists\ \{x^k\} \subset C \text{ and } \{t_k\} \subset \mathbb{R}_+ \text{ such that} \\ x^k \to x,\ t_k \searrow 0, \text{ and } t_k^{-1}(x^k - x) \to d \end{array} \right\} \tag{1.2}$$

$$= \left\{ \gamma d \;\middle|\; \begin{array}{c} \gamma \geq 0, \text{ and there exits } \{x^k\} \subset C \\ \text{with } x^k \to x \text{ such that } d = \displaystyle\lim_{k\to\infty} \frac{x^k - x}{\|x^k - x\|} \end{array} \right\}, \tag{1.3}$$

where $\mathbb{R}_+$ is the set of nonnegative real numbers and $\|\cdot\|$ is any norm on X. By considering $\mathcal{P}_{n-1}$ as a subspace of $\mathcal{P}_n$, we have

$$T_{\mathrm{epi}\,(a)}(p, \mu) \subset \mathcal{P}_{n-1} \times \mathbb{R} \qquad \text{for all } \mu \geq a(p), \tag{1.4}$$

since a is $+\infty$ on $\mathcal{P}_n \backslash \mathcal{M}_n$. In particular,

$$T_{\mathrm{epi}\,(a)}(p, \mu) = \mathcal{P}_{n-1} \times \mathbb{R} \qquad \text{whenever } \mu > a(p), \tag{1.5}$$

since a is continuous on $\mathcal{M}_n$.

In our first lemma we show that the tangential geometry of $\mathrm{epi}\,(a)$ remains essentially unchanged under the linear transformations corresponding to a uniform shift of the roots. For each $\lambda_0 \in \mathbb{C}^n$ define the linear transformation $H_{\lambda_0} : \mathcal{P}_n \to \mathcal{P}_n$ by

$$H_{\lambda_0}(p)(\lambda) = p(\lambda - \lambda_0).$$

Lemma 1.1. *Let λ_0 be a given complex number. Then*

$$T_{epi\,(a)}(H_{\lambda_0}(p), \eta + Re\,(\lambda_0)) = \{(H_{\lambda_0}(v), \mu) \,:\, (v, \mu) \in T_{epi\,(a)}(p, \eta)\}\ .$$

Proof. Define the affine transformation $\hat{H}_{\lambda_0} : \mathcal{P}_n \times \mathbb{R} \to \mathcal{P}_n \times \mathbb{R}$ by

$$\hat{H}_{\lambda_0}(p,\mu) = (H_{\lambda_0}(p), \mu + \operatorname{Re}(\lambda_0)).$$

Clearly, the mapping $\hat{H}_{\lambda_0}$ is invertible (indeed, $\hat{H}_{\lambda_0}^{-1} = \hat{H}_{-\lambda_0}$). In addition,

$$\hat{H}_{\lambda_0}^{-1}(\operatorname{epi}(a)) = \operatorname{epi}(a).$$

Therefore, by [RW98, Exercise 6.7] and the invertibility of $\hat{H}_{\lambda_0}$, we have

$$\begin{aligned} T_{\operatorname{epi}(a)}(H_{\lambda_0}(p), \mu + \operatorname{Re}(\lambda_0)) &= T_{\operatorname{epi}(a)}(\hat{H}_{\lambda_0}(p,\mu)) \\ &= \nabla \hat{H}_{\lambda_0}(p,\mu) T_{\operatorname{epi}(a)}(p,\mu) \\ &= \{(H_{\lambda_0}(v), \mu) \,:\, (v,\mu) \in T_{\operatorname{epi}(a)}(p,\eta)\}. \qquad \square \end{aligned}$$

We now derive a formula for the tangent cone to epi (a) at $(e_{(n,0)}, 0)$. All of our subsequent analysis relies on this key result. The proof is rather long and involved. It is based on an outline provided by Levantovskii [Lev80] for deriving a formula for the tangent cone to the set of stable polynomials.

THEOREM 1.2. *We have* $(v,\eta) \in T_{epi\,(a)}(e_{(n,0)}, 0)$, *with*

$$v = \beta_1 e_{(n-1,0)} + \beta_2 e_{(n-2,0)} + \cdots + \beta_n, \tag{1.6}$$

if and only if

$$Re\,\beta_1 \geq -n\eta, \tag{1.7}$$

$$Re\,\beta_2 \geq 0, \tag{1.8}$$

$$Im\,\beta_2 = 0, \text{ and} \tag{1.9}$$

$$\beta_k = 0 \text{ for } k = 3, \ldots, n. \tag{1.10}$$

Therefore, for $v \in \mathcal{P}_{n-1}$ *given by* (1.6), *we have*

$$da(e_{(n,0)})(v) = \begin{cases} -\frac{Re\,\beta_1}{n} & \text{if (1.8)–(1.10) hold, and} \\ +\infty & \text{otherwise.} \end{cases}$$

Proof. We begin by showing that (1.7)–(1.10) and (1.6) imply that (v,η) is an element of the tangent cone $T_{\operatorname{epi}(a)}(e_{(n,0)}, 0)$. This is done by constructing a curve in epi (a) converging to $(e_{(n,0)}, 0)$ and having derivative equal to (v,η). Consider the polynomials having coefficients that are polynomials in ξ and given by

$$\begin{aligned} p(\lambda,\xi) &= \left(\lambda + \frac{\beta_1}{n}\xi\right)^{n-2} \left(\lambda + \sqrt{-1}(\beta_2\xi)^{\frac{1}{2}} + \frac{\beta_1}{n}\xi\right) \left(\lambda - \sqrt{-1}(\beta_2\xi)^{\frac{1}{2}} + \frac{\beta_1}{n}\xi\right) \\ &= \left(\lambda^{n-2} + (n-2)\frac{\beta_1}{n}\xi\lambda^{n-3} + o(\xi)\right) \left(\lambda^2 + 2\frac{\beta_1}{n}\xi\lambda + \beta_2\xi + o(\xi)\right) \\ &= \lambda^n + \beta_1\xi\lambda^{n-1} + \beta_2\xi\lambda^{n-2} + o(\xi) \\ &= \lambda^n + \xi v(\lambda) + o(\xi)\ . \end{aligned}$$

Let ξ be real and positive. Then $a(p(\lambda,\xi)) = -\frac{\operatorname{Re}(\beta_1)}{n}\xi$. Therefore,

$$\lim_{\xi \searrow 0} \frac{a(p(\lambda,\xi)) - a(\lambda^n)}{\xi} = -\frac{\operatorname{Re}(\beta_1)}{n} \leq \eta,$$

which yields the result.

We now show that any element (v,η) in the tangent cone $T_{\mathrm{epi}\,(a)}(e_{(n,0)},0)$ must satisfy (1.7)–(1.10) if v is given the representation (1.6). To this end, we make use of the following norm on $\mathcal{P}_n \times \mathbb{R}$:

$$\left\|(b_0 e_{(n,0)} + b_1 e_{(n-1,0)} + \cdots + b_n, \mu)\right\| = \max\{\,|b_0|\,,\,|b_1|\,,\ldots,\,|b_n|\,,\,|\mu|\,\}\ .$$

Let $(v,\eta) \in T_{\mathrm{epi}\,(a)}(e_{(n,0)},0)$ with v written as in (1.6). By definition there is a sequence $\{(p_k,\mu_k)\} \in \mathrm{epi}\,(a)$ with $(p_k,\mu_k) \to (e_{(n,0)},0)$ and

$$\text{(1.11)} \qquad \frac{((p_k,\mu_k) - (e_{(n,0)},0))}{\left\|(p_k,\mu_k) - (e_{(n,0)},0)\right\|} \to (\gamma v, \gamma\eta)$$

for some $\gamma > 0$.

Given $\epsilon \in \mathbb{C}^n$, define $\sigma_j\ :\ \mathbb{C}^n \to \mathbb{C}$ for $j = 1,2,\ldots,n$ to be the symmetric functions

$$\text{(1.12)}\,\sigma_1(\epsilon) = \sum_{t=1}^{n} \epsilon_t \quad \text{and} \quad \sigma_j(\epsilon) = \sum_{1\le t_1<t_2<\cdots<t_j\le n} \left(\prod_{s=1}^{j} \epsilon_{t_s}\right) \text{ for } j = 2,\ldots,n,$$

and set $\sigma = (\sigma_1,\sigma_2,\cdots,\sigma_n)^T$. For each $k = 1,2,\ldots$ there exist complex numbers $\epsilon^k = (\epsilon_{k1},\ \epsilon_{k2},\ldots,\epsilon_{kn})^T \to 0$ such that $\mathrm{Re}\,(\epsilon_{kj}) \ge -\mu_k$ for $j = 1,2,\ldots,n$ and

$$\text{(1.13)} \qquad p_k(\lambda) = \prod_{j=1}^{n}(\lambda + \epsilon_{kj}) = (\lambda^n + \sigma_1(\epsilon^k)\lambda^{n-1} + \cdots + \sigma_n(\epsilon^k))\ .$$

For each $k = 1,2,\ldots$, set

$$\nu_k = \left\|(p_k,\mu_k) - (e_{(n,0)},0)\right\| = \max\{(\left\|\sigma(\epsilon^k)\right\|_\infty,\,|\mu_k|\,\}.$$

Then the limit (1.11) can be interpreted componentwise as

$$\gamma\beta_j = \lim_{k\to\infty} \frac{\sigma_j(\epsilon^k)}{\nu_k}$$

for $j = 1,2,\ldots,n$. Set $\tilde\sigma_j = \gamma\beta_j$ for $j = 1,2,\ldots,n$. We establish the result by showing that

$$\text{(1.14)} \quad \mathrm{Re}\,\tilde\sigma_1 \ge -n\gamma\eta,\ \ \mathrm{Re}\,\tilde\sigma_2 \ge 0,\ \ \mathrm{Im}\,\tilde\sigma_2 = 0,\ \text{ and } \tilde\sigma_k = 0 \text{ for } k = 3,4,\ldots,n.$$

Clearly, $\mathrm{Re}\,(\tilde\sigma_1) \ge -n\gamma\eta$ since $\mathrm{Re}\,(\sigma_1(\epsilon^k)) = \sum_{j=1}^{n} \mathrm{Re}\,(\epsilon_{kj}) \ge -n\mu_k$ for all $k = 1,2,\ldots$ and $\mu_k/\nu_k \to \gamma\eta$. We now show that $\tilde\sigma_j = 0$ for $j = 3,4,\ldots,n$. First note that

$$\text{(1.15)} \qquad \sigma_j(\epsilon) = o(\|\epsilon\|_\infty^2) \quad \text{for } j = 3,4,\ldots,n.$$

Define

$$\text{(1.16)} \qquad \alpha_{kj} = \mathrm{Re}\,\epsilon_{kj} \quad \text{and} \quad \delta_{kj} = \mathrm{Im}\,\epsilon_{kj}$$

for $j = 1,2,\ldots,n$ and $k = 1,2,\ldots$. Note that $\alpha_{kj} \ge -\mu_k$ for $j = 1,2,\ldots,n$ and $k = 1,2,\ldots$. In addition, it is easily verified that

$$\mathrm{Re}\,\sigma_2(\epsilon^k) = \sum_{s<t}[\alpha_{ks}\alpha_{kt} - \delta_{ks}\delta_{kt}] \text{ and } \mathrm{Im}\,\sigma_2(\epsilon^k) = \left[\sum_{s<t}\alpha_{ks}\delta_{kt} + \sum_{s<t}\delta_{ks}\alpha_{kt}\right].$$

Then, by definition,

$$\begin{aligned}\left|\sigma_1(\epsilon^k)\right|^2 &= \left\|\epsilon^k\right\|_2^2 + 2\sum_{s<t} \operatorname{Re}(\bar{\epsilon}_{ks}\epsilon_{kt}) \\ &= \left\|\epsilon^k\right\|_2^2 + 2\sum_{s<t} \alpha_{ks}\alpha_{kt} + 2\sum_{s<t} \delta_{ks}\delta_{kt} \\ &= \left\|\epsilon^k\right\|_2^2 + 4\sum_{s<t} \alpha_{ks}\alpha_{kt} + 2\sum_{s<t}[\delta_{ks}\delta_{kt} - \alpha_{ks}\alpha_{kt}] \\ &= \left\|\epsilon^k\right\|_2^2 + 4\sum_{s<t} \alpha_{ks}\alpha_{kt} - 2\operatorname{Re}\sigma_2(\epsilon^k) \\ &\geq \left\|\epsilon^k\right\|_\infty^2 - 2n(n-1)\mu_k^2 - 2\operatorname{Re}(\sigma_2(\epsilon^k)) \\ &\geq \left\|\epsilon^k\right\|_\infty^2 - 4n(n-1)\max\{\,|\mu_k|\,,\,\left|\sigma_2(\epsilon^k)\right|\,\}\,,\end{aligned}$$

whenever $|\mu_k| \leq 1$. Hence, if ϵ^k and μ_k are such that $\left|\sigma_1(\epsilon^k)\right| < \frac{\|\epsilon^k\|_\infty}{2}$ and $|\mu_k| \leq 1$, then, for $\Delta = \frac{3}{16n^2}$, we have

$$\max\{\,|\mu_k|\,,\,\left|\sigma_2(\epsilon^k)\right|\,\} > \Delta\left\|\epsilon^k\right\|_\infty^2.$$

On the other hand, if $\left|\sigma_1(\epsilon^k)\right| \geq \frac{\|\epsilon^k\|_\infty}{2}$ and $\left\|\epsilon^k\right\|_\infty \leq 1$, then $\left|\sigma_1(\epsilon^k)\right| \geq \frac{\|\epsilon^k\|_\infty^2}{4}$. Thus, in either case, we have

$$\max(\,|\mu_k|\,,\,\left|\sigma_1(\epsilon^k)\right|,\,\left|\sigma_2(\epsilon^k)\right|\,) \geq \Delta\left\|\epsilon^k\right\|_\infty^2, \tag{1.17}$$

whenever $\left\|\epsilon^k\right\|_\infty \leq 1$ and $|\mu_k| \leq 1$. This implies that

$$\nu_k \geq \Delta\left\|\epsilon^k\right\|_\infty^2 \tag{1.18}$$

for all k sufficiently large. This bound, in conjunction with (1.15), allows us to conclude that

$$\tilde{\sigma}_j = \lim_{k\to\infty}\frac{\sigma_j(\epsilon^k)}{\nu_k} = 0 \qquad \text{for } j = 3, 4, \ldots, n.$$

We now turn our attention to the coefficient $\tilde{\sigma}_2$. If

$$\max\{\,\left|\sigma_1(\epsilon^k)\right|,\,\left|\sigma_2(\epsilon^k)\right|\,\} = o(\nu_k),$$

we are done since then $\tilde{\sigma} = 0$. Hence, we assume that

$$\max\{\,\left|\sigma_1(\epsilon^k)\right|,\,\left|\sigma_2(\epsilon^k)\right|\,\} \neq o(\nu_k)$$

so that

$$\nu_k = \max\{\,\left|\sigma_1(\epsilon^k)\right|,\,\left|\sigma_2(\epsilon^k))\right|,\,|\mu_k|\,\}$$

for all k sufficiently large. Set $\tilde{\nu}_{kj} = \max\{\,\left|\sigma_j(\epsilon^k)\right|,\,|\mu_k|\,\}$ for $j = 1, 2$. Observe that if $\lim_{k\to\infty}\frac{\sigma_2(\epsilon^k)}{\tilde{\nu}_{k1}} = 0$, then we are done since in this case $\nu_k = \tilde{\nu}_{k1}$ for all k sufficiently large which implies that $\tilde{\sigma}_2 = 0$. Hence, with no loss in generality, we can assume that there is a constant $c > 0$ such that

$$\left|\sigma_2(\epsilon^k)\right| \geq c\tilde{\nu}_{k1} \quad \text{for all } k = 1, 2, \ldots. \tag{1.19}$$

Therefore, there is a constant $K > 0$ such that

$$K\left|\sigma_2(\epsilon^k)\right| \geq \nu_k \quad \text{for all } k \text{ sufficiently large.} \tag{1.20}$$

Now observe that

$$|\sigma_2(\epsilon)| = \left|\sum_{s<t} \epsilon_s \epsilon_t\right| \leq \sum_{s<t} |\epsilon_s|\,|\epsilon_t| \leq \frac{n(n-1)}{2}\,\|\epsilon\|_\infty^2 . \tag{1.21}$$

Therefore, for all k sufficiently large,

$$c\left|\operatorname{Re}(\sigma_1(\epsilon^k))\right| \leq c\left|\sigma_1(\epsilon^k)\right| \leq c\tilde{\nu}_{k1} \leq \left|\sigma_2(\epsilon^k)\right| \leq \frac{n(n-1)}{2}\left\|\epsilon^k\right\|_\infty^2 ,$$

and so, from (1.20), we have

$$|\mu_k| \leq \nu_k \leq \frac{Kn(n-1)}{2}\left\|\epsilon^k\right\|_\infty^2 . \tag{1.22}$$

In particular, this implies that

$$\frac{\mu_k}{\|\epsilon^k\|_\infty} \to 0.$$

In addition, since $\alpha_{kj} + \mu_k \geq 0$ for each $j = 1, 2, \ldots, n$ and all $k = 1, 2, \ldots$ and

$$0 \leq \sum_{j=1}^{n} \frac{\alpha_{kj} + \mu_k}{\|\epsilon^k\|_\infty} \leq \frac{\left|\operatorname{Re}(\sigma_1(\epsilon^k))\right| + n\,|\mu_k|}{\|\epsilon^k\|_\infty} \leq \frac{n(n-1)}{2}\left(\frac{1}{c} + Kn\right)\left\|\epsilon^k\right\|_\infty ,$$

for all $k = 1, 2, \ldots$ (recall the definition of the α_{kj}'s from (1.16)), we obtain

$$\lim_{k\to\infty} \frac{\alpha_{kj}}{\|\epsilon^k\|_\infty} = 0 \quad \text{for } j = 1, 2, \ldots, n. \tag{1.23}$$

Putting together the bounds (1.18), (1.20), and (1.22), we obtain the relation

$$\Delta\left\|\epsilon^k\right\|_\infty^2 \leq \nu_k \leq K\left|\sigma_2(\epsilon^k)\right| \leq K\frac{n(n-1)}{2}\left\|\epsilon^k\right\|_\infty^2 , \tag{1.24}$$

for all $k = 1, 2, \ldots$. In addition, the bound (1.19) implies that

$$\frac{\left|\operatorname{Im}(\sigma_1(\epsilon^k))\right|^2}{|\sigma_2(\epsilon^k)|} \leq \frac{\left|\sigma_1(\epsilon^k)\right|^2}{|\sigma_2(\epsilon^k)|} \leq \frac{1}{c^2}\left|\sigma_2(\epsilon^k)\right|$$

so that

$$\frac{\left|\operatorname{Im}(\sigma_1(\epsilon^k))\right|^2}{|\sigma_2(\epsilon^k)|} \to 0.$$

Now since $\left|\operatorname{Im}(\sigma_1(\epsilon^k))\right|^2 = \sum_{j=1}^{n} \delta_{kj}^2 + 2\sum_{s<t} \delta_{ks}\delta_{kt}$, this implies that

$$\lim_{k\to\infty} \sum_{s<t} \frac{\delta_{ks}\delta_{kt}}{|\sigma_2(\epsilon^k)|} \leq 0. \tag{1.25}$$

Finally, recall that

$$\sigma_2(\epsilon^k) = \left[\sum_{s<t} \alpha_{ks}\alpha_{kt} - \sum_{s<t} \delta_{ks}\delta_{kt}\right] + i\left[\sum_{s<t} \alpha_{ks}\delta_{kt} + \sum_{s<t} \delta_{ks}\alpha_{kt}\right].$$

Therefore, by (1.24), (1.25), and (1.23), we see that

$$\mathrm{Re}\,(\tilde{\sigma}_2) = \lim_{k\to\infty} \frac{\mathrm{Re}\,(\sigma_2(\epsilon^k))}{\nu_k} \geq 0.$$

Similarly, from (1.24) and (1.23), we have

$$\begin{aligned} |\,\mathrm{Im}\,(\tilde{\sigma}_2)| &= \lim_{k\to\infty} \frac{\left|\,\mathrm{Im}\,(\sigma_2(\epsilon^k))\right|}{\nu_k} \\ &\leq \Delta^{-1} \lim_{k\to\infty} \sum_{s<t} \left(\frac{|\alpha_{ks}|}{\|\epsilon^k\|_\infty} \frac{|\delta_{kt}|}{\|\epsilon^k\|_\infty} + \frac{|\alpha_{kt}|}{\|\epsilon^k\|_\infty} \frac{|\delta_{ks}|}{\|\epsilon^k\|_\infty} \right) \\ &= 0, \end{aligned}$$

since $\frac{|\delta_{kj}|}{\|\epsilon^k\|_\infty} \leq 1$ for all $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots$.

The final statement of the theorem concerning the formula for $da(e_{(n,0)})(v)$ now follows immediately from the equivalence of (1.1) and (1.7)–(1.10). □

By combining Lemma 1.1 with Theorem 1.2, we obtain the following corollary.

COROLLARY 1.3. *Given $\lambda_0 \in \mathbb{C}$, we have $(v, \eta) \in T_{epi\,(a)}(e_{(n,\lambda_0)}, Re\,(\lambda_0))$, with*

$$v = \beta_1 e_{(n-1,\lambda_0)} + \beta_2 e_{(n-2,\lambda_0)} + \cdots + \beta_n\ , \tag{1.26}$$

if and only if $\beta_1, \beta_2, \ldots, \beta_n$ satisfy the conditions (1.7)–(1.10). *Therefore, for $v \in \mathcal{P}_{n-1}$ given by* (1.26), *we have*

$$da(e_{(n,\lambda_0)})(v) = \begin{cases} -\frac{Re\,\beta_1}{n} & \textit{if } (1.8)\text{–}(1.10) \textit{ hold,} \\ +\infty & \textit{otherwise.} \end{cases}$$

We now show that the factorization of a polynomial into powers of linear factors (or the *prime factorization*) can be used to obtain a description of the tangent cone to the epigraph of a from Corollary 1.3. We begin by developing a tool that allows us to treat each of the linear factors in the prime factorization separately. We then glue the results for each of the factors back together to obtain a result for the polynomial as a whole. This tool is provided in the next lemma which establishes a local property for factorizations into relatively prime factors.

LEMMA 1.4. *Let $(n_1, n_2, \ldots, n_m)$ be a partition of n, that is, for $j = 1, 2, \ldots, m$ each n_j is a positive integer and $n = \sum_{j=1}^m n_j$. Set*

$$\mathcal{S} = \mathbb{C} \times \mathcal{P}_{n_1-1} \times \mathcal{P}_{n_2-1} \times \cdots \times \mathcal{P}_{n_m-1}$$

and let $p_j \in \mathcal{M}_{n_j}$ for $j = 1, 2, \ldots, m$. Consider the mapping $F : \mathcal{S} \to \mathcal{P}_n$ given by

$$F(v_0, v_1, v_2, \ldots, v_m) = (1 + v_0) \prod_{j=1}^m (p_j + v_j)\ .$$

If the polynomials $p_1, \ldots, p_m$ are relatively prime(i.e., have no common roots), then there exist open neighborhoods U of $0 \in \mathcal{S}$ and W of $F(0) \in \mathcal{P}_n$ such that F is *a homeomorphism between U and W with $\nabla(F^{-1})$ existing, continuous on W, and satisfying $\nabla(F^{-1})(F(u)) = [\nabla F(u)]^{-1}$ for all $u \in U$. Thus, in particular, we have Ran $(\nabla F(0)) = \mathcal{P}_n$; that is, every polynomial $h \in \mathcal{P}_n$ can be written as*

$$h = \nabla F(0)(w_0, w_1, \ldots, w_m) = \sum_{j=0}^{m} r_j w_j \ , \tag{1.27}$$

for some $(w_0, w_1, \ldots, w_m) \in \mathcal{S}$, where

$$r_0 = \prod_{j=1}^{m} p_j \quad \textit{and} \quad r_s = \prod_{j \neq s} p_j \quad \textit{for } s = 1, 2, \ldots, m. \tag{1.28}$$

Proof. Since $\dim(\mathcal{S}) = n + 1 = \dim(\mathcal{P}_n)$, the result follows from the classical inverse function theorem once it is shown that $\ker(\nabla F(0)) = \{0\}$. Let $\mathcal{Z}_j$ denote the set of zeros of the polynomial p_j counting multiplicity, for $j = 1, 2, \ldots, m$, and let $(w_0, w_1, \ldots, w_m) \in \ker(\nabla F(0))$. Since the polynomials $p_1, p_2, \ldots, p_m$ are relatively prime, we have $\mathcal{Z}_j \cap \mathcal{Z}_s = \emptyset$ whenever $j \neq s$. Equations (1.27) and (1.28) and the inclusion $(w_0, w_1, \ldots, w_m) \in \ker(\nabla F(0))$ imply that for each $s = 1, 2, \ldots, m$ the polynomial

$$f_s = r_s w_s$$

has zeros not only at the points $\cup_{j \neq s} \mathcal{Z}_j$ (with the corresponding multiplicities) but also at the points $\mathcal{Z}_s$ (with the corresponding multiplicities). Hence, each f_s is either the zero polynomial or its degree is at least n. However, the degree of each f_s is at most $n-1$, since $w_s \in \mathcal{P}_{n_s - 1}$. Therefore, f_s is the zero polynomial for $s = 1, 2, \ldots, m$. This in turn implies that $w_j = 0$ for $j = 1, 2, \ldots, m$, and finally that $w_0 = 0$. Consequently, $\ker(\nabla F(0)) = \{0\}$. □

As a first application of Lemma 1.4, we show if a polynomial is written as a product of relatively prime factors, then the tangent cone to the epigraph of a at this polynomial is contained within a kind of product of the tangent cones associated with each of the relatively prime factors.

THEOREM 1.5. *Let $(n_1, n_2, \ldots, n_m)$ be a partition of n, and let $p_j \in \mathcal{M}_{n_j}$ for $j = 1, 2, \ldots, m$ be relatively prime. Set $p = \prod_{j=1}^{m} p_j \in \mathcal{M}_n$. Let the space $\mathcal{S}$ and the function $F : \mathcal{S} \to \mathcal{P}_n$ be as given in Lemma* 1.4. *If $(h, \omega) \in T_{epi\,(a)}(p, a(p))$, then there exists $(0, w_1, w_2, \ldots, w_m) \in \mathcal{S}$ such that h is given by* (1.27) *and* (1.28)*, where, for $j = 1, \ldots, m$, $(w_j, \omega) \in T_{epi\,(a_{n_j})}(p_j, a(p))$ and a_{n_j} denotes the abscissa mapping on $\mathcal{P}_{n_j}$.*

Proof. Let (h, ω) be a nonzero element of the tangent cone $T_{\text{epi}\,(a)}(p, a(p))$. Then there is a sequence $\{(q_k, \alpha_k)\} \subset \text{epi}\,(a) \subset \mathcal{M}_n \times \mathbb{R}$ and a scalar $\gamma > 0$ such that

$$(q_k, \alpha_k) \to (p, a(p)) \quad \text{and} \quad \frac{(q_k, \alpha_k) - (p, a(p))}{\|(q_k, \alpha_k) - (p, a(p))\|} \to (\gamma h, \gamma \omega) \ .$$

Let $F : \mathcal{S} \to \mathcal{P}_n$ be as in Lemma 1.4. Then, by trimming finitely many terms from the beginning of the sequence if necessary so that q_k is sufficiently close to p, Lemma 1.4 yields the existence of a sequence $\{(0, v_{k1}, v_{k2}, \ldots, v_{km})\} \subset \mathcal{S}$ such that $(0, v_{k1}, v_{k2}, \ldots, v_{km}) \to 0$ and

$$q_k = F(0, v_{k1}, v_{k2}, \ldots, v_{km}) = \prod_{j=1}^{m} (p_j + v_{kj}) \quad \text{for all } k = 1, 2, \ldots \ ,$$

since $\{q_k\} \subset \mathcal{M}_n = \mathrm{dom}\,(a)$. Since $(q_k, \alpha_k) \in \mathrm{epi}\,(a)$, we have

$$(p_j + v_{kj}, \alpha_k) \in \mathrm{epi}\,(a_{n_j}) \quad \text{for all } j = 1, 2, \ldots, m \text{ and } k = 1, 2, \ldots \tag{1.29}$$

and

$$(p_j + v_{kj}, \alpha_k) \to (p_j, a(p)) \quad \text{for all } j = 1, 2, \ldots, m. \tag{1.30}$$

Set $v^k = (0, v_{k1}, v_{k2}, \ldots, v_{km})$ for $k = 1, 2, \ldots$, and set $\bar{v} = 0$ so that $v^k \to \bar{v}$. Then

$$\begin{aligned} q_k - p &= F(v^k) - F(\bar{v}) \\ &= \nabla F(\bar{v})(v^k - \bar{v}) + o(\|v^k - \bar{v}\|) \, . \end{aligned} \tag{1.31}$$

By Lemma 1.4, $\nabla(F^{-1})$ is continuous in a neighborhood of p so that F^{-1} is Lipschitz continuous near p. Consequently, there is a constant $K > 0$ such that $\|v^k - \bar{v}\| \leq K \|q_k - p\|$ for all $k = 1, 2, \ldots$. This fact, combined with (1.31), yields

$$\begin{aligned} \gamma h &= \lim_{k \to \infty} \frac{q_k - p}{\|(q_k, \alpha_k) - (p, a(p))\|} \\ &= \nabla F(\bar{v}) \lim_{k \to \infty} \frac{v^k - \bar{v}}{\|(q_k, \alpha_k) - (p, a(p))\|} \\ &= \nabla F(0)(0, \hat{w}_1, \hat{w}_2, \ldots, \hat{w}_m), \end{aligned} \tag{1.32}$$

where

$$\hat{w}_j = \lim_{k \to \infty} \frac{v_{kj}}{\|(q_k, \alpha_k) - (p, a(p))\|} \quad \text{for } j = 1, 2, \ldots, m.$$

Equation (1.32) verifies (1.27) with $w_j = \gamma^{-1}\hat{w}_j$ for $j = 1, 2, \ldots, m$. From (1.29), (1.30), and definition (1.2) (here $t_k = \|(q_k, \alpha_k) - (p, a(p))\|$), we have that (w_j, ω) is an element of $T_{\mathrm{epi}\,(a_{n_j})}(p_j, a(p))$, for $j = 1, 2, \ldots, m$, which proves the result. □

We now apply Corollary 1.3, Lemma 1.4, and Theorem 1.5 to obtain a complete representation of the tangent cone to the epigraph of the abscissa mapping at an arbitrary polynomial. This representation involves the prime factorization of the polynomial. For this purpose, and for the application of this result in later sections, it is useful to introduce some more notation.

Let $p \in \mathcal{M}_n$ have prime factorization

$$p = \prod_{j=1}^{m} e_{(n_j, \lambda_j)}, \tag{1.33}$$

where $\lambda_1, \ldots, \lambda_m$ are distinct complex numbers and $(n_1, n_2, \ldots, n_m)$ is a partition of n. Define $\mathcal{S}_p$ to be the product space

$$\mathcal{S}_p = \mathbb{C} \times \mathcal{P}_{n_1 - 1} \times \mathcal{P}_{n_2 - 1} \times \cdots \times \mathcal{P}_{n_m - 1}. \tag{1.34}$$

In conjunction with $\mathcal{S}_p$, we define the mapping $F_p : \mathcal{S}_p \to \mathcal{P}_n$ by

$$F_p(v_0, v_1, \ldots, v_m) = (1 + v_0) \prod_{j=1}^{m} (e_{(n_j, \lambda_j)} + v_j) \quad \text{for all } (v_0, v_1, \ldots, v_m) \in \mathcal{S}_p, \tag{1.35}$$

so that $F_p(0) = p$. By analogy with (1.27), for every $(w_0, w_1, \ldots, w_m) \in \mathcal{S}$, we have

$$\nabla F_p(0)(w_0, w_1, \ldots, w_m) = \sum_{j=0}^{m} r_j w_j \ , \tag{1.36}$$

where

$$r_0 = p \quad \text{and} \quad r_s = \prod_{j \neq s} e_{(n_j, \lambda_j)} \quad \text{for } s = 1, 2, \ldots, m. \tag{1.37}$$

In addition, we define

$$\mathcal{I}(p) = \{ j \in \{1, 2, \ldots, m\} \mid a(p) = \operatorname{Re} \lambda_j \} , \tag{1.38}$$

the set of indices of active roots of p.

We now state and prove the main result of this section.

THEOREM 1.6. *Let $p \in \mathcal{M}_n$ have factorization* (1.33). *Then (h, ω) is an element of $T_{epi\,(a)}(p, a(p))$ if and only if there exists a vector $(w_0, w_1, w_2, \ldots, w_m) \in \mathcal{S}_p$ such that*

$$h = \nabla F_p(0)(w_0, w_1, w_2, \ldots, w_m), \tag{1.39}$$

where $\nabla F_p(0)$ is defined in (1.36)–(1.37),

$$w_0 = 0, \tag{1.40}$$

and

$$(w_j, \omega) \in T_{epi\,(a_{n_j})}(e_{(n_j, \lambda_j)}, a(p)) \quad \textit{for } j = 1, 2, \ldots, m. \tag{1.41}$$

In addition, if for each $j = 1, 2, \ldots, m$, w_j is given the representation

$$w_j = \beta_{j1} e_{(n_j - 1, \lambda_j)} + \beta_{j2} e_{(n_j - 2, \lambda_j)} + \cdots + \beta_{j n_j} \ , \tag{1.42}$$

then, for each $j \in \mathcal{I}(p)$, a necessary and sufficient condition for (1.41) *to hold is that*

$$Re\,\beta_{j1} \geq -n_j \omega, \tag{1.43}$$

$$Re\,\beta_{j2} \geq 0, \tag{1.44}$$

$$Im\,\beta_{j2} = 0, \ \textit{and} \tag{1.45}$$

$$\beta_{js} = 0 \textit{ for } s = 3, 4, \ldots, n_j \ . \tag{1.46}$$

Proof. Let us first assume that $(h, \omega) \in T_{\mathrm{epi}\,(a)}(p, a(p))$ and show that (h, ω) must satisfy (1.39), (1.40), and (1.41). By Lemma 1.4, there must exist a vector $(w_0, w_1, w_2, \ldots, w_m)$ in $\mathcal{S}_p$ such that (1.39) holds. The fact that $w_0 = 0$ follows from (1.4), while (1.41) follows immediately from Theorem 1.5. The conditions (1.43)–(1.46) follow from (1.41) and Corollary 1.3.

Next, let us assume that $(h, \omega) \in \mathcal{P}_{n-1} \times \mathbb{R}$ satisfies (1.39), (1.40), and (1.41). We need to show that $(h, \omega) \in T_{\mathrm{epi}\,(a)}(p, a(p))$. We accomplish this by following the approach taken in Theorem 1.2. That is, we will exhibit a curve in epi (a) passing through $(p, a(p))$ and having the tangent direction (h, ω) at $(p, a(p))$. For $j = 1, 2, \ldots, m$, give each w_j in (1.39) the representation (1.42). Then, by Corollary 1.3, we know that the conditions (1.43)–(1.46) are satisfied for $j \in \mathcal{I}(p)$. For each such $j \in \mathcal{I}(p)$, define

$$\begin{aligned} p_j(\lambda, \xi) &= \left((\lambda - \lambda_j) + \frac{\beta_{j1}}{n_j}\xi \right)^{n_j - 2} \\ &\quad \times \left((\lambda - \lambda_j) + \sqrt{-1}(\beta_{j2}\xi)^{\frac{1}{2}} + \frac{\beta_{j1}}{n_j}\xi \right) \left((\lambda - \lambda_j) - \sqrt{-1}(\beta_{j2}\xi)^{\frac{1}{2}} + \frac{\beta_{j1}}{n_j}\xi \right) \\ &= (\lambda - \lambda_j)^{n_j} + (\beta_{j1}\xi)(\lambda - \lambda_j)^{n_j - 1} + \beta_{j2}\xi(\lambda - \lambda_j)^{n_j - 2} + o(\xi) \\ &= (\lambda - \lambda_j)^{n_j} + \xi w_j(\lambda) + o(\xi), \end{aligned} \tag{1.47}$$

and, for $j \in \{1, 2, \ldots, m\} \backslash \mathcal{I}(p)$, define

$$p_j(\lambda, \xi) = (\lambda - \lambda_j)^{n_j} + \xi w_j(\lambda) \ . \tag{1.48}$$

Set $p(\lambda, \xi) = \prod_{j=1}^{m} p_j(\lambda, \xi)$, so that, from (1.36), (1.37), and (1.39),

$$\begin{aligned} p(\lambda, \xi) &= p(\lambda) + \xi \sum_{j=1}^{m} r_j(\lambda) w_j(\lambda) + o(\xi) \\ &= p(\lambda) + \xi \nabla F_p(0)(0, w_1, \ldots, w_m)(\lambda) + o(\xi) \\ &= p(\lambda) + \xi h(\lambda) + o(\xi). \end{aligned}$$

Then, for all ξ small, positive, and real,

$$\begin{aligned} a(p(\lambda, \xi)) &= \max_{j \in \mathcal{I}(p)} \operatorname{Re} \left(\lambda_j - \frac{\beta_{j1}}{n_j}\xi \right) \\ &\leq a(p) + \xi\omega \end{aligned}$$

so that $(p(\lambda, \xi), a(p) + \xi\omega) \in \operatorname{epi}(a)$ for all ξ small, positive, and real. Therefore, since

$$\lim_{\xi \searrow 0} \frac{(p(\lambda, \xi), a(p) + \xi\omega) - (p(\lambda), a(p))}{\xi} = (h(\lambda), \omega),$$

we have $(h, \omega) \in T_{\operatorname{epi}(a)}(p, a(p))$, completing the proof. □

COROLLARY 1.7. *Let $p \in \mathcal{M}_n$ have factorization* (1.33) *and let $h \in \mathcal{P}_n$. By Lemma* 1.4, *there exists $(w_0, w_1, w_2, \ldots, w_m)$ in $\mathcal{S}_p$ such that* (1.39) *holds, where, for each $j = 1, 2, \ldots, m$, w_j can be written as in* (1.42). *With this representation for h, either $w_0 = 0$ and* (1.44)–(1.46) *hold for $j \in \mathcal{I}(p)$, in which case*

$$da(p)(h) = \max_{j \in \mathcal{I}(p)} \frac{-Re(\beta_{j1})}{n_j} \ ,$$

or

$$da(p)(h) = +\infty.$$

Proof. By Theorem 1.6, we know that $da(p)(h) = +\infty$ if either $w_0 \neq 0$ or the coefficients β_{js}, $s = 1, 2, \ldots, n_j$, do not satisfy one of the conditions in (1.44)–(1.46) for every $j \in \mathcal{I}(p)$. On the other hand, if $w_0 = 0$ and all of the conditions in (1.44)–(1.46) are satisfied for every $j \in \mathcal{I}(p)$, then the inequality (1.43) in Theorem 1.6 implies that $(h, \omega) \in T_{\operatorname{epi}(a)}((p, a(p)))$ if and only if $\omega \geq \frac{-\operatorname{Re}(\beta_{j1})}{n_j}$ for every $j \in \mathcal{I}(p)$. Since $T_{\operatorname{epi}(a)}((p, a(p))) = \operatorname{epi}(da(p))$ [RW98, Theorem 8.2], this proves the corollary. □

2. Regular subgradients and the normal cone. We now turn our attention to the variational objects *dual* to the subderivative and the tangent cone. These are the subgradients and the normal cone. These objects are defined in terms of a duality pairing between the linear space $\mathcal{P}_n$ and its *dual space.* Traditionally the *dual space* is the space of continuous linear functionals on the *primal* space (which in our setting is $\mathcal{P}_n$). The *duality pairing* is then the continuous bilinear functional obtained by evaluating a given linear functional at a given point. However, in general, the dual space may have many possible representations and for each representation there may be any number of bilinear functionals that *pair the spaces in duality.*

In our analysis, we have chosen to regard $\mathcal{P}_n$ as a Hilbert space, in which case the dual of $\mathcal{P}_n$ is itself. However, we will need to consider a whole family of duality pairings, or inner products, on $\mathcal{P}_n$. To describe this family of inner products, recall that for each $\lambda_0 \in \mathbb{C}$, the polynomials

$$e_{(j,\lambda_0)},\ j = 0, 1, \ldots, n, \tag{2.1}$$

form a basis for $\mathcal{P}_n$. Hence, for each $\lambda_0 \in \mathbb{C}$, we can define a real inner product on $\mathcal{P}_n$ associated with the representation in this basis. Given $p = \sum_{j=1}^n a_j e_{(n-j,\lambda_0)}$ and $q = \sum_{j=1}^n b_j e_{(n-j,\lambda_0)}$, define the inner product

$$\langle \cdot\, , \cdot \rangle_{(n,\lambda_0)} : \mathcal{P}_n \times \mathcal{P}_n \to \mathbb{R}$$

by

$$\langle p\, , q \rangle_{(n,\lambda_0)} = \operatorname{Re} \sum_{j=0}^{n} \bar{a}_j b_j. \tag{2.2}$$

Thus, in the case $n = 0$, we recover the real inner product on $\mathbb{C}$. Note that this family of inner products behaves continuously in p, q, and λ_0 in the sense that the mapping

$$(p, q, \lambda) \mapsto \langle p\, , q \rangle_{(n,\lambda)} \tag{2.3}$$

is continuous on $\mathcal{P}_n \times \mathcal{P}_n \times \mathbb{C}$. To see this, note that the expansions of the polynomials p and q in the basis (2.1) are just their Taylor series expansions at λ_0, hence,

$$\langle p\, , q \rangle_{(n,\lambda)} = \operatorname{Re} \left(\sum_{j=0}^{n} \frac{\overline{p^{(j)}(\lambda)}}{j!} \frac{q^{(j)}(\lambda)}{j!} \right),$$

where $f^{(j)}$ denotes the jth derivative of the function f.

By setting $\lambda_0 = 0$ in (2.1), we obtain the *standard basis* for $\mathcal{P}_n$. The inner product (2.2) associated with the standard basis is simply written $\langle \cdot\, , \cdot \rangle$.

The spaces $\mathcal{S}_p$ defined in (1.34) also play a key role in our analysis; therefore, we need an inner product on these spaces as well. We use the inner product

$$\langle (u_0, u_1, \ldots, u_m)\, , (v_0, v_1, \ldots, v_m) \rangle_p = \sum_{s=0}^{m} \langle u_s\, , v_s \rangle_{(n_s - 1, \lambda_s)}\ , \tag{2.4}$$

for every $(u_0, u_1, \ldots, u_m)$ and $(v_0, v_1, \ldots, v_m)$ in $\mathcal{S}_p$, where we define $n_0 = 1$ in this expression and hereafter.

Spaces paired in duality give rise to the notion of the adjoint of a linear transformation. Suppose (X, X^*) and (Y, Y^*) are spaces paired in duality, with the duality pairing between X and X^*, and Y and Y^* given by $\langle \cdot\,,\cdot\rangle_X$ and $\langle \cdot\,,\cdot\rangle_Y$, respectively. If A is a linear transformation mapping X to Y, then the adjoint of A, denoted A^*, is the linear transformation mapping Y^* to X^* defined by the condition that

$$\langle A^*y\,,x\rangle_X = \langle y\,,Ax\rangle_Y \quad \text{for all } x \in X \text{ and } y \in Y^*.$$

The dual variational objects studied in this section are the cone of regular normals and the set of regular subgradients. The cone of regular normal vectors to the epigraph of a at a point $(p,\mu) \in \mathrm{epi}\,(a)$, denoted $\widehat{N}_{\mathrm{epi}\,(a)}(p,\mu)$, is given by

$$\left\{ (z,\eta) \;\middle|\; \begin{array}{c} \langle (z,\eta)\,,(q,\tau)-(p,\mu)\rangle \le o(\|(q,\tau)-(p,\mu)\|) \\ \forall\,(q,\tau)\in \mathrm{epi}\,(a) \end{array} \right\},$$

where $\langle (z,\eta)\,,(q,\tau)\rangle = \eta\tau + \langle z\,,q\rangle$ (note that $\mathrm{epi}\,(a) \subset \mathcal{P}_n \times \mathbb{R}$ so that η and τ are real). The cone of regular normals is defined to be the empty set at points not in the epigraph of a. The set of regular subgradients of a at $p \in \mathrm{dom}\, a = \mathcal{M}_n$ is given by

$$\hat{\partial} a(p) = \{ z \mid a(q) \ge a(p) + \langle z\,,q-p\rangle + o(\|q-p\|)\ \forall\, q \in \mathcal{P}_n \}.$$

If $p \ne \mathcal{M}_n$, we define $\hat{\partial} a(p)$ to be the empty set. By [RW98, Theorem 8.9], we have the following relationship between the cone of regular normals and the set of regular subgradients:

$$\hat{\partial} a(p) = \left\{ z \;\middle|\; (z,-1) \in \widehat{N}_{\mathrm{epi}\,(a)}(p,a(p)) \right\}. \tag{2.5}$$

In addition, [RW98, Proposition 6.5] tells us that the cone of regular normals is the polar of the tangent cone at points $(p,a(p)) \in \mathrm{epi}\,(a)$:

$$\widehat{N}_{\mathrm{epi}\,(a)}(p,a(p)) = T_{\mathrm{epi}\,(a)}(p,a(p))^\circ\ , \tag{2.6}$$

where

$$T_{\mathrm{epi}\,(a)}(p,a(p))^\circ = \left\{ (z,\xi) \;\middle|\; \langle (z,\xi)\,,(h,\omega)\rangle \le 0\ \forall\,(h,\omega) \in T_{\mathrm{epi}\,(a)}(p,a(p)) \right\}.$$

We take a moment to observe two important consequences of the equivalence (2.6). These observations are based on the relations (1.4) and (1.5). By (1.4), we have that the vector $(e_{(n,0)},0)$ is orthogonal to every vector in $T_{\mathrm{epi}\,(a)}(p,a(p))$, regardless of the choice of the polynomial $p \in \mathcal{M}_n$. Therefore, by (2.6),

$$\left\{ (\beta e_{(n,0)},0) \mid \beta \in \mathbb{C} \right\} \subset \widehat{N}_{\mathrm{epi}\,(a)}(p,a(p)) \quad \text{for every } p \in \mathcal{M}_n. \tag{2.7}$$

In addition, (1.5) and (2.6) imply that

$$\left\{ (\beta e_{(n,0)},0) \mid \beta \in \mathbb{C} \right\} = \widehat{N}_{\mathrm{epi}\,(a)}(p,\mu), \quad \text{whenever } \mu > a(p). \tag{2.8}$$

We now proceed to derive an expression for $\widehat{N}_{\mathrm{epi}\,(a)}(p,a(p))$ using (2.6) and Theorem 1.6. We then use the relation (2.5) to determine $\hat{\partial} a(p)$.

THEOREM 2.1. *Let $p \in \mathcal{M}_n$ have factorization* (1.33) *and let $\mathcal{I}(p)$ be as defined in* (1.38). *Then (z,η) is an element of the normal cone $\widehat{N}_{epi\,(a)}(p,a(p))$ if and only if*

$$\eta \le 0 \tag{2.9}$$

and the vector $u \in \mathcal{S}_p$ defined by $u = \nabla F_p(0)^* z$ and given the representation

$$u_j = \sum_{l=1}^{n_j} \mu_{jl} e_{(n_j - l, \lambda_j)} \quad \textit{for } j = 1, \ldots, m \tag{2.10}$$

satisfies

$$u_j = 0 \quad \textit{for } j \notin \mathcal{I}(p) \textit{ and } j \neq 0, \tag{2.11}$$

$$\textit{Re}\,\mu_{j1} \leq 0 \textit{ and } \textit{Im}\,\mu_{j1} = 0 \quad \textit{for } j \in \mathcal{I}(p), \tag{2.12}$$

$$\textit{Re}\,\mu_{j2} \leq 0 \quad \textit{for } j \in \mathcal{I}(p), \quad \textit{and} \tag{2.13}$$

$$\sum_{j \in \mathcal{I}(p)} n_j \mu_{j1} = \eta. \tag{2.14}$$

Proof. Let $(h, \omega) \in T_{\mathrm{epi}\,(a)}(p, a(p))$. By Theorem 1.6, we know that there exists $(0, w_1, w_2, \ldots, w_m) \in \mathcal{S}_p$ such that $h = \nabla F_p(0)(0, w_1, w_2, \ldots, w_m)$, where for $j = 1, 2, \ldots, m$ each w_j has the representation (1.42) with the coefficients β_{js} satisfying (1.43)–(1.46) for $j \in \mathcal{I}(p)$, and, for $j \notin \mathcal{I}(p)$,

$$\beta_{js}, \quad s = 1, 2, \ldots, n_j, \text{ are unrestricted.} \tag{2.15}$$

Now let $(z, \eta) \in \mathcal{P}_n \times \mathbb{R}$ and set $u = (u_0, u_1, \ldots, u_m) = \nabla F_p(0)^* z$, where each u_j, $j = 1, \ldots, m$ is given the representation (2.10). Then, from definition (2.4), we have

$$\begin{aligned} \langle (z, \eta), (h, \omega) \rangle &= \eta\omega + \langle z, h \rangle \\ &= \eta\omega + \langle z, \nabla F_p(0)(0, w_1, w_2, \ldots, w_m) \rangle \\ &= \eta\omega + \langle \nabla F_p(0)^* z, (0, w_1, w_2, \ldots, w_m) \rangle_p \\ &= \eta\omega + \langle (u_0, u_1, \ldots, u_m), (0, w_1, w_2, \ldots, w_m) \rangle_p \\ &= \eta\omega + \sum_{j=1}^{m} \langle u_j, w_j \rangle_{(n_j - 1, \lambda_j)} \\ &= \eta\omega + \sum_{j=1}^{m} \sum_{l=1}^{n_j} \mathrm{Re}\, \bar{\mu}_{jl} \beta_{jl}. \end{aligned} \tag{2.16}$$

Hence, by (2.6), $(z, \eta) \in \widehat{N}_{\mathrm{epi}\,(a)}(p, a(p))$ if and only if

$$\eta\omega + \sum_{j=1}^{m} \sum_{l=1}^{n_j} \mathrm{Re}\, \bar{\mu}_{jl} \beta_{jl} \leq 0 \tag{2.17}$$

for all choices of ω and β_{jl}, $j = 1, \ldots, m$, $l = 1, \ldots, n_j$, satisfying (1.43)–(1.46) for each $j \in \mathcal{I}(p)$.

We first show that any $(z, \eta) \in \mathcal{P}_n \times \mathbb{R}$ for which the associated vector $u = (u_0, u_1, \ldots, u_m) = \nabla F_p(0)^* z$, where each u_j, $j = 1, \ldots, m$, has representation (2.10) and for which η and μ_{jl}, $j = 1, \ldots, m$, $l = 1, \ldots, n_j$, satisfy (2.9) and (2.11)–(2.14) is necessarily an element of the normal cone $\widehat{N}_{\mathrm{epi}\,(a)}(p, a(p))$. For this purpose, suppose that ω and β_{jl}, $j = 1, \ldots, m$, $l = 1, \ldots, n_j$ satisfy (1.43)–(1.46) for each $j \in \mathcal{I}(p)$ so that the corresponding vector (h, ω) is an element of the tangent cone $T_{\mathrm{epi}\,(a)}(p, a(p))$.

Then

$$\begin{aligned}\langle (z,\eta),(h,\omega)\rangle &= \eta\omega + \sum_{j=1}^{m}\sum_{l=1}^{n_j} \operatorname{Re}\bar{\mu}_{jl}\beta_{jl}\\ &= \eta\omega + \sum_{j\in\mathcal{I}(p)} [\mu_{j1}\operatorname{Re}\beta_{j1} + \operatorname{Re}\mu_{j2}\operatorname{Re}\beta_{j2}]\\ &\le \eta\omega - \sum_{j\in\mathcal{I}(p)} n_j\mu_{j1}\left(\frac{-\operatorname{Re}\beta_{j1}}{n_j}\right)\\ &\le \eta\omega - \sum_{j\in\mathcal{I}(p)} n_j\mu_{j1}\omega\\ &= 0,\end{aligned}$$

where the first equality follows from (2.16), the second equality from (2.11), (2.12), (1.45), and (1.46), the first inequality from (2.13) and (1.44), the second inequality from (1.43), and the final equality from (2.14). Therefore, the set of (z,η) satisfying (2.9)–(2.14) is contained in $\widehat{N}_{\operatorname{epi}(a)}(p,a(p))$.

We now show the reverse inclusion. Let $(z,\eta)\in\widehat{N}_{\operatorname{epi}(a)}(p,a(p))$ and set $u=(u_0,u_1,\ldots,u_m)=\nabla F_p(0)^*z$ with each u_j, $j=1,\ldots,m$ given representation (2.10). We show that (z,η) must satisfy the conditions (2.11)–(2.14) by requiring that the inequality (2.17) holds for every (h,ω) in the tangent cone $T_{\operatorname{epi}(a)}(p,a(p))$. To this end, let (h,ω) be any element of the tangent cone $T_{\operatorname{epi}(a)}(p,a(p))$ so that the corresponding vectors w_j, $j=1,\ldots,m$, satisfy (1.43)–(1.46) for each $j\in\mathcal{I}(p)$ and (2.15) for $j\notin\mathcal{I}(p)$. By setting $\omega=1$ and all β_{jl} equal to zero in (2.17), we find that $\eta\le 0$. By (2.15), β_{js} is free for $j\notin\mathcal{I}(p)$, $s=1,2,\ldots,n_j$, so that (2.17) implies that (2.11) holds. Since $\operatorname{Im}\beta_{j1}$ is free whenever $j\in\mathcal{I}(p)$, (2.17) implies that $\operatorname{Im}\mu_{j1}=0$ for all $j\in\mathcal{I}(p)$. In addition, (1.43) and (2.17) imply that $\operatorname{Re}\mu_{j2}\le 0$ for all $j\in\mathcal{I}(p)$. Therefore, (2.9), (2.11), the second half of (2.12) (i.e., the equality), and (2.13) have been verified.

We now establish the first half of (2.12) (i.e., the inequality) and (2.14). By taking $\operatorname{Re}\beta_{j2}=0$ for $j\in\mathcal{I}(p)$, the expression (2.16) can be simplified to

$$\langle (z,\eta),(h,\omega)\rangle = \eta\omega + \sum_{j\in\mathcal{I}(p)} \mu_{j1}\operatorname{Re}\beta_{j1}. \tag{2.18}$$

By combining this with (2.17), we must have

$$\sum_{j\in\mathcal{I}(p)} \mu_{j1}\operatorname{Re}\beta_{j1} \le -\eta\omega \tag{2.19}$$

for all choices of ω and $\operatorname{Re}\beta_{j1}$, $j\in\mathcal{I}(p)$, satisfying (1.43). Observe that (1.43) holds if and only if

$$\omega \ge \max_{j\in\mathcal{I}(p)} \frac{-\operatorname{Re}\beta_{j1}}{n_j}. \tag{2.20}$$

Since $-\eta\ge 0$, we can multiply this inequality through by $-\eta$ to obtain the inequality

$$-\eta\omega \ge -\eta \max_{j\in\mathcal{I}(p)} \frac{-\operatorname{Re}\beta_{j1}}{n_j}. \tag{2.21}$$

Since the right-hand side of this inequality yields the smallest possible value of the product $-\eta\omega$, we find that (1.43) and (2.19) hold if and only if

$$\sum_{j\in\mathcal{I}(p)} \mu_{j1}\operatorname{Re}\beta_{j1} \le (-\eta)\max_{j\in\mathcal{I}(p)} \frac{-\operatorname{Re}\beta_{j1}}{n_j} \qquad \forall \beta_{j1}\in\mathbb{C},\ j\in\mathcal{I}(p). \tag{2.22}$$

Consider two cases: $\eta = 0$ and $\eta < 0$. If $\eta = 0$, then (2.22) implies that $\mu_{j1} = 0$ for all $j \in \mathcal{I}(p)$ so that (2.12) and (2.14) are satisfied. On the other hand, if $\eta < 0$, define $\tilde{\mu}_j = \frac{n_j}{\eta}\mu_{j1}$ and $\tilde{\beta}_j = \frac{-\operatorname{Re}\beta_{j1}}{n_j}$ for $j \in \mathcal{I}(p)$. Substituting into (2.22), we obtain

$$\sum_{j\in\mathcal{I}(p)} \tilde{\mu}_j\tilde{\beta}_j \le \max_{j\in\mathcal{I}(p)} \tilde{\beta}_j \qquad \forall \tilde{\beta}_j\in\mathbb{R}. \tag{2.23}$$

But this holds if and only if $\tilde{\mu}_j \ge 0$ for $j \in \mathcal{I}(p)$ and $\sum_{j\in\mathcal{I}(p)}\tilde{\mu}_j = 1$, or equivalently, (2.12) and (2.14) hold. □

Theorem 2.1 and (2.5) immediately yield the following representation for the set of regular subgradients.

Theorem 2.2. *Let $p \in \mathcal{M}_n$ have factorization* (1.33). *Then $z \in \hat{\partial}a(p)$ if and only if the vector of polynomials*

$$\nabla F_p(0)^* z = (u_0, u_1, \ldots, u_m) \in \mathcal{S}_p,$$

with

$$u_j = \sum_{l=1}^{n_j} \mu_{jl} e_{(n_j-l,\lambda_j)}, \qquad j = 1, 2, \ldots, m,$$

is such that

$$\begin{aligned} u_j &= 0 \quad \textit{for } j \notin \mathcal{I}(p) \textit{ and } j \ne 0,\\ Re\,\mu_{j1} \le 0 \textit{ and } Im\,\mu_{j1} &= 0 \quad \textit{for } j \in \mathcal{I}(p),\\ Re\,\mu_{j2} &\le 0 \quad \textit{for } j \in \mathcal{I}(p), \textit{ and}\\ \sum_{j\in\mathcal{I}(p)} n_j\mu_{j1} &= -1. \end{aligned}$$

A more concise representation for the set of regular subgradients is possible. First note that if $p = e_{(n,\lambda_0)}$, then, for $(h_0, h_1) \in \mathcal{S}_p = \mathbb{C}\times\mathcal{P}_{n-1}$,

$$\nabla F_p(0)(h_0, h_1) = h_0 e_{(n,\lambda_0)} + h_1$$

and

$$\nabla F_p(0)^* \sum_{j=0}^{n} b_j e_{(n-j,\lambda_0)} = \left(b_0, \sum_{j=1}^{n} b_j e_{(n-j,\lambda_0)}\right), \tag{2.24}$$

since

$$\left\langle \sum_{j=0}^{n} b_j e_{(n-j,\lambda_0)}, h_0 e_{(n,\lambda_0)} + h_1 \right\rangle_{(n,\lambda_0)} = \left\langle \left(b_0, \sum_{j=1}^{n} b_j e_{(n-j,\lambda_0)}\right), (h_0, h_1) \right\rangle_p .$$

In this case $\nabla F_p(0)^* = \nabla F_p(0)^{-1}$. Hence, by Theorem 2.2, we have the following formula for the set of regular subgradients of a at $e_{(n,\lambda_0)}$:

$$\hat{\partial} a(e_{(n,\lambda_0)}) = \left\{ z \;\middle|\; \begin{array}{c} z = \sum_{j=0}^n \mu_j e_{(n-j,\lambda_0)}, \\ \text{where } \mu_j \in \mathbb{C},\ j = 0, 1, \ldots, n, \\ \mu_1 = \frac{-1}{n},\ \text{and } \operatorname{Re}(\mu_2) \le 0 \end{array} \right\}. \tag{2.25}$$

In the general case a similar formula can be obtained with the aid of the recession cone of the set $\hat{\partial} a(e_{(n,\lambda_0)})$:

$$\hat{\partial} a(e_{(n,\lambda_0)})^\infty = \left\{ z \;\middle|\; \begin{array}{c} z = \sum_{j=0}^n \mu_j e_{(n-j,\lambda_0)}, \\ \text{where } \mu_j \in \mathbb{C},\ j = 0, 1, \ldots, n, \\ \mu_1 = 0,\ \text{and } \operatorname{Re}(\mu_2) \le 0 \end{array} \right\}. \tag{2.26}$$

Define $\hat{\partial} a(e_{(n,\lambda_0)})^{\tilde{\infty}}$ as the projection of $\hat{\partial} a(e_{(n,\lambda_0)})^\infty$ onto $\mathcal{P}_{n-1}$:

$$\hat{\partial} a(e_{(n,\lambda_0)})^{\tilde{\infty}} = \left\{ z \;\middle|\; \begin{array}{c} z = \sum_{j=1}^n \mu_j e_{(n-j,\lambda_0)}, \\ \text{where } \mu_j \in \mathbb{C},\ j = 1, \ldots, n, \\ \mu_1 = 0,\ \text{and } \operatorname{Re}(\mu_2) \le 0 \end{array} \right\}.$$

Then, given a polynomial $p \in \mathcal{M}_n$ having prime factorization (1.33), the set of regular subgradients of a at p has the form

$$\hat{\partial} a(p) = \nabla F_p(0)^{-*} \left[\operatorname{conv} \{ v_j \mid j \in \mathcal{I}(p) \} + K \right], \tag{2.27}$$

where $v_1, \ldots, v_m \in \mathcal{S}_p$ are given by

$$\begin{aligned} v_1 &= -\frac{1}{n_1}(0, e_{(n_1-1,\lambda_1)}, 0, \ldots, 0), \\ v_2 &= -\frac{1}{n_2}(0, 0, e_{(n_2-1,\lambda_2)}, 0, \ldots, 0), \\ &\vdots \\ v_m &= -\frac{1}{n_m}(0, 0, \ldots, 0, e_{(n_m-1,\lambda_m)}), \end{aligned}$$

and K is the convex cone in $\mathcal{S}_p$ given by

$$K = \mathbb{C} \times \hat{\partial} a(e_{(n_1,\lambda_1)})^{\tilde{\infty}} \times \cdots \times \hat{\partial} a(e_{(n_m-1,\lambda_m)})^{\tilde{\infty}}.$$

Observe that this implies the recession cone of $\hat{\partial} a(p)$ is given by

$$\hat{\partial} a(p)^\infty = \nabla F_p(0)^{-*} K. \tag{2.28}$$

3. Subdifferential regularity.

The set of normal vectors to epi (a) at a point $(p, \mu) \in$ epi (a) is given by

$$N_{\operatorname{epi}(a)}(p, \mu) = \left\{ (z, \omega) \;\middle|\; \begin{array}{c} \exists \{(p_k, \mu_k)\} \subset \operatorname{epi}(a),\ \{(z_k, \omega_k)\} \subset \mathcal{P}_n \times \mathbb{R} \\ \text{with } (z_k, \omega_k) \in \widehat{N}_{\operatorname{epi}(a)}(p_k, \mu_k)\ \forall\, k, \\ \text{such that} \\ (p_k, \mu_k) \to (p, \mu) \text{ and } (z_k, \omega_k) \to (z, \omega) \end{array} \right\}. \tag{3.1}$$

By convention $N_{\mathrm{epi}\,(a)}(p,\mu)=\emptyset$ if $(p,\mu)\notin \mathrm{epi}\,(a)$. The abscissa mapping a is said to be subdifferentially regular at a point $(p,\mu)\in \mathrm{epi}\,(a)$ (equivalently, $\mathrm{epi}\,(a)$ is Clarke regular at (p,μ)) if

$$\widehat{N}_{\mathrm{epi}\,(a)}(p,\mu)=N_{\mathrm{epi}\,(a)}(p,\mu) \tag{3.2}$$

[RW98, Definition 7.25]. The goal of this section is to show that the set $\mathrm{epi}\,(a)$ is everywhere subdifferentially regular.

Some simplification in definition (3.1) is possible due to the continuity of a on its domain $\mathcal{M}_n$. Recall from (2.8) that

$$\widehat{N}_{\mathrm{epi}\,(a)}(p,\mu)=\left\{(\beta e_{(n,0)},0)\mid \beta\in\mathbb{C}\right\}\quad \text{whenever } \mu>a(p).$$

Since this subspace is constant on the set $\{(p,\mu)\mid \mu>a(p)\}$, we find that

$$N_{\mathrm{epi}\,(a)}(p,\mu)=\widehat{N}_{\mathrm{epi}\,(a)}(p,\mu)\quad \text{whenever } \mu>a(p).$$

Therefore, to establish that a is everywhere subdifferentially regular we need only establish the equivalence (3.2) at the points $(p,a(p))$ for $p\in\mathcal{M}_n$. In addition, from (2.7), we have $\left\{(\beta e_{(n,0)},0)\mid \beta\in\mathbb{C}\right\}\subset \widehat{N}_{\mathrm{epi}\,(a)}(p,\mu)$ for all $(p,\mu)\in\mathrm{epi}\,(a)$. Hence, it is always the case that

$$\widehat{N}_{\mathrm{epi}\,(a)}(p,\eta)\subset \widehat{N}_{\mathrm{epi}\,(a)}(p,\mu)\quad \text{whenever } a(p)\le\mu<\eta.$$

Therefore, the representation for the normal cone at the points $(p,a(p))$ for $p\in\mathcal{M}_n$ can be refined to

$$N_{\mathrm{epi}\,(a)}(p,a(p))=\left\{(z,\omega)\;\middle|\;\begin{array}{c}\exists\,\{p_k\}\subset\mathcal{M}_n,\ \{(z_k,\omega_k)\}\subset\mathcal{P}_n\times\mathbb{R}\\ \text{with } (z_k,\omega_k)\in\widehat{N}_{\mathrm{epi}\,(a)}(p_k,a(p_k))\ \forall\,k,\\ \text{such that}\\ p_k\to p \text{ and } (z_k,\omega_k)\to(z,\omega)\end{array}\right\}. \tag{3.3}$$

However, even with this simplification, we are confronted with a significant technical hurdle. Recall from Theorem 2.1 that the regular normals are characterized through the adjoint operator $\nabla F_p(0)^*$. Therefore, we now need to compute limits of these operators along sequences $p_k\to p$. But these adjoints are defined as linear transformations from $\mathcal{P}_n$ to $\mathcal{S}_{p_k}$ and are based on the inner products $\langle\cdot\,,\cdot\rangle_{p_k}$. How can we interpret limits of the adjoints $\nabla F_{p_k}(0)^*$ when the spaces $\mathcal{S}_{p_k}$ and their associated inner products $\langle\cdot\,,\cdot\rangle_{p_k}$ may not even be commensurate? The answer again lies with the local factorization lemma, Lemma 1.4.

Suppose that the polynomial $p\in\mathcal{M}_n$ has prime factorization (1.33) and let $\{p_k\}$ be a sequence of monic polynomials converging to p. Lemma 1.4 says that, by trimming off finitely many elements of the sequence if necessary, we may assume with no loss of generality that each of the polynomials p_k has a factorization of the form

$$p_k=\prod_{j=1}^{m}q_{kj}, \tag{3.4}$$

where the roots of the polynomials q_{kj}, $j=1,\dots,m$, are pairwise disjoint and $q_{kj}\to e_{(n_j,\lambda_j)}$ for each $j=1,\dots,m$. Moreover, since there are only finitely many partitions of n, we may assume with no loss in generality (by extracting a subsequence if necessary) that there exist positive integers ℓ_j, $j = 1, \ldots, m$, and n_{js}, $j = 1, \ldots, m$, $s = 1, \ldots, \ell_j$, with $\sum_{s=1}^{\ell_j} n_{js} = n_j$, such that, for each $k = 1, 2, \ldots$,

$$q_{kj} = \prod_{s=1}^{\ell_j} e_{(n_{js}, \lambda_{kjs})}, \tag{3.5}$$

where the complex numbers λ_{kjs}, $j = 1, \ldots, m$, $s = 1, \ldots, \ell_{kj}$ are distinct and satisfy $\lambda_{kjs} \to \lambda_j$ for $s = 1, \ldots, \ell_j$. Hence, for each $k = 1, 2, \ldots$, we have

$$\mathcal{S}_{p_k} = \mathbb{C} \times \left[\bigtimes_{j=1}^{m} \left(\mathcal{P}_{n_{j1}-1} \times \cdots \times \mathcal{P}_{n_{j\ell_j}-1} \right) \right], \tag{3.6}$$

$$F_{p_k}(v_0, v_{11}, \ldots, v_{1\ell_1}, \ \ldots \ , v_{m1}, \ldots, v_{m\ell_m}) = (1 + v_0) \prod_{j=1}^{m} \prod_{s=1}^{\ell_j} (e_{(n_{js}, \lambda_{kjs})} + v_{js}), \tag{3.7}$$

and

$$\nabla F_{p_k}(0)(h_0, h_{11}, \ldots, h_{1\ell_1}, \ \ldots \ , h_{m1}, \ldots, h_{m\ell_m}) = h_0 r_{k0} + \sum_{j=1}^{m} r_{kj} \left(\sum_{s=1}^{\ell_j} \hat{r}_{kjs} h_{js} \right), \tag{3.8}$$

where

$$r_{k0} = p_k \quad \text{and} \quad r_{kj_0} = \prod_{j \neq j_0} \prod_{s=1}^{\ell_j} e_{(n_{js}, \lambda_{kjs})}, \qquad j_0 = 1, \ldots, m,$$

and

$$\hat{r}_{kj_0 s_0} = \prod_{\substack{s=1 \\ s \neq s_0}}^{\ell_{j_0}} e_{(n_{j_0 s}, \lambda_{kj_0 s})}, \quad j_0 = 1, \ldots, m, \ s_0 = 1, \ldots, \ell_j \ .$$

Let us write $\hat{\mathcal{S}} = \mathcal{S}_{p_k}$, since $\mathcal{S}_{p_k}$ is fixed for all $k = 1, 2, \ldots$. Note that as $k \to \infty$, we have $r_{kj} \to r_j$, where r_j is defined in (1.37), for $j = 0, 1, \ldots, m$, and $\hat{r}_{kjs} \to e_{(\bar{n}_{js}, \lambda_j)}$, where $\bar{n}_{js} = n_j - n_{js}$, for $j = 1, \ldots, m$, $s = 1, \ldots, \ell_j$. Hence, $\nabla F_{p_k}(0) \to \Psi$, where the linear transformation $\Psi : \hat{\mathcal{S}} \to \mathcal{P}_n$ is given by

$$\Psi(h_0, h_{11}, \ldots, h_{1\ell_1}, \ \ldots \ , h_{m1}, \ldots, h_{m\ell_m}) = h_0 r_0 + \sum_{j=1}^{m} r_j \left(\sum_{s=1}^{\ell_j} e_{(\bar{n}_{js}, \lambda_j)} h_{js} \right). \tag{3.9}$$

Observe that the representation of $\nabla F_p(0)$ given in (1.36) and (1.37) enables us to write the operator Ψ as the composition

$$\Psi = \nabla F_p(0) \circ \Xi, \tag{3.10}$$

where the linear operator $\Xi : \hat{\mathcal{S}} \to \mathcal{S}_p$ is given by

$$\begin{aligned} &\Xi(h_0, h_{11}, \ldots, h_{1\ell_1},.,.,h_{m1}, \ldots, h_{m\ell_m}) \\ &\qquad = \left(h_0, \sum_{s=1}^{\ell_1} e_{(\bar{n}_{1s}, \lambda_1)} h_{1s}, \ldots, \sum_{s=1}^{\ell_m} e_{(\bar{n}_{ms}, \lambda_m)} h_{ms} \right). \end{aligned} \tag{3.11}$$

Theorem 2.1 gives us access to the regular normals through the adjoint operators $\nabla F_p(0)^*$. Thus, in order to understand the normal cone, which consists of the limits of the regular normals, we need to come to an understanding of the limit of the adjoints $\nabla F_{p_k}(0)^*$. This limit is an adjoint of the operator Ψ. However, what this means needs clarification since each of the adjoints $\nabla F_{p_k}(0)^*$ arises from a different duality pairing. We need to determine the correct duality pairing for the definition of the adjoint Ψ^* so that it is the limit of the operators $\nabla F_{p_k}(0)^*$.

The duality pairing that we seek is obtained from our earlier observation (2.3) that the mapping $(p, q, \lambda) \mapsto \langle p\,, q\rangle_{(n,\lambda)}$ is continuous. This continuity implies that the pointwise limit of the inner products $\langle\cdot\,,\cdot\rangle_{p_k}$ exists as $k \to \infty$. Indeed, for each

$$u = (u_0, u_{11}, \ldots, u_{1\ell_1}, \ \ldots\ , u_{1\ell_m}, \ldots, u_{m\ell_m})$$

and

$$v = (v_0, v_{11}, \ldots, v_{1\ell_1}, \ \ldots\ , v_{1\ell_m}, \ldots, v_{m\ell_m})$$

in $\hat{\mathcal{S}}$, we have

$$\langle u\,, v\rangle_{p_k} \to \langle u\,, v\rangle_{\infty}\,,$$

where

$$\langle u\,, v\rangle_{\infty} = \langle u_0\,, v_0\rangle + \sum_{j=1}^{m}\sum_{s=1}^{\ell_j} \langle u_{js}\,, v_{js}\rangle_{(n_{js}-1,\lambda_j)}\ . \tag{3.12}$$

Therefore, if we define Ψ^* to be the adjoint of Ψ with respect to the duality pairings $(\mathcal{P}_n, \langle\cdot\,,\cdot\rangle)$ and $(\hat{\mathcal{S}}, \langle\cdot\,,\cdot\rangle_{\infty})$, then

$$\nabla F_{p_k}(0)^* \to \Psi^*. \tag{3.13}$$

Our next task is to derive a representation for the operator Ψ^*. Using the representation for Ψ given in (3.10), this reduces to deriving a representation for the adjoint of the operator Ξ. For this, the following lemma provides the key.

LEMMA 3.1. *Let $\lambda_0 \in \mathbb{C}$ and let $\delta = (n_1, n_2, \ldots, n_m)$ be a partition of n. Define $\mathcal{D}_\delta$ to be the product space*

$$\mathcal{D}_\delta = \mathcal{P}_{(n_1-1)} \times \mathcal{P}_{(n_2-1)} \times \cdots \times \mathcal{P}_{(n_m-1)},$$

endowed with the inner product

$$\langle (u_1, \ldots, u_m)\,, (v_1, \ldots, v_m)\rangle_{(\delta,\lambda_0)} = \sum_{j=1}^{m} \langle u_j\,, v_j\rangle_{(n_j-1,\lambda_0)}\,.$$

For $j = 1, 2, \ldots, m$, define $\bar{n}_j = n - n_j$ and consider the linear transformation $\Phi_{(\delta,\lambda_0)} : \mathcal{D}_\delta \to \mathcal{P}_{n-1}$ given by

$$\Phi_{(\delta,\lambda_0)}(h_1, \ldots, h_m) = \sum_{j=1}^{m} e_{(\bar{n}_j,\lambda_0)} h_j\ .$$

Then the adjoint of $\Phi_{(\delta,\lambda_0)}$ with respect to the duality pairings

$$(\mathcal{P}_{n-1}, \langle\cdot\,,\cdot\rangle_{(n-1,\lambda_0)}) \quad \text{and} \quad (\mathcal{D}_\delta, \langle\cdot\,,\cdot\rangle_{(\delta,\lambda_0)})$$

is given by

$$\Phi^*_{(\delta,\lambda_0)}\left(\sum_{j=1}^{n} b_j e_{(n-j,\lambda_0)}\right) = \left(\sum_{j=1}^{n_1} b_j e_{(n_1-j,\lambda_0)},\ldots,\sum_{j=1}^{n_m} b_j e_{(n_m-j,\lambda_0)}\right).$$

Proof. Define $J_s = \{j \mid n_j \geq s\}$ for $s = 1, 2, \ldots, n$. Note that J_s may be empty for some values of s. For example, if $m \geq 2$, then $J_n = \emptyset$. Let $(h_1, \ldots, h_m) \in \mathcal{D}_\delta$, where each $h_j \in \mathcal{P}_{n_j-1}$ has representation

$$h_j = a_{j1} e_{(n_j-1,\lambda_0)} + a_{j2} e_{(n_j-2,\lambda_0)} + \cdots + a_{jn_j}\ .$$

Given $q \in \mathcal{P}_{n-1}$ with

$$q = b_1 e_{(n-1,\lambda_0)} + \cdots + b_n,$$

we have

$$\begin{aligned}
&\big\langle q\,, \Phi_{(\delta,\lambda_0)}(h_1,\ldots,h_m)\big\rangle_{(n-1,\lambda_0)}\\
&= \Bigg\langle \big(b_1 e_{(n-1,\lambda_0)} + \cdots + b_n\big)\ ,\ \Bigg(e_{(n-1,\lambda_0)}\left(\sum_{j\in J_1} a_{j1}\right)\\
&\qquad\qquad + e_{(n-2,\lambda_0)}\left(\sum_{j\in J_2} a_{j2}\right) + \cdots + \left(\sum_{j\in J_n} a_{jn}\right)\Bigg)\Bigg\rangle_{(n-1,\lambda_0)}\\
&= \operatorname{Re}\left[\bar b_1\left(\sum_{j\in J_1} a_{j1}\right) + \cdots + \bar b_n\left(\sum_{j\in J_n} a_{jn}\right)\right]\\
&= \operatorname{Re}\left[\sum_{s=1}^{n_1} \bar b_s a_{1s} + \sum_{s=1}^{n_2} \bar b_s a_{2s} + \cdots + \sum_{s=1}^{n_m} \bar b_s a_{ms}\right]\\
&= \Bigg\langle \left(\sum_{s=1}^{n_1} b_s e_{(n_1-s,\lambda_0)},\ldots,\sum_{s=1}^{n_m} b_s e_{(n_m-s,\lambda_0)}\right)\ ,\\
&\qquad\qquad \left(\sum_{s=1}^{n_1} a_{1s} e_{(n_1-s,\lambda_0)},\ldots,\sum_{s=1}^{n_m} a_{ms} e_{(n_m-s,\lambda_0)}\right)\Bigg\rangle_{(\delta,\lambda_0)}\\
&= \Bigg\langle \left(\sum_{s=1}^{n_1} b_s e_{(n_1-s,\lambda_0)},\ldots,\sum_{s=1}^{n_m} b_s e_{(n_m-s,\lambda_0)}\right)\ ,(h_1,\ldots,h_m)\Bigg\rangle_{(\delta,\lambda_0)}.
\end{aligned}$$

Since this relation holds for all possible choices of $q \in \mathcal{P}_{n-1}$ and $(h_1, \ldots, h_m) \in \mathcal{D}_\delta$, we have established the result. □

By using the notation developed in Lemma 3.1, we can rewrite the operator $\Xi : \hat{\mathcal{S}} \to \mathcal{S}_p$, defined in (3.11), as

$$\Xi = \big(I, \Phi_{(\delta_1,\lambda_1)}, \ldots, \Phi_{(\delta_m,\lambda_m)}\big)\,, \tag{3.14}$$

where $\delta_j = (n_{j1}, n_{j2}, \ldots, n_{j\ell_j})$ is a partition of n_j for each $j = 1, 2, \ldots, m$. Hence, from (3.10), we have

$$\Psi^* = \Xi^* \circ \nabla F_p(0)^*, \tag{3.15}$$

where $\Xi^* : \mathcal{S}_p \to \hat{\mathcal{S}}$ can be written as

$$\Xi^* = \left(I, \Phi^*_{(\delta_1,\lambda_1)}, \dots, \Phi^*_{(\delta_m,\lambda_m)} \right). \tag{3.16}$$

An explicit representation for the operator Ξ^* can be obtained by applying Lemma 3.1 to each of the operators $\Phi_{(\delta_j,\lambda_j)}$ for $j = 1, 2, \dots, m$.

We now prove the main result of this section.

Theorem 3.2. *The abscissa mapping a is everywhere subdifferentially regular. Equivalently, epi (a) is Clarke regular.*

Proof. Let $p \in \mathcal{P}_n$ have factorization (1.33). Let $(z, \omega) \in N_{\text{epi}\,(a)}(p, a(p))$ so that there exist sequences $\{p_k\} \subset \mathcal{M}_n$ and $\{(z_k, \omega_k)\} \subset \mathcal{P}_n \times \mathbb{R}$ such that $p_k \to p$, $(z_k, \omega_k) \to (z, \omega)$, and $(z_k, \omega_k) \in \widehat{N}_{\text{epi}\,(a)}(p_k, a(p_k))$ for $k = 1, 2, \dots$. We need to show that $(z, \omega) \in \widehat{N}_{\text{epi}\,(a)}(p, a(p))$.

The discussion preceding this theorem shows that we may assume with no loss of generality that (3.4)–(3.16) hold for the sequence $\{p_k\}$. Hence, we make free use of these facts and their associated notations.

Let $\tilde{\mathcal{I}}(p_k) = \{(j, s) \mid a(p_k) = \operatorname{Re} \lambda_{kjs}\}$. Since $(z_k, \omega_k) \in \widehat{N}_{\text{epi}\,(a)}(p_k, a(p_k))$ for $k = 1, 2, \dots$, Theorem 2.1 states that $\omega_k \le 0$ and there exists

$$u_k = (u_{k0}, u_{k11}, \dots, u_{k1\ell_1}, \dots, u_{km1}, \dots, u_{km\ell_m}) \in \hat{\mathcal{S}}$$

with

$$u_{kjs} = \sum_{t=1}^{n_{js}} \mu_{kjst} e_{(n_{js}-t, \lambda_{kjs})}, \qquad \begin{array}{l} k = 1, 2, \dots, \\ j = 1, 2, \dots, m, \\ s = 1, 2, \dots, \ell_j \ , \end{array}$$

such that

$$u_k = \nabla F_{p_k}(0)^* z_k, \tag{3.17}$$

$$u_{kjs} = 0 \quad \text{for } (j, s) \notin \tilde{\mathcal{I}}(p_k), \tag{3.18}$$

$$\operatorname{Re} \mu_{kjs1} \le 0 \text{ and } \operatorname{Im} \mu_{kjs1} = 0 \quad \text{for } (j, s) \in \tilde{\mathcal{I}}(p_k), \tag{3.19}$$

$$\operatorname{Re} \mu_{kjs2} \le 0 \quad \text{for } (j, s) \in \tilde{\mathcal{I}}(p_k), \text{ and} \tag{3.20}$$

$$\sum_{(j,s) \in \tilde{\mathcal{I}}(p_k)} n_{js} \mu_{kjs1} = \omega_k. \tag{3.21}$$

Due to the finiteness of the index sets, we may assume with no loss of generality that $\tilde{\mathcal{I}}(p_k) = \tilde{\mathcal{I}}$ for all $k = 1, 2, \dots$. Define

$$\hat{\mathcal{I}} = \left\{ j \;\middle|\; (j, s) \in \tilde{\mathcal{I}} \text{ for some } s = 1, \dots, \ell_j \right\}.$$

By the continuity of the roots of the monic polynomials (as a multivalued mapping), we have $\hat{\mathcal{I}} \subset \mathcal{I}(p)$, where $\mathcal{I}(p)$ is defined in (1.38).

Using (3.13), let

$$u = \Psi^* z = \lim_{k \to \infty} \nabla F_{p_k}(0)^* z_k = \lim_{k \to \infty} u_k \tag{3.22}$$

and write

$$u = (u_0, u_{11}, \dots, u_{1\ell_1}, \dots, u_{m1}, \dots, u_{m\ell_m}) \in \hat{\mathcal{S}}$$

where

$$u_{js} = \sum_{t=1}^{n_{js}} \mu_{jst} e_{(n_{js}-t,\lambda_j)}, \tag{3.23}$$

with $\mu_{kjst} \to \mu_{jst}$ for $j = 1,\ldots,m,\ s = 1,\ldots,\ell_j,\ t = 1,\ldots,n_{js}$. By (3.22) and (3.17)–(3.21), we have

$$u_{js} = 0 \quad \text{for } (j,s) \notin \tilde{\mathcal{I}}, \tag{3.24}$$

$$\operatorname{Re}\mu_{js1} \le 0 \text{ and } \operatorname{Im}\mu_{js1} = 0 \quad \text{for } (j,s) \in \tilde{\mathcal{I}}, \tag{3.25}$$

$$\operatorname{Re}\mu_{js2} \le 0 \quad \text{for } (j,s) \in \tilde{\mathcal{I}}, \text{ and} \tag{3.26}$$

$$\sum_{(j,s)\in\tilde{\mathcal{I}}} n_{js}\mu_{js1} = \omega. \tag{3.27}$$

Set

$$(w_0, w_1, \ldots, w_m) = \nabla F_p(0)^* z, \tag{3.28}$$

with

$$w_j = \sum_{s=1}^{n_j} b_{js} e_{(n_j-s,\lambda_j)} \qquad \text{for } j = 1,\ldots,m. \tag{3.29}$$

By (3.15), (3.16), (3.22), and (3.28), we have

$$(u_{j1}, u_{j2}, \ldots, u_{j\ell_j}) = \Phi^*_{(\delta_j,\lambda_j)} w_j \qquad \text{for } j = 1,\ldots,m.$$

Consequently, by Lemma 3.1 and (3.23),

$$u_{js} = \sum_{t=1}^{n_{js}} \mu_{jst} e_{(n_{js}-t,\lambda_j)} = \sum_{t=1}^{n_{js}} b_{jt} e_{(n_{js}-t,\lambda_j)}$$

for $j = 1,\ldots,m,\ s = 1,\ldots,\ell_j$, or equivalently,

$$\mu_{jst} = b_{jt} \qquad \text{for } j = 1,\ldots,m,\ s = 1,\ldots,\ell_j,\ t = 1,\ldots,n_{js}. \tag{3.30}$$

Combining this with (3.24) and the definitions (3.23) and (3.29), we find

$$w_j = 0 \quad \text{for } j \notin \hat{\mathcal{I}}, \tag{3.31}$$

and combining (3.30) with (3.25) and (3.26) yields

$$\operatorname{Re} b_{j1} \le 0 \text{ and } \operatorname{Im} b_{j1} = 0 \quad \text{for } j \in \hat{\mathcal{I}} \text{ and} \tag{3.32}$$

$$\operatorname{Re} b_{j2} \le 0 \quad \text{for } j \in \hat{\mathcal{I}}. \tag{3.33}$$

Finally, by combining (3.30) with (3.27), we find

$$\sum_{(j,s)\in\tilde{\mathcal{I}}} n_{js} b_{j1} = \omega. \tag{3.34}$$

Note that the equivalence (3.30) implies that for every $j \in \hat{\mathcal{I}}$ for which $b_{j1} \neq 0$ it must be the case that $\mu_{js1} \neq 0$ for $s = 1, 2, \dots \ell_j$ and so $\{(j1), (j2), \dots, (j\ell_j)\} \subset \tilde{\mathcal{I}}$. Therefore, by (3.27) and (3.34),

$$\omega = \sum_{(j,s)\in\tilde{\mathcal{I}}} n_{js} b_{j1} = \sum_{j\in\hat{\mathcal{I}}} \sum_{s=1}^{\ell_j} b_{j1} n_{js} = \sum_{j\in\hat{\mathcal{I}}} b_{j1} n_j .$$

Consequently, by (3.28), (3.29), (3.31), (3.32), (3.33), the inclusion $\hat{\mathcal{I}} \subset \mathcal{I}(p)$, and Theorem 2.1, we find that $(z, \omega) \in \widehat{N}_{\mathrm{epi}\,(a)}(p, a(p))$, which establishes the result. □

Just as the set of normal vectors is defined to be the set of limits of regular normal vectors, the set of subgradients is defined to be the set of limits of regular subgradients:

$$\partial a(p) = \left\{ q \;\middle|\; \begin{array}{c} \exists\, \{p_k\} \subset \mathrm{dom}\,(a) \text{ and } \{q_k\} \subset \mathcal{P}_n, \\ \text{such that } q_k \in \hat{\partial} a(p_k)\, \forall\, k = 1, 2, \dots, \\ p_k \to p \text{ and } q_k \to q \end{array} \right\}, \tag{3.35}$$

with $\partial a(p) = \emptyset$ if $p \notin \mathrm{dom}\, a = \mathcal{M}_n$. The set of *horizon subgradients*, denoted $\partial^\infty \alpha(p)$, is defined similarly, however, instead of $q_k \to q$ one has $t_k q_k \to q$ for some sequence of positive real numbers $\{t_k\}$ converging to zero. By convention, we have $\partial^\infty a(p) = \{0\}$ if $p \notin \mathrm{dom}\, a$. As in the case of regular subgradients, there is a relationship between these subgradients and the normal cone at a polynomial $p \in \mathcal{M}_n$ [RW98, Theorem 8.9]:

$$\partial a(p) = \left\{ q \;\middle|\; (q, -1) \in N_{\mathrm{epi}\,(a)}(p, a(p)) \right\}$$

and

$$\partial^\infty a(p) = \left\{ q \;\middle|\; (q, 0) \in N_{\mathrm{epi}\,(a)}(p, a(p)) \right\}.$$

Using these relationships, Theorem 3.2 and [RW98, Corollary 8.11] imply that

$$\partial a(p) = \hat{\partial} a(p) \quad \text{and} \quad \partial^\infty a(p) = \hat{\partial} a(p)^\infty \tag{3.36}$$

whenever $p \in \mathcal{M}_n$ (see (2.27) and (2.28)).

The subdifferential regularity of the abscissa mapping implies that it possesses a rich subdifferential calculus. For example, the following chain rule holds.

THEOREM 3.3 (see [RW98, Theorem 10.6]). *Let X be a finite dimensional Euclidean space, and suppose $G : X \to \mathcal{P}_n$ is continuously differentiable in the real sense. Consider the composite mapping $g = a \circ G$. If $x \in X$ is such that $G(x) \in \mathcal{M}_n$ and the only polynomial $q \in \partial^\infty a(G(x))$ with $\nabla G(x)^* q = 0$ is $q = 0$, then*

$$\partial g(x) = \nabla G(x)^* \partial a(G(x)), \qquad \partial^\infty g(x) = \nabla G(x)^* \partial^\infty a(G(x)),$$

and

$$dg(x)(d) = da(G(x))(\nabla G(x) d).$$

To illustrate these results, we apply Theorem 3.3 to the example studied in [BLO]. Let X be $\mathbb{C}^n$ with the standard real inner product, and consider the composition of the abscissa function with the affine mapping $G : \mathbb{C}^n \to \mathcal{P}_n$ given by

$$G(x) = (1 + x_0) e_{(n,0)} + x_1 e_{(n-1,0)} - \sum_{j=2}^{n} x_{j-1} e_{(n-j,0)}.$$

In [BLO, Theorem 2.1], it is shown that $x = 0$ is a strict global minimizer of the function $g = a \circ G$. Since G is affine, we have

$$\nabla G(0)d = d_0 e_{(n,0)} + d_1 e_{(n-1,0)} - \sum_{j=2}^{n} d_{j-1} e_{(n-j,0)},$$

and

$$\nabla G(0)^* \sum_{j=0}^{n} y_j e_{(n-j,0)} = (y_0, y_1 - y_2, -y_3, \ldots, -y_n).$$

The representation for $\partial^\infty a(e_{(n,0)})$ given by (3.36) and (2.26) shows that the only $q \in \partial^\infty a(e_{(n,0)})$ with $\nabla G(0)^* q = 0$ is $q = 0$. Therefore, Theorem 3.3 and the relations (3.36) and (2.25) show that

$$\hat{\partial} g(0) = \partial g(0) = \left\{ \left(z_0, z_1 - \frac{1}{n}, z_2, \ldots, z_{n-1} \right) \mid \operatorname{Re} z_1 \geq 0 \right\}.$$

Finally, observe that since the origin is in the interior of $\hat{\partial} g(0)$, we have, directly from the definition of regular subgradients, that $x = 0$ is a *sharp* minimizer of g in the sense that there exist $\epsilon > 0$ and $\kappa > 0$ such that

$$g(x) \geq g(0) + \kappa \|x\| \qquad \text{whenever} \quad \|x\| \leq \epsilon.$$

Further consequences of these results are explored in [BLO].

Acknowledgment. We gratefully acknowledge a referee for a careful reading of the paper and for several useful suggestions.

REFERENCES

[BLO] J.V. Burke, A.S. Lewis, and M.L. Overton, *Optimizing matrix stability*, Proc. Amer. Math. Soc., to appear.

[BO] J.V. Burke and M.L. Overton, *Variational analysis of non-Lipschitz spectral functions*, Math. Programming, to appear.

[Lev80] L.V. Levantovskii, *On singularities of the boundary of a stability region*, Moscow Univ. Math. Bull., 35 (1980), pp. 19–22.

[RW98] R.T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1677–1696

ON THE CONTROLLABILITY OF THE LINEARIZED BENJAMIN–BONA–MAHONY EQUATION[*]

SORIN MICU[†]

Abstract. We study the boundary controllability properties of the linearized Benjamin–Bona–Mahony equation

$$\begin{cases} u_t - u_{xxt} + u_x = 0, & x \in (0,1),\, t > 0, \\ u(t,0) = 0,\ u(1,t) = f(t), & t > 0. \end{cases}$$

We show that the equation is approximately controllable but not spectrally controllable (no finite linear combination of eigenfunctions, other than zero, is controllable). Next, we prove a finite controllability result and we estimate the norms of the controls needed in this case.

Key words. boundary control, moments, biorthogonals

AMS subject classifications. 93B05, 78M05

PII. S0363012999362499

1. Introduction.

The Benjamin–Bona–Mahony (BBM) equation

$$u_t + u_x + uu_x - u_{xxt} = 0, \tag{1.1}$$

like the Korteweg-de Vries (KdV) equation

$$u_t + u_x + uu_x + u_{xxx} = 0, \tag{1.2}$$

was originally derived as approximation for surface water waves in a uniform channel (see, for instance, [3], [4], and [5]).

Both (1.1) and (1.2) also cover cases of the following type: surface waves of long wavelength in liquids, acoustic-gravity waves in compressible fluids, hydromagnetic waves in cold plasma, acoustic waves in anharmonic crystals, etc. The wide applicability of these equations is the main reason why, during the last decades, they have attracted so much attention from mathematicians.

The main mathematical difference between KdV and BBM models can be most readily appreciated by comparing the dispersion relation for the respective linearized equations. It can be easily seen that these relations are comparable only for small wave numbers (i.e., long waves) and they generate drastically different responses to short waves (which are irrelevant to its role as a physical model). This is one of the reasons why, whereas existence and regularity theory for the KdV equation is difficult, the theory of the BBM equation is comparatively simple. The computing is also much easier for (1.1) than for (1.2).

The existence, uniqueness, and regularity of the BBM equation have been studied, for instance, in [7] and [18]. The large time behavior of the solutions of (1.1) was also intensively analyzed in the last decade (see, for instance, [1], [2], and [3]).

[*] Received by the editors October 20, 1999; accepted for publication November 6, 2000; published electronically February 28, 2001. This research was partially supported by grant PB96-0663 of DGES (Spain) and by grant 303/1999 of CNCSU (Romania).

http://www.siam.org/journals/sicon/39-6/36249.html

[†] Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Universidad Complutense, 28040 Madrid, Spain and Facultatea de Matematica-Informatica, Universitatea din Craiova, 1100, Romania (sorin@sunma4.mat.ucm.es).

Although it is generally considered that the BBM equation is easier to deal with than the KdV equation, it seems that, from the controllability point of view, (1.2) offers greater possibilities than (1.1). While important progress has been made in the last years for the KdV (see, for instance, [20], [24], and [21]), very little is known about the BBM. Some interior unique continuation properties for (1.1) and related problems (the linear case included) were studied in [9]. It is well known that, for the linear equation, by using the Hilbert uniqueness method due to J.-L. Lions (see [15]), the unique continuation property implies approximate controllability. Therefore, from [9], some approximate interior controllability results can be obtained for the linearized BBM equation. Nevertheless, the approximate controllability results for the nonlinear case do not seen to be entirely reducible to a unique continuation property and some estimates are needed on the dependence of the control function with respect to the perturbation introduced by the nonlinear term. In [23], (1.1) posed in $\mathbb{R}_+$ with boundary control is studied. It is proved that approximate controllability holds for the corresponding linear equation.

As far as we know there are no results for the controllability of the nonlinear BBM equation.

The present paper is concerned with the boundary controllability properties of the linearized BBM equation in finite domain. More precisely, given $T > 0$ and $u_0 \in H^{-1}(0,1)$ can we find a control function $f \in L^2(0,T)$ such that the solution u of

$$(1.3) \qquad \begin{cases} u_t - u_{xxt} + u_x = 0, & t > 0, \qquad x \in (0,1), \\ u(t,0) = 0,\ u(t,1) = f(t), & t > 0, \\ u(0,x) = u_0(x), & x \in (0,1), \end{cases}$$

satisfies

$$(1.4) \qquad u(T,x) = 0, \quad x \in (0,1)?$$

We shall first show that (1.3) is not spectrally controllable. This means that no finite linear nontrivial combination of eigenvectors can be driven to zero in finite time by using a control $f \in L^2(0,T)$.

Nevertheless, (1.3) is approximately controllable, i.e., the set of reachable states

$$R(T,u_0) = \{u(T,x) \mid f \in L^2(0,T)\}$$

is dense in $L^2(0,1)$ for any $u_0 \in H^{-1}(0,1)$ and $T > 0$. Hence, given $T > 0$, $u_0 \in H^{-1}(0,1)$, $v_0 \in H^{-1}(0,1)$, and $\varepsilon > 0$, there exists a control function $f \in L^2(0,T)$ such that the solution u of (1.3) satisfies $||u(T) - v_0||_{L^2(0,1)} < \varepsilon$.

These two results can be found at the beginning of the last section (Theorems 4.2 and 4.3).

We refer to [19] for similar negative results in the context of the exact controllability of the linear heat equation in a half-line.

Another interesting problem, with practical relevance, is the following finite controllability property: given $T > 0$, $N > 0$, and $u_0 \in H^{-1}(0,1)$, is there a control $f_N \in L^2(0,T)$ such that the projection of the solution of (1.3) over the finite dimensional space generated by the first $2N$ eigenvectors is equal to zero at $t = T$?

We give a positive answer to this question in Theorem 4.6. Moreover, by using some estimates for the corresponding biorthogonal sequences, we analyze how the norms of the controls change with N. We find an upper bound for the norms of the controls and we prove that this is, in some sense, sharp. More precisely, we prove that for any initial data $u_0 \in H^{-1}(0,1)$ there exists a control f_N such that

$$\| f_N \|_{L^2(0,T)} \leq c_1 e^{\gamma_1 N \ln(N)} \| u_0 \|_{H^{-1}(0,1)}, \tag{1.5}$$

where c_1 and γ_1 do not depend on N. Moreover, there are initial data $u_0 \in H_0^1(0,1)$ for which any corresponding control f_N satisfies

$$c_2 e^{\gamma_2 N \ln(N)} \| u_0 \|_{H_0^1(0,1)} \leq \| f_N \|_{L^2(0,T)}, \tag{1.6}$$

where c_2 and γ_2 do not depend on N.

We remark that the norms of the controls f_N may increase very rapidly as N goes to infinity. Hence, the cost needed to drive to zero the first $2N$ eigenmodes can be very high when N is large.

The controllability of the KdV equation has been studied in [20], [21], and [24]. It has been proved that exact controllability holds for the linearized equation with different boundary conditions and number of controls. Hence, a Sobolev space of initial data can be controlled from the boundary. This implies that the linearized KdV equation is not only spectrally controllable but also N-partially controllable with uniformly bounded controls. Therefore the boundary controllability properties of the linearized KdV are much "nicer" than the corresponding ones for the linearized BBM (which is not spectrally controllable and not uniformly N-partially controllable). We also remark that, based on the linear case, local or global controllability results (depending on the number of controls) can be obtained for the nonlinear KdV equation.

The paper is organized in the following way. In the second section we study the differential operator A corresponding to (1.3). We prove that A has a sequence of purely imaginary eigenvalues $(i\lambda_n)_{n\in\mathbb{Z}^*}$ such that $\lim_{|n|\to\infty} \lambda_n = 0$.

In the third section we analyze the biorthogonal sequences to the exponentials family $\{e^{i\lambda_n t}\}_{n\in\mathbb{Z}^*}$ or to a subset of it. First, we prove that there is no full biorthogonal sequence. Next we concentrate our attention on the finite families $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N \\ n\neq 0}}$. In this case various biorthogonal sequences can be constructed. We give an example and we analyze the behavior of the norms of the biorthogonals as N goes to infinity. The techniques used in this section combine classical elements from the theory of analytic functions with constructions specific to our problem.

Finally, in the last section, we use the previous results to solve the controllability problems mentioned above.

2. Linearized BBM equation: Elementary properties.

Let us consider the following equation

$$\begin{cases} u_t - u_{xxt} + u_x = 0, & x \in (0,1), \quad t > 0, \\ u(t,0) = u(t,1) = 0, & t > 0, \\ u(0,x) = u_0(x), & x \in (0,1), \end{cases} \tag{2.1}$$

representing the linearized BBM equation.

In order to put (2.1) in an abstract Cauchy form, we apply the operator $(\mathcal{I}-\partial_x^2)^{-1}$. The following equivalent equation is obtained:

$$\begin{cases} u_t + Au = 0, & x \in (0,1), \quad t > 0, \\ u(0) = u_0, & x \in (0,1), \end{cases} \tag{2.2}$$

where $A : H_0^1(0,1) \to H_0^1(0,1)$ is given by

$$Au = (\mathcal{I} - \partial_x^2)^{-1} \partial_x u. \tag{2.3}$$

Here ∂_x^2 denote the Laplace operator

$$\partial_x^2 : H^2(0,1) \cap H_0^1(0,1) \to L^2(0,1), \quad \partial_x^2 u = u_{xx}.$$

The main properties of the operator A are given in the following proposition.

PROPOSITION 2.1. *A is a compact, skew-adjoint operator in $H_0^1(0,1)$.*

Proof. Due to the regularizing effect of the operator $(\mathcal{I} - \partial_x^2)^{-1}$ it follows immediately that A takes values in $H^2(0,1) \cap H_0^1(0,1)$ which is compactly embedded in $H_0^1(0,1)$. Hence A is compact.

Let us consider in $H_0^1(0,1)$ the inner product given by

$$(u,v) = \int_0^1 \partial_x u \partial_x v + \int_0^1 uv. \tag{2.4}$$

For any $u, v \in H^2(0,1) \cap H_0^1(0,1)$, we obtain

$$(Au,v) = \left(\left(\mathcal{I} - \partial_x^2\right)^{-1} \partial_x u, v \right) = \int_0^1 \partial_x \left[\left(\mathcal{I} - \partial_x^2\right)^{-1} \partial_x u \right] \partial_x v + \int_0^1 \left[\left(\mathcal{I} - \partial_x^2\right)^{-1} \partial_x u \right] v$$

$$= \int_0^1 \left(\mathcal{I} - \partial_x^2\right)^{-1} \left(\partial_x^2 u\right) \partial_x v - \int_0^1 \left[\left(\mathcal{I} - \partial_x^2\right)^{-1} u \right] \partial_x v = - \int_0^1 u(\partial_x v)$$

$$= \int_0^1 (\partial_x u) v = - \int_0^1 \partial_x u \left(\mathcal{I} - \partial_x^2\right)^{-1} \left(\partial_x^2 v\right) + \int_0^1 \partial_x u \left[\left(\mathcal{I} - \partial_x^2\right)^{-1} v \right]$$

$$= - \int_0^1 \partial_x u \partial_x \left[\left(\mathcal{I} - \partial_x^2\right)^{-1} \partial_x v \right] - \int_0^1 u \left[\left(\mathcal{I} - \partial_x^2\right)^{-1} \partial_x v \right] = - \left(u, \left(\mathcal{I} - \partial_x^2\right)^{-1} \partial_x v \right)$$

$$= -(u, Av).$$

By density we obtain that $(Au,v) = -(u,Av) \quad \forall u, v \in H_0^1(0,1)$ and therefore A is skew-adjoint in $H_0^1(0,1)$. □

Since A is compact, (2.2) can be treated like an ordinary differential equation in the Hilbert space $H_0^1(0,1)$. By using Cauchy–Lipschitz–Picard theorem the following properties concerning the solutions of (2.2) are immediate.

PROPOSITION 2.2. *Equation* (2.2) *has a unique solution $u \in C^1\left([0,\infty); H_0^1(0,1)\right)$ which satisfies*

$$\int_0^1 \mid \partial_x u(x,t) \mid^2 + \int_0^1 \mid u(x,t) \mid^2 = \int_0^1 \mid \partial_x u_0 \mid^2 + \int_0^1 \mid u_0 \mid^2 . \tag{2.5}$$

Proof. Since A is a bounded linear operator the existence and uniqueness of solutions follow from Cauchy–Lipschitz–Picard theorem (see [8, p. 104]).

On the other hand, since A is skew-adjoint, we have

$$\frac{1}{2} \frac{d}{dt} \parallel u \parallel_{H_0^1}^2 = Re(u, u_t) = Re(u, -Au) = 0.$$

Hence, the H_0^1 norm of the solution is conserved. □

Remark 2.1. *In fact much more can be said about the regularity of solutions of* (2.2). *Since* (2.2) *is linear and A is a bounded operator we can easily deduce that $u \in C^\omega\left([0,\infty);\, H_0^1(0,1)\right)$, where $C^\omega\left([0,\infty);\, H_0^1(0,1)\right)$ represents the class of analytic functions defined in $[0,\infty)$ with values in $H_0^1(0,1)$. Indeed, for $t_0 \in [0,\infty)$,*

$$\left\|\sum_{n=0}^{\infty} u^{(n)}(t_0)\frac{(t-t_0)^n}{n!}\right\|_{H_0^1} \le \sum_{n=0}^{\infty} \frac{|t-t_0|^n}{n!}\left\|u^{(n)}(t_0)\right\|_{H_0^1}$$

$$\le \sum_{n=0}^{\infty} \frac{|t-t_0|^n}{n!}\,||A||^n\,||u(t_0)||_{H_0^1} < \infty.$$

Hence the series $\sum_{n=0}^{\infty} u^{(n)}(t_0)\frac{(t-t_0)^n}{n!}$ is (absolutely) convergent and

$$u(t) = \exp\left(-A(t-t_0)\right)u(t_0) = \sum_{n=0}^{\infty}(-1)^n\frac{(t-t_0)^n}{n!}A^n u(t_0) = \sum_{n=0}^{\infty}\frac{(t-t_0)^n}{n!}u^{(n)}(t_0).$$

Our next objective is to express the solution u of (2.2) in Fourier series. To do so we need the spectral decomposition of the operator A.

Proposition 2.3. *A has a sequence of purely imaginary eigenvalues $(\mu_n)_{n\in\mathbb{Z}^*}$,*

$$\mu_n = \operatorname{sgn}(n)\frac{i}{2\sqrt{1+n^2\pi^2}}, \quad n \in \mathbb{Z}^*. \tag{2.6}$$

Moreover, to each eigenvalue μ_n corresponds an unique eigenfunction U_n,

$$U_n(x) = \frac{1}{\sqrt{n^2\pi^2+1}}e^{-i\operatorname{sgn}(n)\sqrt{n^2\pi^2+1}x}\sin(n\pi x), \quad n \in \mathbb{Z}^*, \tag{2.7}$$

such that $\| U_n \|_{H_0^1}= 1$. The family $(U_n)_{n\in\mathbb{Z}^*}$ forms an orthonormal basis in $H_0^1(0,1)$.

Proof. We are looking for $\mu \in \mathbb{C}$ and $\tau \in H_0^1(0,1)$ such that $A\tau = \mu\tau$, which is equivalent to

$$\begin{cases} \mu\tau - \mu\tau_{xx} - \tau_x = 0, \\ \tau(0) = \tau(1) = 0. \end{cases} \tag{2.8}$$

Hence, $\tau(x) = c_1 e^{x\frac{-1-\sqrt{1+4\mu^2}}{2\mu}} + c_2 e^{x\frac{-1+\sqrt{1+4\mu^2}}{2\mu}}$.

From the boundary conditions we obtain, from one hand, that $c_1 = -c_2$ and from the other hand, that the eigenvalues of the operator are given by the equation

$$e^{\frac{\sqrt{1+4\mu^2}}{\mu}} = 1. \tag{2.9}$$

It results that the eigenvalues of the operator A are

$$\mu_n = \operatorname{sgn}(n)\frac{i}{2\sqrt{1+n^2\pi^2}}, \quad n \in \mathbb{Z}^*.$$

To each μ_n corresponds an eigenfunction

$$U_n(x) = \frac{1}{\sqrt{n^2\pi^2+1}}e^{-i\operatorname{sgn}(n)\sqrt{1+n^2\pi^2}x}\sin(n\pi x)$$

with $||U_n||_{H_0^1} = 1.$ □

REMARK 2.2. *Let us remark that, for each $n \in \mathbb{Z}^*$,*

$$(U_n)_x(1) = \frac{(-1)^n n\pi}{\sqrt{n^2\pi^2+1}} e^{-i\,\mathrm{sgn}(n)\sqrt{1+n^2\pi^2}} \neq 0$$

and $|(U_n)_x(1)| \sim 1$ as $|n| \to \infty$.

REMARK 2.3. *We have obtained that $\lim_{|n|\to\infty} \lambda_n = 0$. This is due to the compactness of the operator A and will have some very important consequences for the controllability properties of the BBM equation.*

If we consider an initial data, $u_0 \in H_0^1(0,1)$, $u_0 = \sum_{n\in\mathbb{Z}^*} a_n U_n$, the solution of (2.2) corresponding to this initial data can be written as

$$u = \sum_{n\in\mathbb{Z}^*} a_n U_n e^{i\lambda_n t},$$

where $\lambda_n = \frac{\mathrm{sgn}(n)}{2\sqrt{1+n^2\pi^2}}$ and $\mu_n = i\lambda_n$ are the eigenvalues of the operator A found in Proposition 2.3.

3. Biorthogonal sequences.

Let $\lambda_n i$, $n \in \mathbb{Z}^*$, be the eigenvalues of the operator A. In this section we study the sequences biorthogonal to the family of exponentials $\{e^{i\lambda_n t}\}_{n\in\mathbb{Z}^*}$ or to some subset of it. All the results of this section will be used to study the boundary controllability properties of the BBM equation in the last section.

Let us first recall the following definition.

DEFINITION 3.1. *Let $(f_n)_{n\geq 1}$ be a sequence of vectors from a Hilbert space H. The sequence $(g_n)_{n\geq 1} \subset H$ is biorthogonal to $(f_n)_{n\geq 1}$ if and only if $(f_n, g_m) = \delta_{nm}$ $\forall n, m \geq 1$.*

We begin with the following negative result.

THEOREM 3.2. *Let $T > 0$ and $m \in \mathbb{Z}^*$. There is no function $\Theta_m \in L^2(-T,T)$ such that*

$$\int_{-T}^{T} \Theta_m(t) e^{i\lambda_n t} dt = \begin{cases} 0 & \text{if } n \in \mathbb{Z}^*,\ n \neq m, \\ 1 & \text{if } n = m. \end{cases} \tag{3.1}$$

Proof. Let us suppose that there exists a function $\Theta_m \in L^2(-T,T)$ such that (3.1) is satisfied.

We define $F : \mathbb{C} \to \mathbb{C}$ by

$$F(z) = \int_{-T}^{T} \Theta_m(t) e^{izt} dt. \tag{3.2}$$

From the Paley–Wiener theorem, F is an entire function. Moreover, from (3.2) and (3.1) we obtain that

$$F(\lambda_n) = \delta_{nm} \quad \forall n \in \mathbb{Z}^*. \tag{3.3}$$

Since $\lim_{n\to\infty} \lambda_n = 0$, it follows that F is zero on a set with a finite accumulation point. Therefore $F = 0$ which contradicts the fact that $F(\lambda_m) = 1$.

Hence, there is no function $\Theta_m \in L^2(-T,T)$ such that (3.1) is satisfied and the proof ends. □

REMARK 3.1. *From Theorem* 3.2 *the following result of nonobservability can be obtained: there is no sequence* $(\rho_n)_{n\in\mathbb{Z}^*}$ *of positive constants such that the following inequality*

$$\sum_{n\in\mathbb{Z}^*}\rho_n \mid a_n \mid^2 \leq \int_{-T}^{T}\left|\sum_{n\in\mathbb{Z}^*}a_n e^{i\lambda_n t}\right|^2 dt \tag{3.4}$$

is true for any sequence $(a_n)_{n\in\mathbb{Z}^*}$ *with a finite number of nonzero terms.*

This is a direct consequence of Theorem 3.2 *and the following result in moments theory (see* [22, p. 151]*).*

THEOREM A. *Let* H *be a Hilbert space,* $(f_n)_n$ *a vector family from* H*, and* $(c_n)_n$ *a sequence of scalars. In order for a vector* $f\in H$ *to exist such that* $\parallel f \parallel\leq M$ *and* $(f, f_n) = c_n$ $\forall n$*, it is necessary and sufficient that*

$$\left|\sum_n a_n\bar{c}_n\right| \leq M\left\|\sum_n a_n f_n\right\|^2 \tag{3.5}$$

for any finite number of scalars $a_1, a_2, \ldots$.

Let us suppose now that (3.4) *is true. Then, for each* $m\in\mathbb{Z}^*$,

$$|a_m|^2 \leq \frac{1}{\rho_m}\int_{-T}^{T}\left|\sum_{n\in\mathbb{Z}^*}a_n e^{i\lambda_n t}\right|^2 dt$$

for any sequence $(a_n)_{n\in\mathbb{Z}^*}$ *with a finite number of nonzero terms.*

From Theorem A *it follows that there exists* $\Theta_m \in L^2(-T,T)$ *such that* $\int_{-T}^{T}\Theta_m(t)e^{i\lambda_n t}dt = \delta_{mn}$ *which contradicts Theorem* 3.2.

REMARK 3.2. *Theorem* 3.2 *proves that there is no sequence biorthogonal to* $\{e^{i\lambda_n t}\}_{n\in\mathbb{Z}^*}$ *in* $L^2(-T,T)$*. This is related to the fact that* $\lim_{|n|\to\infty}\lambda_n = 0$ *which affects the linear independency of the exponential family* $\{e^{i\lambda_n t}\}_{n\in\mathbb{Z}^*}$ *in* $L^2(-T,T)$*.*

We shall use this result in the last section to prove that no eigenfunction of equation (1.3) *can be driven to zero by using a control function in* $L^2(0,T)$ *(see Theorem* 4.2*).*

Let $N\in\mathbb{N}^*$. We shall pass now to prove the existence of a biorthogonal to the finite family of exponentials $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N\\ n\neq 0}}$.

THEOREM 3.3. *Let* $T > 0$ *and* $N\in\mathbb{N}^*$*. There exists a biorthogonal sequence* $\{\Psi_m\}_{\substack{|n|\leq N\\ n\neq 0}}$ *to the family of exponentials* $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N\\ n\neq 0}}$ *in* $L^2(-T,T)$.

Proof. Let us first prove that there is a constant $C_1(N) > 0$ such that, for any scalars $(a_n)_{\substack{|n|\leq N\\ n\neq 0}}$,

$$C_1(N)\sum_{\substack{|n|\leq N\\ n\neq 0}} \mid a_n \mid^2 \leq \int_{-T}^{T}\left|\sum_{\substack{|n|\leq N\\ n\neq 0}}a_n e^{i\lambda_n t}\right|^2 dt. \tag{3.6}$$

We consider the space generated by $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N\\ n\neq 0}}$,

$$X = \underset{L^2(-T,T)}{\text{Span}}\left\{e^{i\lambda_n t}\right\}_{\substack{|n|\leq N\\ n\neq 0}}.$$

X is a finite-dimensional space of dimension $2N$. Moreover, the application

$$\sum_{\substack{|n|\leq N\\ n\neq 0}} a_n e^{i\lambda_n t}\in X \longrightarrow \sqrt{\sum_{\substack{|n|\leq N\\ n\neq 0}} \mid a_n \mid^2}$$

is a norm in X. Since X is finite dimensional this new norm and the one induced from $L^2(T,T)$ are equivalent. It follows that there is a constant $C_1(N)$ such that (3.6) is satisfied for any scalars (a_n).

Now, for each m, $\mid m \mid\leq N$, $m\neq 0$, we can apply Theorem A from Remark 3.1 by taking $c_n=\delta_{nm}$, $f_n=e^{i\lambda_n t}$, and $H=L^2(-T,T)$. It follows that there exists a function $\Psi_m\in L^2(-T,T)$ such that $\int_{-T}^{T}\Psi_m(t)e^{i\lambda_n t}dt=\delta_{nm}$ $\forall n$, $\mid n\mid\leq N$, $n\neq 0$.

Hence we get a biorthogonal sequence $\{\Psi_m\}_{\substack{|m|\leq N\\ m\neq 0}}\subset L^2(-T,T)$ to the family of exponentials $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N\\ n\neq 0}}$ and the proof finishes. □

REMARK 3.3. *The following inequality is also true:*

$$\int_{-T}^{T}\left|\sum_{\substack{|n|\leq N\\ n\neq 0}} a_n e^{i\lambda_n t}\right|^2 dt\leq 4NT\sum_{\substack{|n|\leq N\\ n\neq 0}}\mid a_n\mid^2 . \tag{3.7}$$

Indeed, from the Cauchy inequality

$$\int_{-T}^{T}\left|\sum_{\substack{|n|\leq N\\ n\neq 0}} a_n e^{i\lambda_n t}\right|^2 dt\leq\int_{-T}^{T}\left(\sum_{\substack{|n|\leq N\\ n\neq 0}}\mid a_n\mid^2\right)\left(\sum_{\substack{|n|\leq N\\ n\neq 0}}\mid e^{i\lambda_n t}\mid^2\right)dt=4NT\sum_{\substack{|n|\leq N\\ n\neq 0}}\mid a_n\mid^2$$

and (3.7) *is proved.*

REMARK 3.4. *The proof of Theorem* 3.3 *shows that there exists at least a biorthogonal sequence to any finite family of exponentials.*

REMARK 3.5. *From Theorem* A *(Remark* 3.1*) we also obtain that the norm of the biorthogonal sequence* $\{\Psi_m\}_{\substack{|m|\leq N\\ m\neq 0}}$ *is bounded by* $C_1(N)$. *Since Theorem* 3.2 *proves that there is no biorthogonal sequence to* $\{e^{i\lambda_n t}\}_{n\in\mathbb{Z}^*}$ *we deduce again from Theorem* A *that* $C_1(N)$ *degenerates when* N *goes to infinity. We shall analyze how this constant changes when* N *increases.*

THEOREM 3.4. *Let* $T>0$ *and* $N\in\mathbb{N}^*$. *There exists a biorthogonal sequence* $\{\Theta\}_{\substack{|n|\leq N\\ n\neq 0}}$ *to the family of exponentials* $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N\\ n\neq 0}}$ *in* $L^2(-T,T)$ *such that*

$$\parallel\Theta_m\parallel^2_{L^2(-T,T)}\leq C_1 e^{\alpha N\ln(N)}, \tag{3.8}$$

where C_1 *and* α *are two constants which do not depend on* N.

Proof. Let us first define, for each m such that $\mid m\mid\leq N$ and $m\neq 0$,

$$\xi_m(z)=\left(\prod_{\substack{|n|\leq N\\ n\neq 0,m}}\frac{z-\lambda_n}{\lambda_m-\lambda_n}\right)\left(\frac{\sin\frac{T(z-\lambda_m)}{2N}}{\frac{T(z-\lambda_m)}{2N}}\right)^{2N}. \tag{3.9}$$

Each function ξ_m has the following properties:

- ξ_m is an entire function,

- $\xi_m(\lambda_n) = \delta_{nm}$,
- $\xi_m(x) \in L^2(-\infty,\infty)$,
- ξ_m is of the exponential type at most T, i.e., there exists a constant $A > 0$ such that $\forall \varepsilon > 0$, we have

$$|\xi_m(z)| \leq Ae^{(T+\varepsilon)|z|} \quad \forall z \in \mathbb{C}.$$

Let us now define

$$\Theta_m(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \xi_m(x) e^{-ixt} dx, \tag{3.10}$$

and we shall show that $\{\Theta_m\}_{\substack{|n|\leq N \\ n\neq 0}}$ is the biorthogonal sequence we are looking for.

From the properties of ξ_m, by using Paley–Wiener theorem, it follows that Θ_m has compact support in $[-T,T]$, it belongs to $L^2(-T,T)$, and

$$\int_{-T}^{T} \Theta_m(t) e^{i\lambda_n t} dt = \xi_m(\lambda_n) = \delta_{nm}.$$

It follows that $(\Theta_m)_{\substack{|m|\leq N \\ m\neq 0}}$ is a biorthogonal sequence to $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N \\ n\neq 0}}$.

Our next objective is to estimate the norm of Θ_m and to see that it satisfies (3.8).

From Plancherel's theorem we have that

$$\| \Theta_m \|_{L^2(-T,T)} = \| \xi_m \|_{L^2(-\infty,\infty)}. \tag{3.11}$$

Let us now estimate $\| \xi_m \|_{L^2(-\infty,\infty)}$.

$$\| \xi_m \|^2_{L^2(-\infty,\infty)} = \int_{-\infty}^{\infty} \left| \left(\prod_{\substack{|n|\leq N \\ n\neq 0,m}} \frac{x-\lambda_n}{\lambda_m-\lambda_n} \right) \left(\frac{\sin \frac{T(x-\lambda_m)}{2N}}{\frac{T(x-\lambda_m)}{2N}} \right)^{2N} \right|^2$$

$$= \left(\prod_{\substack{|n|\leq N \\ n\neq 0,m}} \frac{1}{|\lambda_n-\lambda_m|^2} \right) \int_{-\infty}^{\infty} \left| \left(\prod_{\substack{|n|\leq N \\ n\neq 0,m}} (x-\lambda_n) \right) \left(\frac{\sin \frac{T(x-\lambda_m)}{2N}}{\frac{T(x-\lambda_m)}{2N}} \right)^{2N} \right|^2 dx.$$

Let us first evaluate the constant

$$\gamma_1(N) = \prod_{\substack{|n|\leq N \\ n\neq 0,m}} \frac{1}{|\lambda_m-\lambda_n|^2}.$$

We have

$$\frac{1}{|\lambda_m-\lambda_n|} = \frac{1}{\left| \frac{\operatorname{sgn}(n)}{\sqrt{1+n^2\pi^2}} - \frac{\operatorname{sgn}(m)}{\sqrt{1+m^2\pi^2}} \right|} = \frac{2\sqrt{1+n^2\pi^2}\sqrt{1+m^2\pi^2}}{\left|\operatorname{sgn}(n)\sqrt{1+m^2\pi^2} - \operatorname{sgn}(m)\sqrt{1+n^2\pi^2}\right|}$$

$$\leq \frac{2\sqrt{1+n^2\pi^2}\sqrt{1+m^2\pi^2}}{\left|\sqrt{1+m^2\pi^2}-\sqrt{1+n^2\pi^2}\right|} = \frac{2\sqrt{1+n^2\pi^2}\sqrt{1+m^2\pi^2}\left(\sqrt{1+m^2\pi^2}+\sqrt{1+n^2\pi^2}\right)}{|m^2-n^2|\,\pi^2}$$

$$\leq \frac{2\sqrt{(1+n^2\pi^2)(1+m^2\pi^2)}\left(\frac{\pi^2}{2}m + \frac{\pi^2}{2}n\right)}{|m^2-n^2|\,\pi^2} \leq \sqrt{(1+n^2\pi^2)(1+m^2\pi^2)} \leq (2\pi N)^2.$$

It follows that

$$\gamma_1(N) = \prod_{\substack{|n|\leq N\\ n\neq 0,m}} \frac{1}{\mid \lambda_m - \lambda_n \mid^2} \leq (2\pi N)^{8N-4}. \tag{3.12}$$

Let us now evaluate the integral

$$\gamma_2(N) = \int_{-\infty}^{\infty} \left| \left(\prod_{\substack{|n|\leq N\\ n\neq 0,m}} (x-\lambda_n) \right) \left(\frac{\sin \frac{T(x-\lambda_m)}{2N}}{\frac{T(x-\lambda_m)}{2N}} \right)^{2N} \right|^2 dx.$$

We have

$$\gamma_2(N) = \int_{|x|\leq \frac{1}{2}} \left| \prod_{\substack{|n|\leq N\\ n\neq 0,m}} (x-\lambda_n) \right|^2 \left| \frac{\sin \frac{T(x-\lambda_m)}{2N}}{\frac{T(x-\lambda_m)}{N}} \right|^{4N} dx$$

$$+ \int_{|x|\geq \frac{1}{2}} \left| \prod_{\substack{|n|\leq N\\ n\neq 0,m}} (x-\lambda_n) \right|^2 \left| \frac{\sin \frac{T(x-\lambda_m)}{2N}}{\frac{T(x-\lambda_m)}{2N}} \right|^{4N} dx$$

$$\leq \int_{|x|\leq \frac{1}{2}} \left| \prod_{\substack{|n|\leq N\\ n\neq 0,m}} (x-\lambda_n) \right|^2 dx + \int_{|x|\geq \frac{1}{2}} \left| \prod_{\substack{|n|\leq N\\ n\neq 0,m}} (x-\lambda_n) \right|^2 \left(\frac{2N}{T(x-\lambda_m)} \right)^{4N} dx.$$

However, since $|\lambda_n| < \frac{1}{2}$,

$$\left| \prod_{\substack{|n|\leq N\\ n\neq 0,m}} (x-\lambda_n) \right| \leq \begin{cases} 1 & \text{if } |x| \leq \frac{1}{2}, \\ (2 \mid x \mid)^{2N-1} & \text{if } \mid x \mid \geq \frac{1}{2}. \end{cases}$$

It follows that

$$\gamma_2(N) \leq 1 + \int_{|x|\geq \frac{1}{2}} \frac{2^{4N-2}}{\mid x \mid^2} \left| \frac{x}{x-\lambda_m} \right|^{4N} 2^{4N} \left(\frac{N}{T} \right)^{4N} dx$$

$$\leq 1 + 2^{12N-2} \left(\frac{N}{T} \right)^{4N} \int_{|x|\geq \frac{1}{2}} \frac{1}{\mid x \mid^2} = 1 + 2^{12N} \left(\frac{N}{T} \right)^{4N}.$$

Hence,

$$\gamma_2(N) \leq 1 + 2^{12N} \left(\frac{N}{T} \right)^{4N}. \tag{3.13}$$

From (3.12) and (3.13) it follows that, for N large enough,

$$\parallel \xi_m \parallel_{L^2(-\infty,\infty)} \leq C_1 N^{\alpha N}, \tag{3.14}$$

where $C_1 > 0$ and $\alpha > 3$ are two constants which do not depend on N.

From (3.11) it follows that

$$\| \Theta_m \|_{L^2(-T,T)} \leq C_1 N^{\alpha N}$$

and (3.8) is obtained. □

REMARK 3.6. *In Theorem* 3.4 *we construct an explicit biorthogonal sequence which norm increases as* $\exp(\alpha N \ln(N))$ *as* $N \to \infty$. *Nevertheless, many other biorthogonals can be found. What can be said about the norms of these biorthogonals? We shall prove in the next theorem that the norm of any biorthogonal to* $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N\\ n\neq 0}}$ *is bounded from below by a constant of the type* $\exp(\beta N \ln(N))$. *In this sense,* (3.8) *is sharp.*

THEOREM 3.5. *Let* $(\psi_n)_{\substack{|n|\leq N\\ n\neq 0}}$ *be biorthogonal to* $\{e^{i\lambda_n t}\}_{\substack{|n|\leq N\\ n\neq 0}}$ *in* $L^2(-T,T)$. *Then there exist two positive constants* C_2 *and* ω, *not depending on* N, *such that*

$$\| \psi_m \|_{L^2(-T,T)} \geq C_2 e^{\omega N \ln(N)} \tag{3.15}$$

$\forall m \neq 0$ *such that* $| m | \leq N$.

Proof. In order to prove the theorem some arguments from [11] will be used. We shall give the proof in several steps.

Step 1. Let us define the following sequence of functions:

$$\tau_m(z) = \int_{-T}^{T} \psi_m(t) e^{itz} dt, \quad | m | \leq N, \quad m \neq 0. \tag{3.16}$$

From the Paley–Wiener theorem it follows that τ_m is an entire function of exponential type at most T. Moreover,

$$| \tau_m(x) | \leq \sqrt{2T} \, \| \psi_m \|_{L^2(-T,T)} \quad \forall x \in \mathbb{R}. \tag{3.17}$$

Since τ_m is a function of exponential type it follows from Hadamard's factorization theorem that

$$\tau_m(z) = a z^p e^{bz} \prod_{z_k \in E} \left(1 - \frac{z}{z_k}\right) e^{z/z_k}, \tag{3.18}$$

where E is the set of the zeros z_k of τ_m with $z_k \neq 0$, $E = \{z_k \in \mathbb{C} \mid \tau_m(z_k) = 0, \quad z_k \neq 0\}$.

From the definition of the function τ_m it follows that $\tau_m(\lambda_n) = \delta_{m,n}$. Therefore $\{\lambda_n : | n | \leq N, n \neq 0, n \neq m\} \subseteq E$. Let $E' = \{\lambda_n : | n | \leq N, n \neq 0, n \neq \pm m\}$ and define the polynomial function

$$P_m(z) = \prod_{\substack{|n|\leq N\\ n\neq 0, \pm m}} \frac{\lambda_n - z}{\lambda_n - \lambda_m}. \tag{3.19}$$

Let us now define function $\phi_m(z)$ by

$$\phi_m(z) = \frac{\tau_m(z)}{P_m(z)}. \tag{3.20}$$

The function ϕ_m has the following properties:

- it is an entire function of exponential type at most T,
- $\phi_m(\lambda_m) = 1$,
- $\tau_m(z) = P_m(z)\phi_m(z)$.

Step 2. In this step we shall give some estimates for $\mid P_m(z) \mid$.

$$\mid P_m(z) \mid = \left| \prod_{\substack{|n|\leq N \\ n\neq 0,\pm m}} \frac{\lambda_n - z}{\lambda_n - \lambda_m} \right| = \left(\prod_{\substack{|n|\leq N \\ n\neq 0,\pm m}} \mid \lambda_n - z \mid \right) \left[\prod_{\substack{|n|\leq N \\ n\neq 0,\pm m}} \mid \lambda_n - \lambda_m \mid \right]^{-1}.$$

By taking $z \in \mathbb{C}$ such that $\mid z \mid \geq 2$ we obtain that

$$\prod_{\substack{|n|\leq N \\ n\neq 0,\pm m}} \mid \lambda_n - z \mid \geq (\mid z \mid -1)^{2N-2}. \tag{3.21}$$

On the other hand

$$\prod_{\substack{|n|\leq N \\ n\neq 0,\pm m}} \mid \lambda_n - \lambda_m \mid \leq \prod_{\substack{|n|\leq N \\ n\neq 0,\pm m}} \frac{1}{2}\left(\frac{1}{\sqrt{1+n^2\pi^2}} + \frac{1}{\sqrt{1+m^2\pi^2}} \right) \leq 1. \tag{3.22}$$

From (3.21) and (3.22) we deduce that

$$\mid P_m(z) \mid \geq (\mid z \mid -1)^{2N-2} \quad \forall z \in \mathbb{C},\ \mid z \mid \geq 2. \tag{3.23}$$

Step 3. From (3.17) and (3.23) we obtain that

$$\mid \phi_m(z) \mid = \frac{\mid \tau_m(z) \mid}{\mid P_m(z) \mid} \leq \frac{\sqrt{2T} e^{T \operatorname{Im} z} \parallel \psi_m \parallel_{L^2(-T,T)}}{(|z|-1)^{2N-2}} \quad \forall z \in \mathbb{C},\ |z| \geq 2. \tag{3.24}$$

It follows that

$$\mid \phi_m(x) \mid \leq \sqrt{2T} \parallel \psi_m \parallel_{L^2(-T,T)} \frac{1}{(\mid x \mid -1)^{2N-2}} \quad \forall x \in \mathbb{R},\ \mid x \mid \geq 2. \tag{3.25}$$

We shall show that (3.25) is not possible unless $\parallel \psi_m \parallel$ grows rapidly with N.

Let us first recall the following result (see [14, p. 21] and [6, p. 52]).

THEOREM B. *Let $f(z)$ be holomorphic in the circle $\mid z \mid \leq 2eR$ $(R > 0)$ with $f(0) = 1$ and let $\eta \in (0, \frac{3e}{2})$. Then inside the circle $\mid z \mid \leq R$, but outside of a family of excluded circles the sum of whose radii is not greater than $4\eta R$, we have*

$$\ln(\mid f(z) \mid) > -\left(2 + \ln\left(\frac{3e}{2\eta} \right) \right) \ln(M_f(2eR)), \tag{3.26}$$

where $M_f(2eR) = \max\limits_{|z|=2eR} \mid f(z) \mid$.

We apply this result to our case. Let us define $\varphi_m : \mathbb{C} \to \mathbb{C}, \quad \varphi_m(z) = \phi_m(\lambda_m - z)$.

Evidently, φ_m is an entire function such that $\varphi_m(0) = 1$. Hence, φ_m satisfies the hypothesis of Theorem B. It follows that, $\forall R > 0$ and $\eta \in (0, \frac{3e}{2})$,

$$\ln(\mid \varphi_m(z) \mid) > -2e\left(2 + \ln\left(\frac{3e}{2\eta} \right) \right) \ln(M_{\varphi_m}(2eR)) \quad \forall z \in \mathbb{C},\ \mid z \mid \leq R, \tag{3.27}$$

outside of a set of circles the sum of whose radii is not greater than $4\eta R$.

Let us denote $\delta = 2e(2+\ln(\frac{3e}{2\eta})) > 1$. Also, remark that, from (3.24),

$$M_{\varphi_m}(2eR) \le e^{4eRT}||\psi_m||_{L^2(-T,T)}$$

if $2eR \ge 2$.

Hence, $\forall R > 0$ and $\eta \in (0, \frac{3e}{2})$ such that $2eR \ge 2$,

$$\ln(|\ \varphi_m(z)\ |) > -\delta \ln\left(e^{4eRT}||\psi_m||_{L^2(-T,T)}\right) \quad \forall z \in \mathbb{C},\ |\ z\ |\le R, \tag{3.28}$$

outside of a set of circles the sum of whose radii is not greater than $4\eta R$.

Let us consider $R > 6$ and $\eta \in (0, \frac{1}{8})$.

It follows that there exists $x_0 \in [\frac{R}{2}, R]$ such that

$$\ln(|\ \varphi_m(x_0)\ |) > -\delta \ln\left(e^{4eRT}||\psi_m||_{L^2(-T,T)}\right). \tag{3.29}$$

On the other hand, from (3.25),

$$|\varphi_m(x_0)| = |\phi_m(\lambda_m - x_0)| \le \sqrt{2T}\ \|\ \psi_m\ \|_{L^2(-T,T)}\ \frac{1}{(|\ \lambda_m - x_0\ | -1)^{2N-2}}. \tag{3.30}$$

From (3.30) and (3.29) the following estimate is obtained:

$$\ln\left(\sqrt{2T}\ \|\ \psi_m\ \|_{L^2(-T,T)}\ \frac{1}{(|\ \lambda_m - x_0\ | -1)^{2N-2}}\right) > -\delta \ln\left(e^{4eRT}||\psi_m||_{L^2(-T,T)}\right).$$

Hence

$$(1+\delta)\ln\left(||\psi_m||_{L^2(-T,T)}\right) > -4e\delta TR - \ln(\sqrt{2T}) + (2N-2)\ln(|x_0-\lambda_m|-1). \tag{3.31}$$

Let us now analyze the expression

$$G(N, x_0, R) = (2N-2)\ln(|\ \lambda_m - x_0\ | -1) - 4e\delta TR.$$

Remark that, for $R = N > 6$,

$$\begin{aligned} G(N, x_0, R) &\ge (2N-2)\ln(|\ x_0\ | - |\ \lambda_m\ | -1) - 4e\delta TN \\ &\ge (2N-2)\ln\left(\frac{N}{2}-2\right) - 4e\delta TN \\ &= 2N\left(\frac{N-1}{N}\ln\left(\frac{N}{2}-2\right) - \underbrace{2e\delta T}_{\text{cte}}\right). \end{aligned}$$

It follows that there exists $\omega > 0$, not depending on N, such that

$$G(N, x_0, R) \ge \omega N \ln(N) \tag{3.32}$$

for any N sufficiently large.

From (3.31) it follows that

$$\ln\left(||\psi_m||_{L^2(-T,T)}\right) > -\frac{\ln(\sqrt{2T})}{1+\delta} + \omega N \ln(N)$$

and the proof finishes. □

4. Controllability results. In this section we study some boundary controllability properties of the BBM equation. We begin with the following exact controllability problem: given $T < 0$ and an initial data $u_0 \in H^{-1}(0,1)$ find a control $f \in L^2(0,T)$ such that the solution u of

$$\left\{\begin{array}{lll} u_t - u_{xxt} + u_x = 0, & x \in (0,1), & t > 0, \\ u(t,0) = 0, u(t,1) = f(t), & t > 0, & \\ u(0,x) = u_0(x), & x \in (0,1), & \end{array}\right. \tag{4.1}$$

satisfies

$$u(T,x) = 0, \quad x \in (0,1). \tag{4.2}$$

REMARK 4.1. *Equation* (4.1) *has to be understood in a weak sense. For instance, the solution of* (4.1) *can be defined by transpositions (see* [16], [17]*). Let us briefly recall how can this be done.*

Consider $g \in L^1(0,T,L^2(0,1))$ and v the solution of the adjoint equation

$$\left\{\begin{array}{lll} v_t - v_{xxt} + v_x = g, & x \in (0,1), & t > 0, \\ v(t,0) = v(t,1) = 0, & t > 0, & \\ v(T,x) = 0, & x \in (0,1). & \end{array}\right. \tag{4.3}$$

By multiplying (formally) (4.1) *by v and integrating by parts we obtain*

$$0 = \int_0^T \int_0^1 (u_t - u_{xxt} + u_x)v = \int_0^1 \left(uv|_0^T - u_{xx}v|_0^T\right) + \int_0^T (u_x - uv_x + uv)\,|_0^1$$
$$- \int_0^T \int_0^1 u(v_t - v_{txx} + v_x) = \int_0^1 [-u_0 v(0) + (u_0)_{xx} v(0)] - \int_0^T f v_{tx} - \int_0^T \int_0^1 ug.$$

Therefore we can say that u is the solution of (4.1) *if and only if*

$$\int_0^T \int_0^1 ug + \langle u_0, v(0)\rangle_{H^{-1},H_0^1} = -\int_0^T f(t) v_{tx}(t,1)dt \tag{4.4}$$

$\forall g \in L^1(0,T;L^2(0,1))$ and v the solution of (4.3). *$\langle \cdot, \cdot \rangle$ represents the duality product between H_0^1 and H^{-1}. As in* [16], [17] *it can be proved that* (4.4) *has a unique solution $u \in C([0,T];L^2(0,1))$. On the other hand we have just seen that a classical solution of* (4.1) *is the solution of* (4.4).

Concerning the controllability of (4.1) let us begin with the following result which transforms the control problem into a moments problem.

LEMMA 4.1.

(i) *The initial data $u_0 \in H^{-1}(0,1)$ is controllable to zero in time $T > 0$ with a control $f \in L^2(0,T)$ if and only if*

$$\langle u_0, v(0)\rangle_{H^{-1},H_0^1} = -\int_0^T f(t) v_{tx}(t,1)dt \tag{4.5}$$

for any solution v of the equation

$$\left\{\begin{array}{l} v_t - v_{txx} + v_x = 0, \\ v(t,0) = v(t,1) = 0, \\ v(T,x) = v^T(x) \in H_0^1. \end{array}\right. \tag{4.6}$$

(ii) *The initial data $u_0 \in H^{-1}(0,1)$, $u_0(x) = \sum_{n\in\mathbb{Z}^*} a_n U_n(x)$, is controllable to zero in time $T > 0$ if and only if there exists $f \in L^2(0,T)$ such that*

$$\int_0^T f(t)e^{-i\lambda_n t}dt = \frac{i}{\lambda_n^2 (U_n)_x(1)} a_n \quad \forall n \in \mathbb{Z}^*. \tag{4.7}$$

Proof. (i) Let u be the solution of (4.1) and v the solution of (4.6). It follows that

$$0 = \int_0^T \int_0^1 (u_t - u_{xxt} + u_x)v = -\int_0^T \int_0^1 u(v_t - v_{txx} + v_x)$$

$$+ \int_0^1 (uv - u_{xx}v)\Big|_0^T + \int_0^T (u_x v_t - uv_{xt} + uv)\Big|_0^1 = -\int_0^1 (u_0 v(0) + (u_0)_x v_x)$$

$$+ \int_0^1 (u(T)v(T) + u_x(T)v_x(T)) - \int_0^T f(t)v_{xt}(t,1)dt.$$

We obtain that

$$\int_0^T f(t)v_{xt}(t,1)dt + \langle u_0, v(0)\rangle_{H^{-1},H_0^1} = \langle u(T), v^T\rangle_{H^{-1},H_0^1}$$

$\forall v^T \in H_0^1$.

Hence, u_0 is controllable to zero in time $T > 0$ if and only if (4.5) is satisfied.

(ii) For the second part let us put $v^T = \sum_{n\neq 0} b_n U_n$ and use (4.5). It follows that

$$\sum_{n\neq 0} \frac{1}{\lambda_n} a_n b_n e^{i\lambda_n T} = -\int_0^T f(t) \sum_{n\neq 0} i\lambda_n e^{i\lambda_n(T-t)} b_n (U_n)_x(1)dt$$

which is equivalent to

$$\sum_{n\neq 0} b_n e^{i\lambda_n T}\left[\int_0^T f(t) i\lambda_n e^{-i\lambda_n t}(U_n)_x(1)dt + \frac{1}{\lambda_n}a_n\right] = 0$$

for any $(b_n)_{n\neq 0} \in \ell^2$.

It follows that the control problem is equivalent to finding $f \in L^2(0,T)$ such that

$$\int_0^T f(t)e^{-i\lambda_n t}dt = \frac{i}{(\lambda_n)^2 (U_n)_x(1)} a_n \quad \forall n \in \mathbb{Z}^*. \qquad \square$$

By using Lemma 4.1 and Theorem 3.3 from section 3 the following negative result can be easily proved.

THEOREM 4.2. *No eigenfunction of the operator A can be driven to zero in finite time.*

Proof. The controllability of an eigenfunction U_m is equivalent, by Lemma 4.1, to finding $f \in L^2(0,T)$ such that

$$\int_0^T f(t)e^{-i\lambda_n t} = \begin{cases} 0 & \forall n \in \mathbb{Z}^*, \quad n \neq m, \\ \frac{i}{(\lambda_m)^2 (U_m)_x(1)}, & n = m. \end{cases}$$

Let us suppose that there exists $f \in L^2(0,T)$ with these properties. We define $g \in L^2(-\frac{T}{2}, \frac{T}{2})$ such that $g(t) = f(\frac{T}{2} - t)e^{-\frac{i\lambda_m T}{2}}$ almost everywhere in $(-\frac{T}{2}, \frac{T}{2})$. Then

$$\int_{\frac{T}{2}}^{\frac{T}{2}} g(t)e^{i\lambda_n t}dt = e^{\frac{iT}{2}(\lambda_n - \lambda_m)} \int_0^T f(t)e^{-i\lambda_n t}dt = \begin{cases} 0 & \forall n \in \mathbb{Z}^*, \quad n \neq m, \\ \frac{i}{(\lambda_m)^2 (U_m)_x(1)}, & n = m. \end{cases}$$

However, in Theorem 3.2, we have proved that this is not possible and the proof finishes. □

REMARK 4.2. *From Theorem* 4.2 *it follows that* (4.1) *is not spectrally controllable. This means that no finite linear nontrivial combination of eigenvectors can be driven to zero in finite time by using a control $f \in L^2(0,T)$.*

Let us now study the approximate controllability of (4.1). We recall that (4.1) is approximate controllable in time $T > 0$ if the set of reachable states

$$R(u_0, T) = \{u(T,x) \mid f \in L^2(0,T)\} \tag{4.8}$$

is dense in $L^2(0,1)$ for any $u_0 \in H^{-1}(0,1)$.

In other words, given $T > 0$, $u_0 \in H^{-1}(0,1)$, $v_0 \in L^2(0,1)$, and $\varepsilon > 0$ there exists a control function $f \in L^2(0,T)$ such that the solution u of (4.1) satisfies $||u(T) - v_0||_{L^2(0,1)} < \varepsilon$.

THEOREM 4.3. *Equation* (4.1) *is approximate controllable in any time $T > 0$ with controls in $L^2(0,T)$.*

Proof. From the linearity of (4.1) it follows that it is sufficient to prove that the set $R(0,T)$ is dense in $H_0^1(0,1)$ for any $T > 0$. Therefore we shall consider only the case $u_0 = 0$. Let $u \in C([0,T], H_0^1(0,1))$ be the corresponding solution of (4.1).

Let also v be the solution of the adjoint equation

$$\begin{cases} v_t - v_{txx} + v_x = 0, & x \in (0,1), \quad t < T, \\ v(t,0) = v(t,1) = 0, & t < T, \\ v(T,x) = v^T(x) \in H_0^1(\Omega). \end{cases} \tag{4.9}$$

It follows that

$$\int_0^T f(t) v_{xt}(t,1) dt = (u(T,x), v^T(x))_{H_0^1}. \tag{4.10}$$

Suppose that $R(0,T)$ is not dense in $H_0^1(0,1)$. Hence, there exists $v^T \in H_0^1(0,1)$, $v^T \neq 0$, such that

$$(u(T,x), v^T(x)) = 0 \quad \forall f \in L^2(0,T).$$

From (4.10) it follows that

$$\int_0^T f(t) v_{xt}(t,1) = 0 \quad \forall f \in L^2(0,T).$$

Therefore $v_{xt}(t,1) = 0 \quad \forall t \in (0,T)$. We show now that this contradicts the fact that $v^T \neq 0$. Hence, the problem is reduced to a unique continuation property.

Let us consider the Fourier decomposition of v^T:

$$v^T = \sum_{n \in \mathbb{Z}^*} a_n U_n,$$

where $(a_n)_{n \in \mathbb{Z}^*} \in \ell^2$ and the series converges in $H_0^1(0,1)$.

It follows that the corresponding solution of (4.9) is

$$v(t,x) = \sum_{n \in \mathbb{Z}^*} a_n e^{i\lambda_n (T-t)} U_n(x), \quad t \in (0,T).$$

From the equation v verifies it follows that $v \in C^{\omega}\left([0,\infty);\, H_0^1(0,1)\right)$ (see Remark 2.1).

Hence, from the fact that $v_{xt}(t,1) = 0 \quad \forall t \in (0,T)$, we obtain that $v_{xt}(t,1) = 0 \quad \forall t \in \mathbb{R}$, i.e.,

$$\sum_{n\in\mathbb{Z}^*} a_n e^{i\lambda_n(T-t)}(U_n)_x(1)(-i\lambda_n) = 0 \quad \forall t \in \mathbb{R}.$$

For each $m \in \mathbb{Z}^*$,

$$\begin{aligned} 0 &= \lim_{S\to\infty} \frac{1}{2S}\int_{-S}^{S}\left[\sum_{n\in\mathbb{Z}^*} a_n e^{i\lambda_n(T-t)}(U_n)_x(1)(-i\lambda_n)\right] e^{i\lambda_m t}dt \\ &= a_m(U_m)_x(1)(-i\lambda_m)e^{i\lambda_m T}. \end{aligned}$$

From Remark 2.2 $(U_m)_x(1) \neq 0$. This implies that $a_m = 0 \quad \forall m \in \mathbb{Z}^*$ and therefore $v^T = 0$, which represents a contradiction. Hence, $R(0,T)$ is dense in $H_0^1(0,1)$ and the proof finishes. □

As we have seen in Theorem 4.2 no finite linear combination of eigenfunctions can be driven to zero. In this case the following question arises naturally: can we control to zero at least a part of the solution u of (4.1)? And if we can do this, what is the cost we have to pay?

Therefore we shall now investigate the following special type of controllability.

DEFINITION 4.4. *Equation* (4.1) *is N-partially controllable to zero in time $T > 0$ if, for any $u_0 \in H^{-1}(0,1)$, there exists a control $f \in L^2(0,T)$ such that the projection of the corresponding solution u of* (4.1) *over the space generated by the eigenvectors $(U_n)_{\substack{|n|\leq N\\ n\neq 0}}$ is zero at time $t = T$.*

Let $X_N = \operatorname{Span}\{U_n :| n |\leq N,\, n \neq 0\}$ and let

$$\Pi_N : H^{-1}(0,1) \to X_N, \quad \Pi_N\left(\sum_{n\neq 0} a_n U_n\right) = \sum_{\substack{|n|\leq N\\ n\neq 0}} a_n U_n,$$

be the projection operator.

Evidently, u is N-partially controllable to zero if and only if

$$\Pi_N(u(T)) = 0. \tag{4.11}$$

By using the same argument as in Lemma 4.1, the following result can be obtained immediately.

LEMMA 4.5. *The initial data $u_0 = \sum_{n\neq 0} a_n U_n$ is N-partially controlled to zero in time $T > 0$ if and only if there exists $f \in L^2(0,T)$ such that*

$$\int_0^T f(t)e^{i\lambda_n t}dt = \frac{i}{\lambda_n^2 (U_n)_x(1)} a_n \quad \forall \mid n \mid\leq N, n \neq 0. \tag{4.12}$$

Now, the following theorem can be proved.

THEOREM 4.6. *Any initial data $u_0 \in H^{-1}(0,1)$ can be N-partially controlled to zero in time $T > 0$ by using a control $f_N \in L^2(0,T)$ such that*

$$\| f_N \|_{L^2(0,T)}^2 \leq c_1 \| u_0 \|_{H^{-1}}^2 e^{\alpha_1 N \ln(N)}, \tag{4.13}$$

where c_1 and α_1 are two constants which do not depend on N.

Moreover, there exists initial data $u_0 \in H_0^1(0,1)$ such that any control f_N satisfies

$$\| f_N \|^2_{L^2(0,T)} \geq c_2 \| u_0 \|^2_{H_0^1} e^{\omega_1 N \ln(N)}, \tag{4.14}$$

where c_2 and ω_1 are two constants which do not depend on N.

Proof. Let us consider the initial data $u_0 = \sum_{n\neq 0} a_n U_n$ from $H^{-1}(0,1)$. We prove that there exists a function $f_N \in L^2(0,T)$ such that (4.12) is satisfied. This will be the control we are looking for.

Let $(\Theta_n)_{\substack{n\leq N\\ n\neq 0}}$ be the biorthogonal sequence to $(e^{i\lambda_n t})_{\substack{n\leq N\\ n\neq 0}}$ in $L^2(-\frac{T}{2},\frac{T}{2})$ constructed in Theorem 3.4.

Then we can define

$$f_N(t) = \sum_{\substack{n\neq 0\\ |n|\leq N}} \frac{ia_n}{\lambda_n^2 (U_n)_x(1)} \Theta_n\left(\frac{T}{2} - t\right) e^{\frac{i\lambda_n T}{2}}.$$

Evidently, $f_N \in L^2(0,T)$ and $\int_0^T f_N(t) e^{-i\lambda_n t} dt = \frac{ia_n}{\lambda_n^2 (U_n)_x(1)}$ $\forall \mid n \mid\leq N, n \neq 0$.

From Lemma 4.5 it follows that f_N is the control we are looking for.

By using inequality (3.8) from Theorem 3.4 it follows that

$$\| f_N \|^2_{L^2(0,T)} \leq \sum_{\substack{|n|\leq N\\ n\neq 0}} \frac{\mid a_n \mid^2}{\mid \lambda_n \mid^2 \mid (U_n)_x(1) \mid^2} \sum_{\substack{|n|\leq N\\ n\neq 0}} \| \Theta_n \|^2$$

$$\leq \sum_{\substack{|n|\leq N\\ n\neq 0}} \frac{\mid a_n \mid^2}{\mid \lambda_n \mid^2 \mid (U_n)_x(1) \mid^2} \sum_{\substack{|n|\leq N\\ n\neq 0}} c_1 e^{\alpha N \ln(N)} \leq c_1 \| u_0 \|^2_{H^{-1}} e^{\alpha_1 N \ln(N)}$$

for any $\alpha_1 > \alpha$.

On the other hand let us consider $u_0 = U_m$, $\mid m \mid\leq N$, $m \neq 0$. From Lemma 4.5 u_0 is N-partially controllable to zero in time $T > 0$ if and only if there exists a control $f_N^m \in L^2(0,T)$ such that

$$\int_0^T f_N^m(t) e^{-i\lambda_n t} dt = \begin{cases} 0, & n \neq m, \\ \frac{i}{(\lambda_m)^2 (U_m)_x(1)}, & n = m. \end{cases}$$

We define $g_N^m \in L^2(-\frac{T}{2},\frac{T}{2})$ such that $g_N^m(t) = f_N^m(\frac{T}{2} - t) e^{-\frac{i\lambda_m T}{2}}$ almost everywhere in $(-\frac{T}{2},\frac{T}{2})$. Then

$$\int_{\frac{T}{2}}^{\frac{T}{2}} g_N^m(t) e^{i\lambda_n t} dt = e^{\frac{iT}{2}(\lambda_n - \lambda_m)} \int_0^T f_N^m(t) e^{-i\lambda_n t} dt = \begin{cases} 0 & \forall n \in \mathbb{Z}^*, \quad n \neq m, \\ \frac{i}{(\lambda_m)^2 (U_m)_x(1)}, & n = m. \end{cases}$$

Now, by using Theorem 3.5, it follows that

$$\| f_N^m \|^2_{L^2(0,T)} = \| g_N^m \|^2_{L^2(-\frac{T}{2},\frac{T}{2})} \geq C_2^2 \mid \lambda_m \mid^4 \mid (U_m)_x(1) \mid^2 e^{2\omega N \ln(N)}.$$

It follows that (4.14) is true for any $\omega_1 < 2\omega$ and $c_2 = C_2^2$ and the proof finishes. □

Remark 4.3. *Theorem 4.6 proves that the cost (the norm of the control functions) needed to drive to zero the projection of the solutions of (4.1) over the space generated by the first $2N$ eigenfunctions may increase very rapidly when N goes to infinity. Theorem 4.6 gives an upper bound for these norms (essentially, $e^{\alpha N \ln(N)}$) and shows that there exists a lower bound of the same order.*

5. Comments. As we have mentioned in the introduction, based on the linear case, local or global controllability results (depending on the number of controls) have been obtained for the nonlinear KdV equation in [20], [21], and [24].

The same cannot be said for the nonlinear equation (1.1). In fact, to our knowledge, no result for the controllability of the BBM equation is available. The controllability properties of the nonlinear systems are usually studied by linearizing the problem at an equilibrium state, by proving exact controllability results for this linear problem and by applying next the implicit function theorem. This method was first used in [13] for the ordinary differential equations and next generalized for the nonlinear wave equation (see, for instance, [12]). In [10] and [25] exact and local controllability results were given by using Schauder's fixed point theorem instead of the implicit function theorem. All approaches use the exact controllability result for the linearized equation. Taking into account the negative results (like nonspectral controllability) obtained in this paper for the linearized BBM equation it is not possible to study the controllability properties of (1.1) by using one of the classical techniques mentioned above. Probably, the controllability results for (1.1) are not better than the ones for the corresponding linear case but this is still to be proved.

REFERENCES

[1] J. Albert, *On the decay of solutions of the generalized Benjamin-Bona-Mahony equation*, J. Math. Anal. Appl., 141 (1989), pp. 527–537.

[2] J. Albert, *Dispersion of low-energy waves for the generalized Benjamin-Bona.Mahony equation*, J. Differential Equations, 63 (1986), pp. 117–134.

[3] C. J. Amick, J. L. Bona, and M. E. Schonbeks, *Decay of solutions of some nonlinear wave equations*, J. Differential Equations, 81 (1989), pp. 1–49.

[4] T. B. Benjamin, *Lectures on Nonlinear Wave Motion*, Lectures in Appl. Math. 15, AMS, Providence, RI, 1974.

[5] T. B. Benjamin, J. L. Bona, and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Philos. Trans. Roy. Soc. London Ser. A, 272 (1972), p. 47.

[6] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[7] J. L. Bona and V. A. Dougalis, *An initial and boundary value problem for a model equation propagation of long waves*, J. Math. Anal. Appl., 75 (1980), pp. 503–522.

[8] H. Brezis, *Analyse Fonctionnelles: théorie et applications*, Masson, Paris, 1987.

[9] M. Davila and G. Perla Menzala, *Unique continuation for the Benjamin-Bona-Mahony and Boussinesq's equations*, NoDEA Nonlinear Differential Equations Appl., 5 (1998), pp. 367–182.

[10] C. Fabre, J. P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[11] H. O. Fattorini, *Estimates for sequences biorthogonal to certain complex exponentials and boundary control of the wave equation*, Lecture Notes in Control and Inform. Sci., 2 (1979), pp. 111–124.

[12] H. O. Fattorini, *Local controllbility of a nonlinear wave equation*, Math. Systems Theory, 9 (1975), pp. 35–40.

[13] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley & Sons, New York, 1967.

[14] B. J. Levin, *Distribution of Zeros of Entire Functions*, Transl. Math. Monogr. 5, AMS, Providence, RI, 1950.

[15] J.-L. Lions, *Contrôlabilité exacte perturbations et stabilisation de systèmes distribués*, Tome 1, Masson, Paris, 1988.

[16] J.-L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications*, Vol. 1, Dunod, Paris, 1968.

[17] J.-L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications*, Vol. 2, Dunod, Paris, 1968.

[18] L. A. Medeiros and M. Milla Miranda, *Weak solutions for a nonlinear dispersive equation*, J. Math. Anal. Appl., 59 (1977), pp. 792–799.

[19] S. Micu and E. Zuazua, *On the lack of null-controllability of the heat equation on the half-line*,

Trans. Amer. Math. Soc., 353 (2001), pp. 1635–1659.

[20] L. Rosier, *Exact boundary controllability for the Korteweg-de Vries equation on a bounded domain*, ESAIM: Control Optim. Calc. Var., 2 (1997), pp. 33–55.

[21] D. L. Russell and B.-Y. Zhang, *Controllability and stabilizability of the third-order linear dispersion equation on a periodic domain*, SIAM J. Control Optim., 31 (1993), pp. 659–676.

[22] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[23] B.-Y. Zhang, *Some Results for Nonlinear Dispersive Wave Equations with Applications to Control*, Ph.D. Thesis, University of Wisconsin, Madison, WI, 1990.

[24] B.-Y. Zhang, *Exact boundary controllability of the Korteweg-de Vries equation*, SIAM J. Control Optim., 37 (1999), pp. 543–565.

[25] E. Zuazua, *Exact controllability of the semilinear wave equation*, J. Math. Pures Appl., 69 (1990), pp. 1–31.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1697–1710

SINGULAR STOCHASTIC CONTROL, LINEAR DIFFUSIONS, AND OPTIMAL STOPPING: A CLASS OF SOLVABLE PROBLEMS*

LUIS H. R. ALVAREZ†

Abstract. We consider a class of singular stochastic control problems arising frequently in applications of stochastic control. We state a set of conditions under which the optimal policy and its value can be derived in terms of the minimal r-excessive functions of the controlled diffusion, and demonstrate that the optimal policy is of the standard local time type. We then state a set of weak smoothness conditions under which the value function is increasing and concave, and demonstrate that given these conditions increased stochastic fluctuations decrease the value and increase the optimal threshold, thus postponing the exercise of the irreversible policy. In line with previous studies of singular stochastic control, we also establish a connection between singular control and optimal stopping, and show that the marginal value of the singular control problem coincides with the value of the associated stopping problem whenever 0 is not a regular boundary for the controlled diffusion.

Key words. singular stochastic control, linear diffusions, optimal stopping, harvesting, cash flow management

AMS subject classifications. 49L05, 60H30, 93E20, 92D25

PII. S0363012900367825

1. Introduction.

Singular stochastic control problems arise quite naturally in many fields applying stochastic control theory. Good examples of such fields are, for example, rational harvesting planning (cf. [1], [5], [6], [15], [24], [25], and [26]) and optimal cash flow management (cf. [3], [19], and [29]). Somewhat surprisingly, while the mathematical literature on singular stochastic control is very extensive (cf. [4], [7], [8], [9], [10], [17], [20], [21], [22], [23], [30], and [31]), it has not been applied yet to its full extent in these applications even though there are many important unanswered questions left. In order to illustrate this argument more precisely, consider a harvester facing the problem of having to determine the harvesting strategy maximizing the expected cumulative yield from the present up to extinction given the form of the underlying stochastic population dynamics. The recent literature on singular stochastic harvesting policies indicates that the depletion and, therefore, the extinction of a harvested population is seldom an optimal policy for a future-oriented harvester and that in most cases a part of the population should be left unharvested in order to be able to harvest in the future as well. Since harvesting is suboptimal as long as the marginal yield accrued from harvesting an extra individual falls short of its opportunity cost (measuring the marginal yield accrued from preserving an extra individual) and the latter factor usually dominates the former at sufficiently low densities, we find that the instantaneous depletion of a population is not optimal. Unfortunately, even though this conclusion is appealing, it has been rigorously proved only for diffusion models with logistic expected growth rates. Similarly, most studies of cash flow management rely on cash flows with affine growth and diffusion parameters (geomet-

*Received by the editors February 14, 2000; accepted for publication (in revised form) October 4, 2000; published electronically February 28, 2001. This work was supported by the Foundation for the Promotion of the Actuarial Profession (Aktuaaritoiminnan Kehittämissäätiö), the Yrjö Jahnsson Foundation, and Academy of Finland grant 42753.

http://www.siam.org/journals/sicon/39-6/36782.html

†Department of Economics, Economic Mathematics and Statistics, Turku School of Economics and Business Administration, FIN-20500 Turku, Finland and the Institute of Applied Mathematics, University of Turku, FIN-20014 Turku, Finland (luis.alvarez@tukkk.fi).

ric Brownian motion, Ornstein–Uhlenbeck process, and Brownian motion). Although that assumption is advantageous due to its mathematical simplicity, it overlooks a broad class of processes appearing frequently in reality (for example, *mean reverting diffusions*) and, therefore, leaves unanswered important questions on the nature of the optimal solution and its value in the presence of more complex cash flow processes.

In light of these arguments, we plan to consider the singular stochastic control problem of a regular linear diffusion when the cumulative payoff of the decision-maker depends solely on the implemented policy, and the expected percentage growth rate of the controlled diffusion is decreasing. Thus, the considered problem can be interpreted as the determination of the dividend policy maximizing the expected cumulative present value of the dividends from the present up to the random liquidation date or, alternatively, as the determination of the harvesting strategy maximizing the expected cumulative yield from the present up to the extinction date of the harvested population. By relying on a combination of the classical theory of diffusions and the results of the recent study [4], we show that *the results obtained by relying on logistic growth rates are typically qualitatively robust* in the sense that the optimal policy is to reflect the controlled diffusion at a single threshold satisfying an ordinary first order necessary condition for an optimum in most models subject to decreasing percentage growth rates. In contrast to [4], we are also able to state *a set of very weak sufficient conditions under which the value of the singular stochastic control problem is increasing and concave.* To the best knowledge of the author, these conditions are the weakest under which the value of the singular stochastic control problem has been shown to be concave since no Novikov-type condition is required nor does the drift have to be globally concave. Moreover, the result is shown to be valid independently of the boundary behavior of the controlled diffusion at the lower boundary (for a comparison, see [4]). Given this finding, we show that *increased stochastic fluctuations (i.e., volatility) decrease the value of the singular stochastic control problem and increase the threshold at which the diffusion should optimally be reflected.* Put differently, we demonstrate that *the sign of the relationship between stochasticity and the optimal policy is unambiguously negative*, a result which is in accordance with observations in reality. An economically and biologically important consequence of this finding is that our results support the argument that *increased stochastic fluctuations increase the required exercise premium of an irreversible policy.* We also establish a connection between singular stochastic control and optimal stopping by first demonstrating that the marginal value of the control problem dominates the value of an associated optimal stopping problem. By relying on the classical theory of diffusions, we then prove that these quantities coincide whenever the lower boundary of the state-space of the controlled diffusion is not regular (i.e., when it is either natural, entrance, or exit). This result is of interest since it extends previous results obtained under the assumption that the lower boundary is unattainable for the controlled diffusion (cf. [30] and [31]). In line with recent studies considering singular stochastic control and optimal stopping, we then present an alternative interpretation of the marginal value of the associated stopping problem in terms of the valuation of a perpetual American forward contract and demonstrate that determining the optimal singular stochastic control is closely related to the determination of an optimal irreversible exit policy in the presence of uncertainty.

2. The singular stochastic control problem. Consider the process $\{X(t); t \in [0, \tau(0))\}$, where $\tau(0) = \inf\{t \geq 0 : X(t) \leq 0\}$ denotes the possibly infinite exit date from $\mathbb{R}_+$, defined on a complete filtered probability space $(\Omega, P, \{\mathcal{F}_t\}_{t\geq 0}, \mathcal{F})$ satisfying the usual conditions and described on $\mathbb{R}_+$ by the (Itô-) stochastic differential equation

$$dX(t) = \mu(X(t))X(t)dt + \sigma(X(t))dW(t) - dZ(t), \quad X(0) = x, \tag{1}$$

where the mapping $\mu : \mathbb{R}_+ \mapsto \mathbb{R}$ denotes the expected percentage growth rate of X, $Z(t)$ denotes the implemented control, and $\sigma : \mathbb{R}_+ \mapsto \mathbb{R}_+$ denoting the infinitesimal diffusion coefficient of X is a given Lipschitz-continuous mapping on $\mathbb{R}_+$. We assume throughout this study that the expected percentage growth rate $\mu(x)$ is continuous, decreasing, and satisfies the conditions $\lim_{x\downarrow 0} \mu(x) > 0$, $\lim_{x\downarrow 0} x\mu(x) = 0$, and $\lim_{x\to\infty} \mu(x) < 0$. In other words, (1) describes a dynamic system subject to *pure compensation.* These assumptions imply that the equation $\mu(x) = 0$ has a unique root $K = \mu^{-1}(0)$ and that $\mu(x)x < 0$ for $x > K$ (in mathematical biology, K is known as the *carrying capacity of the environment*; cf. [12], [13], and [14]). For simplicity, we also assume that $\sigma(x) > 0$ on $(0, \infty)$ (this assumption can be relaxed as long as the boundary behavior of the diffusion is specified at the singularities $\{\sigma^{-1}(0)\}$; cf. [4] and [5]). Moreover, in accordance with reality (cf. [1], [5], [6], [24], [25], and [26]), we assume that the upper boundary ∞ of the state-space of X is natural. Thus, even while X may be expected to increase, it is never expected to become infinitely high in finite time. Finally, we call a control Z *admissible* if it is nonnegative, nondecreasing, right-continuous, and $\{\mathcal{F}_t\}$-adapted, and denote the set of admissible controls as Λ.

We observe that when $Z(t) \equiv 0$, X evolves according to a regular linear time homogeneous diffusion with basic characteristics

$$S'(x) = \exp\left(-\int^x \frac{2\mu(y)y}{\sigma^2(y)}dy\right)$$

denoting the density of its scale function S and

$$m'(x) = \frac{2}{\sigma^2(x)S'(x)}$$

denoting the density of its speed measure m. It is worth observing that applying Itô's theorem on the mapping $x \mapsto \ln x$ yields that

$$X(t) = x \exp\left(\int_0^t \mu(X(s))ds\right) M(t), \tag{2}$$

where

$$M(t) = \exp\left(\int_0^t \frac{\sigma(X(s))}{X(s)}dW(s) - \int_0^t \frac{1}{2}\frac{\sigma^2(X(s))}{X^2(s)}ds\right).$$

Thus, if $\sigma(x)/x$ is square-integrable, that is, if

$$E_x \int_0^{t\wedge\tau(0)} \frac{\sigma^2(X(s))}{X^2(s)}ds < \infty,$$

then $M(t)$ is a positive local martingale and, therefore, a supermartingale. In that case, we observe that $0 \leq X(t) \leq xe^{\mu(0)t}M(t)$ almost surely and that $0 \leq E[X(t)] \leq xe^{\mu(0)t}$.

We now define the *net convenience yield accrued from retaining a stock x undistributed* as the mapping $\theta : \mathbb{R}_+ \mapsto \mathbb{R}$ defined in

$$\theta(x) = (\mu(x) - r)x. \tag{3}$$

As is clear intuitively, the mapping $\theta(x)$ measures the net income flow accrued from storing a marginal unit of stock x. In order to guarantee the boundedness of the considered functionals, we assume throughout this study that for all $x \in \mathbb{R}_+$ we have

$$E_x \int_0^{\tau(0)} e^{-rs}\theta(X(s))ds < \infty, \tag{4}$$

where the expectation is taken with respect to the law of the uncontrolled diffusion.

Given the assumptions above, we now plan to consider the stochastic control problem

$$V(x) = \sup_{Z \in \Lambda} E_x \int_0^{\tau(0)} e^{-rs} dZ(s), \tag{5}$$

where $r > 0$ denotes the exogeneously determined discount rate. As in [4], we find by applying the generalized Itô's theorem for semimartingales to the identity map $x \mapsto x$ that

$$V(x) \leq x + \sup_{Z \in \Lambda} E_x \int_0^{\tau(0)} e^{-rs}\theta(X(s))ds \tag{6}$$

for all $x \in \mathbb{R}_+$. Thus, we immediately find the following.

Lemma 1. *If $\mu(0) \leq r$, then the optimal policy is $Z(0) = x$. Under the optimal policy, we have $\tau(0) = 0$ and $V(x) = x$.*

Proof. The assumption $\mu(0) \leq r$ implies that $\theta(x) \leq 0$ for all $x \in \mathbb{R}_+$. The required result is then a straightforward consequence of the inequality (6). □

Lemma 1 states that if the percentage growth rate $\mu(x)$ is smaller than the discount rate r for all x, then waiting is never optimal and the current stock x should be instantaneously depleted since no intertemporal gains can be accrued by postponing the decision into the future. Before proceeding in our analysis, we state the following auxiliary definition.

Definition 2 (see [11, chapter II], [18, section 4.6], and [27, section II.3]). *The Green-kernel $G_r : \mathcal{I}^2 \mapsto \mathbb{R}_+$ of the diffusion X with state-space $\mathcal{I} \subseteq \mathbb{R}$ is defined as*

$$G_r(x,y) = \int_0^\infty e^{-rt} p(t;x,y)dt,$$

where $p(t;x,y)$ is the transition density of X defined with respect to its speed measure m. There are two linearly independent fundamental solutions, $\hat{\psi}(x)$ and $\hat{\varphi}(x)$, with $\hat{\psi}(x)$ increasing and $\hat{\varphi}(x)$ decreasing, spanning the set of solutions of the ordinary differential equation $((\mathcal{A} - r)u)(x) = 0$, where

$$\mathcal{A} = \frac{1}{2}\sigma^2(x)\frac{d^2}{dx^2} + \mu(x)x\frac{d}{dx} \tag{7}$$

is the second order differential operator representing the infinitesimal generator of the underlying diffusion X. The Green-kernel $G_r(x,y)$ can be rewritten in terms of these solutions in the alternative form

$$G_r(x,y) = \begin{cases} B^{-1}\hat{\psi}(x)\hat{\varphi}(y), & x < y, \\ B^{-1}\hat{\psi}(y)\hat{\varphi}(x), & x \geq y, \end{cases}$$

where

$$B = \frac{\hat\psi'(x)}{S'(x)}\hat\varphi(x) - \frac{\hat\varphi'(x)}{S'(x)}\hat\psi(x) > 0$$

is the constant Wronskian determinant of the fundamental solutions.

In order to solve the singular stochastic control problem (5), we now define the auxiliary Markovian functional $F : \mathbb{R}^2_+ \mapsto \mathbb{R}$ as (cf. [1], [4], and [5])

$$F(x,b) = \begin{cases} R(b) - \frac{R'(b)}{\psi'(b)}\psi(b), & x \geq b, \\ R(x) - \frac{R'(b)}{\psi'(b)}\psi(x), & x < b, \end{cases} \tag{8}$$

where

$$\psi(x) = \begin{cases} \hat\psi(x) & \text{if 0 is either natural, exit, or entrance for } X, \\ \hat\psi(x) - \frac{\hat\psi(0)}{\hat\varphi(0)}\hat\varphi(x) & \text{if 0 is regular for } X \end{cases}$$

denotes the increasing fundamental solution of the ordinary differential equation $((\mathcal{A}-r)u)(x) = 0$ defined in the domain of the generator of the diffusion $\{X(t); t \in [0, \tau(0))\}$ and

$$R(x) = B^{-1}\hat\varphi(x)\int_0^x \psi(y)\theta(y)m'(y)dy + B^{-1}\psi(x)\int_x^\infty \hat\varphi(y)\theta(y)m'(y)dy \tag{9}$$

denotes the expected cumulative present value of the future net convenience yields $\theta(x)$ (condition (4) guarantees the convergence of the integrals in (9)). Our first results characterizing the value and optimal policy are now presented in the following (generalizing the results of Theorem 4 in [4]).

Theorem 3. *Assume that $\mu(0) > r$ and that the convenience yield $\theta(x)$ is increasing for $x < x^*$ and decreasing for $x > x^*$, where $x^* = \operatorname{argmax}_{x\in\mathbb{R}_+}\{\theta(x)\} \in (0,\infty)$. Then the optimal policy is*

$$Z(t) = \begin{cases} \mathcal{L}(t, b^*), & t > 0, \\ (x - b^*)^+, & t = 0, \end{cases} \tag{10}$$

where $\mathcal{L}(t,b^*)$ denotes the local time of the process $X(t)$ at the state b^*, and

$$b^* = \operatorname*{argmin}_{x\in\mathbb{R}_+}\left\{\frac{R'(x)}{\psi'(x)}\right\} \in (x^*, \mu^{-1}(r))$$

is the unique interior root of the ordinary first order necessary condition

$$r\int_0^{b^*} \psi(y)\theta(y)m'(y)dy = \theta(b^*)\frac{\psi'(b^*)}{S'(b^*)}. \tag{11}$$

Moreover, the value of the optimal policy is twice continuously differentiable on $\mathbb{R}_+$ and it reads as

$$V(x) = \begin{cases} x + \frac{\theta(b^*)}{r}, & x \geq b^*, \\ x + F(x,b^*), & x < b^*, \end{cases} \tag{12}$$

which can be rewritten alternatively as

$$V(x)=\begin{cases} x+\frac{\theta(b^*)}{r}, & x\geq b^*,\\ \frac{\psi(x)}{\psi'(b^*)}, & x<b^*.\end{cases} \tag{13}$$

Proof. Consider first the Markovian functional

$$\frac{R'(x)}{\psi'(x)}=B^{-1}\frac{\hat{\varphi}'(x)}{\psi'(x)}\int_0^x\psi(y)\theta(y)m'(y)dy+B^{-1}\int_x^\infty\hat{\varphi}(y)\theta(y)m'(y)dy.$$

Standard differentiation yields

$$\frac{d}{dx}\left[\frac{R'(x)}{\psi'(x)}\right]=\frac{2S'(x)}{\sigma^2(x)\psi'^2(x)}\left[r\int_0^x\psi(y)\theta(y)m'(y)dy-\theta(x)\frac{\psi'(x)}{S'(x)}\right].$$

Define now the functional $J:\mathbb{R}_+\mapsto\mathbb{R}$ as

$$J(x)=r\int_0^x\psi(y)\theta(y)m'(y)dy-\theta(x)\frac{\psi'(x)}{S'(x)}. \tag{14}$$

If $z>x>x^*$, the monotonicity of $\theta(x)$ implies that

$$\begin{aligned}\frac{1}{r}\left[J(z)-J(x)\right]&=\int_x^z\psi(y)\theta(y)m'(y)dy-\frac{\theta(z)}{r}\frac{\psi'(z)}{S'(z)}+\frac{\theta(x)}{r}\frac{\psi'(x)}{S'(x)}\\ &>\frac{\theta(z)}{r}\left[\frac{\psi'(z)}{S'(z)}-\frac{\psi'(x)}{S'(x)}\right]-\frac{\theta(z)}{r}\frac{\psi'(z)}{S'(z)}+\frac{\theta(x)}{r}\frac{\psi'(x)}{S'(x)}\\ &=\frac{[\theta(x)-\theta(z)]}{r}\frac{\psi'(x)}{S'(x)}>0,\end{aligned}$$

proving that $I(x)$ is monotonically increasing on (x^*,∞). Similarly, if $z<x<x^*$, the monotonicity of $\theta(x)$ implies that

$$\begin{aligned}\frac{1}{r}\left[J(x)-J(z)\right]&=\int_z^x\psi(y)\theta(y)m'(y)dy-\frac{\theta(x)}{r}\frac{\psi'(x)}{S'(x)}+\frac{\theta(z)}{r}\frac{\psi'(z)}{S'(z)}\\ &<\frac{[\theta(z)-\theta(x)]}{r}\frac{\psi'(z)}{S'(z)}<0,\end{aligned}$$

proving that $I(x)$ is monotonically decreasing on $(0,x^*)$. Moreover, it is clear that the assumption $\mu(0)>r$ and the monotonicity of $\mu(x)$ imply that $\theta(x)>0$ on $(0,\mu^{-1}(r))$. Thus, we find that $J(0)=0$,

$$J(\mu^{-1}(r))=r\int_0^{\mu^{-1}(r)}\psi(y)\theta(y)m'(y)dy,$$

and

$$J(x^*)=\int_0^{x^*}\psi(y)(\theta(y)-\theta(x^*))m'(y)dy-\frac{\theta(x^*)}{r}\frac{\psi'(0)}{S'(0)}<0.$$

Combining this finding with the proven monotonicity and continuity of the functional $J(x)$ implies that there is a unique threshold $b^*\in(x^*,\mu^{-1}(r))$ at which the functional $R'(x)/\psi'(x)$ is minimized and that $R'(x)/\psi'(x)$ is monotonically decreasing for $x<b^*$

and monotonically increasing for $x > b^*$. Given these results, we observe that the proposed value function is twice continuously differentiable on $\mathbb{R}_+$. Moreover, we observe that on $(0, b^*)$ the proposed value function satisfies the inequality

$$V'(x) = 1 + R'(x) - \frac{R'(b^*)}{\psi'(b^*)}\psi'(x) \geq 1,$$

since $\psi'(x) > 0$ and $R'(x)/\psi'(x) \geq R'(b^*)/\psi'(b^*)$ for all $x \in \mathbb{R}_+$. Moreover, we also find that

$$((\mathcal{A} - r)V)(x) = \begin{cases} \theta(x) - \theta(b^*), & x \geq b^* \\ 0, & x < b^* \end{cases} \leq 0,$$

since $((\mathcal{A} - r)R)(x) + \theta(x) = 0$ and b^* is attained on the set where $\theta(x)$ is decreasing. Thus, the proposed value function satisfies the conditions of Lemma 1 in [4] and, therefore, dominates the value of the singular stochastic control problem (5). The rest of the proof is then completed as in the proof of Theorem 4 in [4]. □

Theorem 3 states a set of considerably weak sufficient conditions under which the optimal policy is of the threshold type in the sense that there is a unique optimal threshold at which the control policy should be instantaneously applied at a maximal rate in order to maintain the process below the critical threshold b^*. It is clear that if $\mu(x)$ is continuously differentiable on $\mathbb{R}_+$, then partial integration of (11) implies that the optimal reflection threshold has to satisfy the equivalent condition

$$\int_0^{b^*} \frac{\psi'(y)}{S'(y)}[\mu(y) - r + \mu'(y)y]dy = 0. \tag{15}$$

It is also worth noticing that the optimal threshold b^* is below the equilibrium density $\mu^{-1}(r)$ at which the percentage growth rate is equal to the discount rate, but above the density x^* at which the net convenience yield $\theta(x)$ is maximized. This result is of interest since the difference $b^* - x^*$ measures the *required exercise premium from exercising the irreversible policy*, since x^* is the threshold at which the policy is irreversibly exercised in the absence of stochastic fluctuations (i.e., in the deterministic case).

The results of Theorem 3 are interesting from the point of view of studies considering *rational harvesting planning in the presence of extinction risk*, since it demonstrates that the basic conclusions of models relying on logistic growth remain valid even in models subject to more complex and general per capita growth rates. More specifically, Theorem 3 shows that there is a unique optimal threshold density at which harvesting should be initiated in most models of pure compensation (cf. [5], [6], [24], [25], and [26]). Thus, we find that the *results obtained by relying on logistic models are qualitatively robust*. Similarly, Theorem 3 extends the older results obtained in studies considering the determination of the *rational dividend policy of a firm facing the risk of liquidation* by stating a simple monotonicity condition in terms of the discount rate and the expected growth rate of the cash flow of the firm (cf. [3], [19], and [29]). Moreover, if $\mu(x)$ is assumed to be differentiable, then the result of Theorem 3 is of interest from a *capital theoretic* point of view as well, since it demonstrates that in the presence of stochastic fluctuations we have that $\mu(b^*) + \mu'(b^*)b^* < r$, thus violating the standard *deterministic golden rule of capital accumulation* stating that the marginal yield $\mu(b^*) + \mu'(b^*)b^*$ accrued from retaining yet another marginal unit of stock x should be equal to the interest rate r (cf. [28, pp. 594–595]). In the present case, we find by ordinary differentiation that

$$\mu(b^*) + \mu'(b^*)b^* = r - \frac{1}{2}\sigma^2(b^*)V'''(b^*-).$$

It is also of interest to point out that Theorem 3 demonstrates that on $x < b^*$ the value function satisfies the ordinary differential equation $((\mathcal{A}-r)V)(x) = 0$ subject to the (*Von Neumann-type*) boundary condition $V'(b^*) = 1$ and the variational inequality $V'(x) > 1$. Especially, Theorem 3 shows that the value function $V \in C^2(\mathbb{R}_+)$ is the solution of the quasi-variational inequality

$$\min\{((r-\mathcal{A})V)(x), V'(x) - 1\} = 0.$$

Remark 1. It is worth pointing out that the results of Theorem 3 are also valid in the case where the infinitesimal diffusion coefficient $\sigma(x)$ vanishes at a point $\xi \geq K = \mu^{-1}(0)$ (*independently on the boundary behavior of* X *at* ξ). The reason for this finding is that the analysis in the proof of Theorem 3 is independent of the upper boundary of the state-space of the controlled diffusion as long as this boundary is not below $\mu^{-1}(0)$. This type of diffusion usually appears in studies considering random percentage growth rates, that is, in models of the form (cf. [4], [5], [13], [14], and [26])

$$\frac{dX(t)}{X(t)} = (\alpha dt + \sigma dW(t))\mu(X(t)).$$

A set of sufficient conditions under which the conditions of Theorem 1 are always satisfied are now summarized in the following.

COROLLARY 4. *If $\mu(x)$ is continuously differentiable, $\mu(x)x$ is strictly concave, and $\mu(0) > r$, then the conclusions of Theorem* 3 *are valid.*

Proof. It is sufficient to prove that the net convenience yield satisfies the monotonicity properties of Theorem 3. It is clear that under the assumptions of our corollary, the convenience yield $\theta(x)$ is strictly concave and satisfies the condition $\theta(0) = \theta(\hat{x}) = 0$, where $\hat{x} = \mu^{-1}(r)$. *Rolle's theorem* then implies that there is at least one point $x^* \in (0, \hat{x})$ where the marginal net convenience yield $\theta'(x)$ vanishes. Since $\theta(x)$ is strictly concave, x^* is unique and constitutes a global maximum of $\theta(x)$. Moreover, $\theta'(x) > 0$ on $(0, x^*)$ and $\theta'(x) < 0$ on (x^*, ∞) completing the proof of our corollary. □

Corollary 4 states an usually satisfied concavity condition under which the results of our Theorem 3 are always valid (for example, logistic and Gompertz-type growth). However, it is worth emphasizing that the results of Theorem 3 are more generally valid than the results of Corollary 4, since the concavity of the mapping $\mu(x)x$ is not a necessary condition for the existence of a well-defined global maximum for the net convenience yield $\theta(x)$ (for example, the gamma-response model $\mu(x)x = (x^{-\alpha}e^{-\beta x} - \delta)x$, where $0 < \alpha < 1$, $\beta > 0$, and $\delta > 0$).

In order to describe unambiguously the sign of the relationship between stochastic fluctuations and the optimal policy, we first prove the following important result summarizing the monotonicity and curvature properties of the value.

THEOREM 5. *Assume that $\mu(0) > r$ and that the net convenience yield $\theta(x)$ is increasing for $x < x^*$ and decreasing for $x > x^*$. Then, $V'(x) > 0$ and $V''(x) \leq 0$ for all $x \in \mathbb{R}_+$, and $F(x, b^*)$ is concave on $(0, b^*)$.*

Proof. As was shown in Theorem 3, our assumptions imply that the value is twice continuously differentiable and that it can be written as in (13). Since $\psi(x)$

is increasing, ordinary differentiation of (13) then proves the alleged monotonicity of $V(x)$. Differentiating (13) twice then yields

$$V''(x) = \begin{cases} 0, & x \geq b^*, \\ \frac{\psi''(x)}{\psi'(b^*)}, & x < b^*. \end{cases}$$

Thus, it is sufficient to show that $\psi(x)$ is concave on $(0, b^*)$. To accomplish this task, we first observe that the ordinary differential equation $((\mathcal{A} - r)\psi)(x) = 0$ can be rewritten as

$$\frac{1}{2}\sigma^2(x)\frac{\psi''(x)}{S'(x)} = r\frac{\psi(x)}{S'(x)} - \mu(x)\frac{\psi'(x)}{S'(x)}. \tag{16}$$

Adding and subtracting $rx\psi'(x)/S'(x)$ from (16) then yield that

$$\frac{1}{2}\sigma^2(x)\frac{\psi''(x)}{S'(x)} = r\left[\frac{\psi(x)}{S'(x)} - x\frac{\psi'(x)}{S'(x)}\right] - \theta(x)\frac{\psi'(x)}{S'(x)}.$$

Consider now the mapping

$$g(x) = \frac{\psi(x)}{S'(x)} - x\frac{\psi'(x)}{S'(x)}.$$

Since $0 \leq \lim_{x\downarrow 0} \frac{\psi'(x)}{S'(x)} < \infty$ (cf. [11, p. 19]), we find that the second term on the right-hand side of the equation above vanishes as $x \downarrow 0$. Thus, we observe that

$$\lim_{x\downarrow 0} g(x) = \lim_{x\downarrow 0} \frac{\psi(x)}{S'(x)}.$$

If 0 is either natural, exit, or regular, we find that $\lim_{x\downarrow 0} \frac{\psi(x)}{S'(x)} = 0$ whenever $\lim_{x\downarrow 0} S'(x) > 0$, since $\psi(0) = 0$ (cf. [11, p. 19]). If 0 is entrance, then $\lim_{x\downarrow 0} S'(x) = \infty$ and $\psi(0) \in [0, \infty)$, implying that $\lim_{x\downarrow 0} \frac{\psi(x)}{S'(x)} = 0$ (cf. [11, p. 19]). Thus, it remains to deal with the case when 0 is either natural, exit, or regular and $\lim_{x\downarrow 0} S'(x) = 0$. The latter condition implies that $\frac{\mu(x)x}{\sigma^2(x)}$ has to tend towards infinity as $x \downarrow 0$. L'Hospital's rule then yields that

$$\lim_{x\downarrow 0} g(x) = -\lim_{x\downarrow 0} \frac{\sigma^2(x)}{2\mu(x)x}\frac{\psi'(x)}{S'(x)} = 0,$$

proving that

$$\lim_{x\downarrow 0} g(x) = 0.$$

Ordinary differentiation of $g(x)$ then yields that

$$g'(x) = \theta(x)\psi(x)m'(x),$$

implying that

$$\frac{\psi(x)}{S'(x)} - x\frac{\psi'(x)}{S'(x)} = \int_0^x \theta(y)\psi(y)m'(y)dy.$$

Thus, we have proved that

$$\frac{1}{2}\sigma^2(x)\frac{\psi''(x)}{S'(x)} = r\int_0^x \theta(y)\psi(y)m'(y)dy - \frac{\psi'(x)}{S'(x)}\theta(x). \tag{17}$$

Combining (17) with (11) and the analysis in the proof of Theorem 3 then implies that $\psi''(x) \leq 0$ for all $x \in (0, b^*]$, proving the alleged concavity of both the value $V(x)$ and the cumulative net convenience yields $R(x, b^*)$. □

Theorem 5 states a set of sufficient conditions under which *the marginal value $V'(x)$ is positive but diminishing* as a mapping of the current stock x. It is worth emphasizing that Theorem 5 is valid under a set of very weak conditions, since it does not require a Novikov-type condition, nor does it require concavity or differentiability assumptions on the form of the drift $\mu(x)x$ (for a comparison, see [16]). Thus, Theorem 5 demonstrates that it is only the first order monotonicity properties of the yield $\theta(x)$ which determine both the monotonicity and concavity of the value of the optimal policy. To the best knowledge of the author, the conditions of Theorem 5 are the weakest under which the concavity of the value function has been proven.

Our main results on the effect of increased volatility both on the optimal policy and the value are now summarized in the following.

THEOREM 6. *Assume that the conditions of Theorem* 5 *are met. Then, increased stochastic fluctuations decrease or leave unchanged the value $V(x)$ and increase or leave unchanged the optimal threshold. That is, if $\tilde{\sigma} : \mathbb{R}_+ \mapsto \mathbb{R}$ satisfies the condition $\tilde{\sigma}(x) \geq \sigma(x)$ on $\mathbb{R}_+$, $\tilde{V}(x)$ denotes the value, and $\tilde{b}$ denotes the optimal threshold in the presence of greater stochastic fluctuations, then $\tilde{b} \geq b^*$ and $\tilde{V}(x) \leq V(x)$ on $\mathbb{R}_+$.*

Proof. The alleged result is proved by applying the same technique as in the proof of Theorem 6 in [4]. □

Theorem 6 states a set of usually satisfied conditions under which increased uncertainty has a negative impact on both the value and the optimal policy. As Theorem 6 demonstrates, increased uncertainty increases the optimal threshold and, therefore, postpones the exercise of the singular policy whenever the net convenience yield is increasing below a given point x^* and decreasing above it. It is also worth pointing out that Theorem 6 clearly shows that the value of the associated singular stochastic control problems can be completely ordered in terms of the volatilities of the controlled diffusions whenever the value functions are concave. The result of Theorem 6 is very important from the point of view of applications, since it confirms the intuitively clear argument that *the sign of the relationship between uncertainty and an irreversible policy is unambiguously negative.* Summarizing, we have the following.

COROLLARY 7. *Assume that $\mu(0) > r$ and that the net convenience yield $\theta(x)$ is increasing for $x < x^*$ and decreasing for $x > x^*$. Then, increased stochastic fluctuations increase the required exercise premium $b^* - x^*$.*

Proof. Since x^* is independent of $\sigma(x)$, the conclusion is a straightforward implication of Theorem 6. □

3. Optimal stopping and the marginal value.

It is well known that singular stochastic control is closely related to optimal stopping (cf. [7], [8], [10], [20], [21], [22], [30], and [31]). In order to verify this connection in the present case, we assume throughout this section that both the diffusion coefficient $\sigma(x)$ and the percentage growth rate $\mu(x)$ are continuously differentiable on $\mathbb{R}_+$. Given these extra assumptions, we now plan to consider the optimal stopping problem

$$H(x) = \sup_{\tau < \tilde{\tau}(0)} E_x\left[\exp\left(\int_0^\tau \theta'(\tilde{X}(s))ds\right)\right], \tag{18}$$

where $\tilde{X}(t)$ evolves according to the diffusion described by the stochastic differential equation

$$(19)\quad d\tilde{X}(t) = (\mu(\tilde{X}(t))\tilde{X}(t) + \sigma'(\tilde{X}(t))\sigma(\tilde{X}(t)))dt + \sigma(\tilde{X}(t))dW(t), \quad \tilde{X}(0) = x,$$

and $\tilde{\tau}(0) = \inf\{t \geq 0 : \tilde{X}(t) \leq 0\}$. We can now show the following.

LEMMA 8. *Assume that $\mu(0) > r$ and that the net convenience yield $\theta(x)$ is increasing for $x < x^*$ and decreasing for $x > x^*$. Then, $V'(x) \geq H(x)$ for all $x \in \mathbb{R}_+$.*

Proof. It is clear that $V' \in C^1(\mathbb{R}_+) \cap C^2(\mathbb{R}_+ \backslash \{b^*\})$, that $V'(x) \geq 1$ for all $x \in \mathbb{R}_+$, that $V'''(b^*+) = 0$, and that

$$V'''(b^*-) = -\frac{2\theta'(b^*)}{\sigma^2(b^*)} < \infty.$$

Moreover, since

$$((\mathcal{A} - r)V)(x) = \begin{cases} \theta(x) - \theta(b^*), & x > b^*, \\ 0, & x < b^*, \end{cases}$$

we find that

$$\frac{d}{dx}((\mathcal{A} - r)V)(x) = \frac{1}{2}\sigma^2(x)V'''(x) + (\mu(x)x + \sigma'(x)\sigma(x))V''(x) + \theta'(x)V'(x) \leq 0,$$

since b^* is attained on the set where $\theta(x)$ is decreasing. Thus, the alleged result follows from Theorem 10.4.1 in [32]. □

Lemma 8 demonstrates, by relying on a simple variational argument, that under the assumptions of our paper the marginal value $V'(x)$ of the singular stochastic control problem (5) dominates the value of the associated stopping problem (18). It is, of course, of interest to find out the cases when these two values may coincide in the present case. The main conclusion of this section is now summarized in the following.

THEOREM 9. *Assume that $\mu(0) > r$, that the convenience yield $\theta(x)$ is increasing for $x < x^*$, decreasing for $x > x^*$, and that 0 is either natural, exit, or entrance for the diffusion X. Then, $V'(x) = H(x)$ for all $x \in \mathbb{R}_+$.*

Proof. We know from Lemma 8 that $V'(x) \geq H(x)$ so it is sufficient to show the opposite inequality. Since the stopping time τ in the optimal stopping problem (18) is arbitrary, we observe that

$$H(x) \geq E_x \left[\exp\left(\int_0^{\tau(0,b^*)} \theta'(\tilde{X}(s))ds \right) \right],$$

where $\tau(0, b^*) = \inf\{t \geq 0 : \tilde{X}(t) \notin (0, b^*)\}$. Define now the functional $G(x, a, b^*)$ as

$$G(x, a, b^*) = E_x \left[\exp\left(\int_0^{\tau(a,b^*)} \theta'(\tilde{X}(s))ds \right) \right],$$

where $\tau(a, b^*) = \inf\{t \geq 0 : \tilde{X}(t) \notin (a, b^*)\}$. Since $\hat{\varphi}''(x)\psi'(x) - \psi''(x)\hat{\varphi}'(x) = 2rB\tilde{S}'(x) > 0$, where $\tilde{S}'(x) = S'(x)/\sigma^2(x)$, we observe that the general solution of the ordinary second order differential equation

$$\frac{1}{2}\sigma^2(x)u''(x) + (\mu(x)x + \sigma'(x)\sigma(x))u'(x) + \theta'(x)u(x) = 0$$

is $u(x) = c_1\psi'(x) + c_2\hat{\varphi}'(x)$, where c_1 and c_2 are unknown real constants. It is now an elementary exercise in linear algebra to demonstrate that if $x \in (a, b^*)$, then

$$G(x,a,b^*) = \frac{\hat{\varphi}'(x) - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}\psi'(x)}{\hat{\varphi}'(a) - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}\psi'(a)} + \frac{\psi'(x) - \frac{\psi'(a)}{\hat{\varphi}'(a)}\hat{\varphi}'(x)}{\psi'(b^*) - \frac{\psi'(a)}{\hat{\varphi}'(a)}\hat{\varphi}'(b^*)}. \tag{20}$$

Consider now the second term on the right-hand side of (20). Invoking the alleged boundary behavior of X then implies that (cf. [11, p. 19])

$$\lim_{a\downarrow 0} \frac{\psi'(x) - \frac{\psi'(a)}{\hat{\varphi}'(a)}\hat{\varphi}'(x)}{\psi'(b^*) - \frac{\psi'(a)}{\hat{\varphi}'(a)}\hat{\varphi}'(b^*)} = \lim_{a\downarrow 0} \frac{\psi'(x) - \frac{\psi'(a)/S'(a)}{\hat{\varphi}'(a)/S'(a)}\hat{\varphi}'(x)}{\psi'(b^*) - \frac{\psi'(a)/S'(a)}{\hat{\varphi}'(a)/S'(a)}\hat{\varphi}'(b^*)} = \frac{\psi'(x)}{\psi'(b^*)}.$$

Consider now the first term on the right-hand side of (20). Since $\hat{\varphi}'(x)/\psi'(x)$ is increasing on $\mathbb{R}_+$ and $\psi(x)$ is increasing and concave on $(0, b^*)$, we find that

$$0 \leq \frac{\hat{\varphi}'(x) - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}\psi'(x)}{\hat{\varphi}'(a) - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}\psi'(a)} = \frac{\psi'(x)}{\psi'(a)} \frac{\frac{\hat{\varphi}'(x)}{\psi'(x)} - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}}{\frac{\hat{\varphi}'(a)}{\psi'(a)} - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}} \leq \frac{\frac{\hat{\varphi}'(x)}{\psi'(x)} - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}}{\frac{\hat{\varphi}'(a)}{\psi'(a)} - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}}.$$

Letting $a \downarrow 0$ and invoking again the alleged boundary behavior of X then yields

$$0 \leq \lim_{a\downarrow 0} \frac{\hat{\varphi}'(x) - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}\psi'(x)}{\hat{\varphi}'(a) - \frac{\hat{\varphi}'(b^*)}{\psi'(b^*)}\psi'(a)} \leq 0$$

implying that

$$\lim_{a\downarrow 0} G(x,a,b^*) = \frac{\psi'(x)}{\psi'(b^*)}$$

and, therefore, that

$$H(x) \geq G(x,0,b^*) = V'(x)$$

completing the proof of our theorem. □

Theorem 9 demonstrates, by relying on the classical theory of linear diffusions, that the marginal value $V'(x)$ of the singular stochastic control problem (5) coincides with the value of an associated optimal stopping problem (18) except in the case where the lower boundary is regular. It is worth emphasizing that the result of Theorem 9 is rather general since it is valid also in the case where the lower boundary may be attainable in finite time (exit) for the controlled diffusion X. It is also clear that whenever 0 is regular, we have that $V'(x) \geq H(x)$ and, therefore, that $\{x \in \mathbb{R}_+ : H(x) > 1\} \subseteq \{x \in \mathbb{R}_+ : V'(x) > 1\}$. In other words, the continuation region of the stopping problem (18) is a subset of the do-nothing-region in the associated singular stochastic control problem (5).

Applying now the fundamental theorem of calculus to the mapping $\exp(\int_0^t \theta'(\tilde{X}(s))ds)$ then implies that the value of the optimal stopping problem (18) coincides with the value of a perpetual American-type forward contract (an optimal exit problem; see [2] and references therein, see also [8] for an associated interpretation) of the form

$$H(x) = 1 + \sup_{\tau < \tilde{\tau}(0)} E_x \int_0^\tau \exp\left(\int_0^s \theta'(\tilde{X}(t))dt\right) \theta'(\tilde{X}(s))ds.$$

Thus, we find the following.

COROLLARY 10. *Assume that $\mu(0) > r$, that the convenience yield $\theta(x)$ is increasing for $x < x^*$, decreasing for $x > x^*$, and that 0 is either natural, exit, or entrance for the diffusion X. Then, for all $x \in \mathbb{R}_+$ we have that*

$$V'(x) = 1 + \sup_{\tau<\tilde{\tau}(0)} E_x \int_0^\tau \exp\left(\int_0^s \theta'(\tilde{X}(t))dt\right) \theta'(\tilde{X}(s))ds.$$

Proof. The result is a direct consequence of Theorem 9. □

Corollary 10 presents an interesting result from the point of view of applications. Namely, it states that under the optimal policy (given the conditions of Theorem 9), the marginal expected cumulative present value of the future convenience yields $\theta(x)$ is equal to the expected cumulative present value of the future marginal convenience yields $\theta'(x)$ evaluated from the present up to the Markov time at which the accumulation of the marginal yields is optimally stopped.

REFERENCES

[1] L. H. R., ALVAREZ, *Optimal harvesting under stochastic fluctuations and critical depensation*, Math. Biosci., 152 (1998), pp. 63–85.

[2] L. H. R. ALVAREZ, *Exit strategies and price uncertainty: A Greenian approach*, J. Math. Econom., 29 (1998), pp. 43–56.

[3] L. H. R. ALVAREZ, *Managerial Compensation and Corporate Behavior*, University of Turku, Institute for Applied Mathematics, Research report A25, 1998.

[4] L. H. R. ALVAREZ, *A class of solvable singular stochastic control problems*, Stochastics Stochastics Rep., 67 (1999), pp. 83–122.

[5] L. H. R. ALVAREZ, *On the option interpretation of rational harvesting planning*, J. Math. Biol., 40 (2000), pp. 383–405.

[6] L. H. R. ALVAREZ AND L. A. SHEPP, *Optimal harvesting of stochastically fluctuating populations*, J. Math. Biol., 37 (1998), pp. 155–177.

[7] F. M. BALDURSSON, *Singular stochastic control and optimal stopping*, Stochastics Stochastics Rep., 21 (1987), pp. 1–40.

[8] F. M. BALDURSSON AND I. KARATZAS, *Irreversible investment and industry equilibrium*, Finance and Stochastics, 1 (1997), pp. 69–89.

[9] V. E. BENES, L. A. SHEPP, AND H. S. WITSENHAUSEN, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.

[10] F. BOETIUS AND M. KOHLMANN, *Connections between optimal stopping and singular stochastic control*, Stochastic Processes Appl., 77 (1998), pp. 253–281.

[11] A. BORODIN AND P. SALMINEN, *Handbook on Brownian Motion—Facts and Formulae*, Birkhäuser, Basel, 1996.

[12] C. A. BRAUMANN, *General models of fishing with random growth parameters*, in Mathematics Applied to Biology and Medicine, J. Demongeot and V. Capasso, eds., Wuerz Publishing Ltd, Winnipeg, MB, 1993.

[13] C. A. BRAUMANN, *Variable effort fishing models in random environments*, Math. Biosci., 156 (1999), pp, 1–19.

[14] C. A. BRAUMANN, *Applications of stochastic differential equations to population growth*, in Proceedings of the Ninth International Colloquium on Differential Equations, D. Bainov, ed., VSP BV, Utrecht, The Netherlands, 1999, pp. 47–52.

[15] C. W. CLARK, *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*, Wiley, New York, 1976.

[16] N. EL KAROUI, M. JEANBLANC-PICQUÉ, AND S. E. SHREVE, *Robustness of the Black-Scholes formula*, Math. Finance, 8 (1998), pp. 93–126.

[17] W. H. FLEMING AND H. M.SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer, New York, 1993.

[18] K. ITÔ AND H. P. MCKEAN JR., *Diffusion Processes and Their Sample Paths*, Springer, Berlin, 1965.

[19] M. JEANBLANC-PICQUÉ AND A. N. SHIRYAEV, *Optimization of the flow of dividends*, Russian Math. Surveys, 50 (1995), pp. 257–277.

[20] I. KARATZAS, *A class of singular stochastic control problems*, Adv. Appl. Probab., 15 (1983), pp. 225–254.

[21] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control.* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.

[22] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control.* II. *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.

[23] T. Ø Kobila, *A class of solvable stochastic investment problems involving singular controls*, Stochastics Stochastics Rep., 43 (1993), pp. 29–63.

[24] R. Lande, S. Engen, and B.-E. Sæther, *Optimal harvesting, economic discounting and extinction risk in fluctuating populations*, Nature, (1994), pp. 88–90.

[25] R. Lande, S. Engen, and B.-E. Sæther, *Optimal harvesting of fluctuating populations with a risk of extinction*, The American Naturalist, 145 (1995), pp. 728–745.

[26] E. M. Lungu and B. Øksendal, *Optimal harvesting from a population in a stochastic crowded enviroment*, Math. Biosci., 145 (1996), pp. 47–75.

[27] P. Mandl, *Analytical Treatment of One-Dimensional Markov Processes*, Springer, Prague, 1968.

[28] R. C. Merton, *Continuous-Time Finance*, Basil Blackwell, Oxford, UK, 1990.

[29] A. Milne and D. Robertson, *Firm behaviour under the threat of liquidation*, J. Econom. Dynam. Control, 20 (1996), pp. 1427–1449.

[30] T. Myhre, *A Connection between Singular Stochastic Control and Optimal Stopping*, MSc thesis, Department of Mathematics, University of Oslo, Oslo, Norway, 1997.

[31] T. Myhre, *Connections between Optimal Stopping and Singular Stochastic Control*, Department of Mathematics, University of Oslo, Oslo, Norway, manuscript.

[32] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 5th ed., Springer, Berlin, 1998.

SIAM J. Control Optim.
Vol. 39, No. 6, pp. 1711–1735

AN LMI-BASED ALGORITHM FOR DESIGNING SUBOPTIMAL STATIC $\mathcal{H}_2/\mathcal{H}_\infty$ OUTPUT FEEDBACK CONTROLLERS*

F. LEIBFRITZ[†]

Abstract. We consider the problem of designing a suboptimal $\mathcal{H}_2/\mathcal{H}_\infty$ feedback control law for a linear time-invariant control system when a complete set of state variables is not available. This problem can be necessarily restated as a nonconvex optimization problem with a bilinear, multiobjective functional under suitably chosen linear matrix inequality (LMI) constraints. To solve such a problem, we propose an LMI-based procedure which is a sequential linearization programming approach. The properties and the convergence of the algorithm are discussed in detail. Finally, several numerical examples for static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problems demonstrate the applicability of the considered algorithm and also verify the theoretical results numerically.

Key words. static controllers, output feedback, suboptimal control, robust control, linear systems, linear matrix inequalities, nonconvex programming

AMS subject classifications. 90C22, 90C26, 93A99, 93B36, 93B51, 93B52, 93C05, 93D09, 49N99, 65K05

PII. S0363012999349553

1. Introduction.

The static or reduced fixed order dynamic output feedback control problem that meets desired performance and/or robustness specifications is an active research area of the control community. In this paper we consider the computational design of static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controllers. This is an important example of a nonconvex control problem. It consists of determining a static output feedback gain which achieves a certain nominal (suboptimal) performance measure subject to a robustness constraint. The static output feedback problems are important, since it is not always possible to have full access to the state vector and a controller must be used which is based only on the available observations. Moreover, they are important because other problems are reducible to some variations of the static output feedback problem and relevant when a simple controller must be used due to cost and reliability.

During the past decade, control problems with combined $\mathcal{H}_2$ and $\mathcal{H}_\infty$ design criteria have gained a great deal of attention. Concerning continuous-time systems, [7] provides the solution of standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems in terms of algebraic Riccati equations, where both state feedback and full order compensator-based output feedback are considered. The design of feedback controllers that satisfy both $\mathcal{H}_\infty$ and $\mathcal{H}_2$ specifications is interesting because it offers robust stability and nominal performance. In 1989, Bernstein and Haddad [2] introduced a mixed $\mathcal{H}_2/\mathcal{H}_\infty$ problem. Their approach is to minimize an auxiliary cost subject to an $\mathcal{H}_\infty$ norm constraint, and this cost yields an upper bound on the $\mathcal{H}_2$ norm. The work of [2] is extended in [46] and [6], where another mixed $\mathcal{H}_2/\mathcal{H}_\infty$ problem is addressed. The system considered therein is dual to the Bernstein–Haddad setup (see [45]). Other related works on the design of $\mathcal{H}_2/\mathcal{H}_\infty$ controllers by state or full order output feedback can be found, for example, in [23], [33], [35], [38], and [40]. Only [44] considers a mixed $\mathcal{H}_2/\mathcal{H}_\infty$ problem for the static output feedback case. The solvability conditions and

*Received by the editors January 7, 1999; accepted for publication (in revised form) September 5, 2000; published electronically February 28, 2001.
http://www.siam.org/journals/sicon/39-6/34955.html

[†]FB-IV Mathematics, University of Trier, D-54286 Trier, Germany (leibfr@uni-trier.de).

the algorithms discussed in the above literature are based on coupled Riccati and/or Lyapunov equations. Recently, linear matrix inequalities (LMIs) have attained much attention in control engineering [3], [5], [36], since many control problems can be formulated in terms of LMIs and thus solved via convex programming approaches. For example, this includes $\mathcal{H}_\infty$ [4], [11], [12], [25], $\mathcal{H}_2$ [18], [37], and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ [19], [20], [27], [29], [39]. However, the resulting controllers are state feedback or of order n_x equal to the plant. The difficulties arise if we want to design a static (or reduced fixed order) output feedback controller. Then the problem of determining a static output feedback controller including $\mathcal{H}_2$ and $\mathcal{H}_\infty$ can be restated as a linear algebra problem, which involves two coupled LMIs. In this case, the solution of one should be the inverse of the other. The problem is then no longer convex [12], [25], [31], [30], [42], [43], and finding a solution numerically to these nonconvex problems is a difficult task.

In this paper we will develop an LMI-based computational procedure for solving a mixed $\mathcal{H}_2/\mathcal{H}_\infty$ problem by a static output feedback controller, which is an extension of the algorithm proposed by Leibfritz [31] for the design of stabilizing static $\mathcal{H}_2$ and $\mathcal{H}_\infty$ output feedback gains. The suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem considered in this paper can be necessarily rewritten to a nonconvex, multiobjective programming problem. In particular, this problem consists of minimizing a (nonconvex) functional of the form $\mathcal{J}(P, Q, Y) = \mathrm{Tr}(PQ) + \mathrm{Tr}(Y)$ subject to suitably chosen LMI constraints. Then, using the solution of this problem (if any exists), the existence of a static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain F can be tested by solving a suitably chosen LMI feasibility problem in F. Therefore, the problem of finding a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain reduces to an optimization problem with a nonconvex objective over a convex set and an LMI feasibility problem.

Similar to Leibfritz [31], we will derive the so-called sequential linear programming matrix method (SLPMM) for solving the resulting nonconvex programming problem. This approach is motivated by successive minimization of a linearization of $\mathcal{J}(P, Q, Y)$ subject to LMI constraints as proposed by [1] for general nonconvex bilinear programming problems. Note that the SLPMM algorithm is closely related to the cone complementarity linearization method developed by [9] for solving the static output feedback stabilization problem. As shown in [31], the theoretical advantages of the SLPMM algorithm over the cone complementarity algorithm are the following. First, the SLPMM algorithm always generates a strictly decreasing sequence of the objective function values which is bounded below by an integer n_x, and thus it is convergent. Second, the sequence of iterates generated by the SLPMM is contained in a compact level set, and therefore it is always bounded. Finally, if a corresponding bilinear matrix problem is nonempty, then every accumulation point of the generated sequence solves this bilinear matrix problem. In this case, it is always possible to reconstruct a static output feedback gain from this solution. On the other hand, if there exists no matrix pair satisfying the bilinear matrix problem, then the SLPMM algorithm always terminates with an objective function value of a bilinear matrix inequality minimization problem which is greater than n_x. This indicates for the considered plant that there exists no static output feedback controller. But in this case, one can construct a reduced fixed order dynamic controller from the computed solution of the bilinear matrix inequality minimization problem. For more details, we refer to Leibfritz [31]. In contrast to these strong theoretical convergence results, the authors in [9] can guarantee only that the sequence of a linear approximation of the objective function values is a monotonically (nonstrict) decreasing sequence, which is bounded below by $2n_x$. Moreover, in the last few years a number of numerical procedures have been proposed for solving static $\mathcal{H}_2$ output feedback problems. For example, the LMI-based methods also include the Min-Max algorithm of Geromel, de Souza, and Skelton [16], [17], the XY-centering algorithm of Iwasaki and Skelton [26], and the alternating projection method of Grigoriadis and Skelton [21], but the convergence of these algorithms is not always guaranteed. Particularly, the Min-Max algorithm guarantees only sequences of upper and lower bounds to the maximal and minimal eigenvalues of PQ, which are strictly decreasing, and increasing sequences under strong technical assumptions. In addition, for ensuring the global convergence of the Min-Max algorithm, they must assume that the generated sequences are contained in a compact set. On the other hand, it may occur that the Min-Max algorithm generates an unbounded sequence even if the bilinear matrix problem is nonempty. In this case the Min-Max method breaks down [9], [17]. Since the XY-centering algorithm is closely related to the Min-Max procedure, the global convergence of this approach can be shown only under similar technical assumptions that are as strong as the ones imposed for the Min-Max algorithm [26, Theorem 2]. Finally, the alternating projection method is guaranteed to converge only locally. These observations motivate us to derive the SLPMM algorithm for more complicated problems such as the static $\mathcal{H}_2/\mathcal{H}_\infty$ problem. For this problem class, we will show that the SLPMM procedure behaves theoretically as well as numerically in a similar way to that described above.

The paper is organized as follows. Section 2 defines the considered system realization and describes the static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem considered in this paper. Section 3 contains the necessary and sufficient conditions for the existence of stabilizing static $\mathcal{H}_\infty$ output feedback controllers. Moreover, the formulation of the LMI-based nonconvex optimization problem, which must be necessarily solvable if the static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem has a solution, can be found therein. Section 3.1 presents the main part of this paper. Therein we motivate the nonconvex multiobjective programming problem, and, similarly as in [31], we derive the SLPMM algorithm for finding a numerical solution of this problem class. Thereafter, we discuss the properties and global convergence of this procedure. Finally, in section 4, we present several examples for the design of suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback control laws, which will demonstrate the applicability of the SLPMM algorithm applied to this problem class. We also verify numerically the theoretical results and demonstrate the design of reduced fixed order dynamic controllers if the algorithm terminates with an optimal value greater than $n_x + \mathrm{Tr}(Y^*)$.

We will use the following notation. I_r denotes the $(r \times r)$ identity matrix. The set of real symmetric $(n \times n)$ matrices is denoted by $\mathcal{S}_n$, and $\mathcal{S}_n^+$ describes the cone of symmetric positive definite $(n \times n)$ matrices. For $A \in \mathcal{S}_n$, $A \succ 0$ $(A \succeq 0)$ means that A is positive definite (semidefinite). Similarly, $A \prec 0$ $(A \preceq 0)$ denotes that A is negative definite (semidefinite). For $A, B \in \mathcal{S}_n$, $A \preceq B$ $(A \succeq B)$ denotes the usual Loewner ordering [24]. The symbol $\mathrm{Tr}(A) = \sum_{i=1}^n a_{ii}$ is the trace operator of a matrix $A \in \mathbb{R}^{n \times n}$. $||A||_F$ is the Frobenius norm of a matrix. Finally, $||T_{zw}||_\infty$ denotes the $\mathcal{H}_\infty$ norm of a proper and real rational stable transfer matrix, i.e., $T_{zw} \in \mathcal{RH}_\infty$ [10], and $||T_{zw}||_{\mathcal{H}_2}$ is the usual $\mathcal{H}_2$ norm of a strictly proper and real rational stable transfer matrix, i.e., $T_{zw} \in \mathcal{RH}_2$ [10].

2. The static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem.

In this section, we focus on the static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem as formulated by Doyle et al. [6] for the full order dynamic output feedback case. However, we take a “suboptimal” approach for designing a static output feedback controller which is similar to [29]. The problem solved by Doyle et al. has been shown to be a dual problem of Bernstein and Haddad [2] in some sense; see, for example, Yeh, Banda, and Chang [45]. Khargonekar and Rotea [29] have obtained a nice solution to the dual problem for a class of suboptimal full order output feedback compensators. For example, a convex optimization approach is proposed to solve the full order output feedback dual problem. However, the static output feedback case considered in this paper is much more difficult, since the problem is then no longer convex.

Consider a finite dimensional linear time–invariant plant Σ_P with the state space realization

$$\Sigma_P \left\{ \begin{array}{rcl} \dot{x}(t) &=& Ax(t) + B_0 w_0(t) + B_1 w_1(t) + B_2 u(t), \quad x(0) = 0, \\ z(t) &=& C_1 x(t) + D_{10} w_0(t) + D_{11} w_1(t) + D_{12} u(t), \\ y(t) &=& C_2 x(t) + D_{20} w_0(t) + D_{21} w_1(t), \end{array} \right.$$

where $x(t) \in \mathbb{R}^{n_x}$ is the state, $w_i(t) \in \mathbb{R}^{n_w}$, $i = 0, 1$, are the disturbance inputs, $u(t) \in \mathbb{R}^{n_u}$ is the control input, $z(t) \in \mathbb{R}^{n_z}$ is the regulated output, and $y(t) \in \mathbb{R}^{n_y}$ is the measured output of the control system. The static output feedback controller Σ_C with the state space realization

$$\Sigma_C \left\{ \begin{array}{rcl} u(t) &=& Fy(t) \end{array} \right.$$

is to be designed, where $F \in \mathbb{R}^{n_u \times n_y}$ denotes the unknown static output feedback gain. Substituting Σ_C into the plant, Σ_P yields the corresponding closed loop system given by

$$\Sigma_{cl} \left\{ \begin{array}{rcl} \dot{x}(t) &=& A_F x(t) + B_{0F} w_0(t) + B_F w_1(t), \quad x(0) = 0, \\ z(t) &=& C_F x(t) + D_{0F} w_0(t) + D_F w_1(t), \end{array} \right.$$

where the closed loop matrices $A_F, B_F, C_F, D_F, B_{0F}$, and D_{0F} are defined as follows:

$$A_F = A + B_2 F C_2,\ B_F = B_1 + B_2 F D_{21},\ C_F = C_1 + D_{12} F C_2,\ D_F = D_{11} + D_{12} F D_{21},$$
$$B_{0F} = B_0 + B_2 F D_{20},\ D_{0F} = D_{10} + D_{12} F D_{20}.$$

Throughout the whole paper, the following assumptions are imposed on the system.

ASSUMPTION 2.1.

(1) *The pair* (A, B_2) *is stabilizable, and the pair* (A, C_2) *is detectable.*

(2) *The data matrices in* Σ_P*, especially the following, are real constant matrices:*

$$A \in \mathbb{R}^{n_x \times n_x}, B_1 \in \mathbb{R}^{n_x \times n_w}, B_2 \in \mathbb{R}^{n_x \times n_u}, C_1 \in \mathbb{R}^{n_z \times n_x}, C_2 \in \mathbb{R}^{n_y \times n_x},$$
$$D_{11} \in \mathbb{R}^{n_z \times n_w}, D_{12} \in \mathbb{R}^{n_z \times n_u}, D_{21} \in \mathbb{R}^{n_y \times n_w}, B_0 \in \mathbb{R}^{n_x \times n_w},$$
$$D_{10} \in \mathbb{R}^{n_z \times n_w}, D_{20} \in \mathbb{R}^{n_y \times n_w}.$$

(3) $F \in \mathbb{R}^{n_u \times n_y}$, $n_u < n_x$, $n_y < n_x$, *and* $\operatorname{rank}(B_2) = n_u$, $\operatorname{rank}(C_2) = n_y$.

We start by considering the analysis problem. Assume the static output feedback law Σ_C is fixed such that the closed loop system Σ_{cl} is internally stable. For example, there exists a static output feedback gain in the set

$$\mathcal{F}_s = \{F \in \mathbb{R}^{n_u \times n_y} \mid A_F \text{ is Hurwitz}\}, \tag{2.1}$$

the so-called stability set. Let $T_{zw_1}(s) = C_F(sI - A_F)^{-1}B_F + D_F$ ($T_{zw_0}(s) = C_F(sI - A_F)^{-1}B_{0F} + D_{0F}$), $s \in \mathbb{C}$, denote the closed loop transfer matrix from w_1 to z (from w_0 to z). The concepts of $\mathcal{H}_\infty$ norm and $\mathcal{H}_2$ norm/cost are well known (cf. [7]). Therefore, we will omit detailed discussion and content ourselves with starting the following definitions for reference. Since all coefficient matrices are assumed to be real and $F \in \mathcal{F}_s$ is fixed, the transfer matrix $T_{zw_1} \in \mathcal{RH}_\infty$ and the $\mathcal{H}_\infty$ norm of T_{zw_1} is defined by

$$||T_{zw_1}||_\infty = \sup_{\omega\in\mathbb{R}} \sigma_{\max}(T_{zw_1}(i\omega)), \tag{2.2}$$

where $\sigma_{\max}(T_{zw_1}(\cdot))$ denotes the largest singular value of T_{zw_1} and i denotes the imaginary unit. Recall that $T_{zw_0} \in \mathcal{RH}_2$ and the $\mathcal{H}_2$ norm of T_{zw_0} is finite, for example, $||T_{zw_0}||_{\mathcal{H}_2} < \infty$, if and only if $D_{0F} \equiv 0$. In this case, if L_o denotes the observability Gramian of the pair (A_F, C_F), the $\mathcal{H}_2$ norm of T_{zw_0} can be computed by

$$||T_{zw_0}||^2_{\mathcal{H}_2} = \operatorname{Tr}(B^T_{0F} L_o B_{0F}); \tag{2.3}$$

for example, L_o satisfies $\operatorname{Lyap}(L_o, F) = A^T_F L_o + L_o A_F + C^T_F C_F = 0$.

Now let the scalar $\gamma > 0$ be given and assume that $||T_{zw_1}||_\infty < \gamma$. Define $R_F = I_{n_w} - \gamma^{-2} D^T_F D_F$, and then it is a standard fact (see, for example, [30], [47]), that there exist a unique real symmetric matrix X and a gain $F \in \mathbb{R}^{n_u \times n_y}$ such that $R_F \succ 0$,

$$\begin{aligned}\operatorname{Ric}(X, F) = {}& A^T_F X + X A_F + \gamma^{-1} C^T_F C_F \\ &+ \gamma^{-1}(X B_F + \gamma^{-1} C^T_F D_F) R^{-1}_F (B^T_F X + \gamma^{-1} D^T_F C_F) = 0,\end{aligned} \tag{2.4}$$

and $A_F + \gamma^{-1} B_F R^{-1}_F (B^T_F X + \gamma^{-1} D^T_F C_F)$ is Hurwitz. Moreover, X satisfies (cf. [2])

$$0 \preceq L_o \preceq X \preceq P,$$

where P fulfills $\operatorname{Ric}(P, F) \preceq 0$. Thus, if the $\mathcal{H}_2$ norm of T_{zw_0} is finite, then we have

$$||T_{zw_0}||^2_{\mathcal{H}_2} = \operatorname{Tr}(B^T_{0F} L_o B_{0F}) \le \operatorname{Tr}(B^T_{0F} X B_{0F}) \le \operatorname{Tr}(B^T_{0F} P B_{0F}). \tag{2.5}$$

Similarly as in [2], [46], [29], or [35], these inequalities motivate us to define the following auxiliary $\mathcal{H}_2/\mathcal{H}_\infty$ cost function for the linear time–invariant closed loop system Σ_{cl}:

$$C_\gamma(P, F) = \operatorname{Tr}(B^T_{0F} P B_{0F}), \tag{2.6}$$

which is an upper bound on $||T_{zw_0}||^2_{\mathcal{H}_2}$ if and only if $D_{0F} \equiv 0$. Moreover, $C_\gamma(P, F) \le \operatorname{Tr}(Y)$ whenever the symmetric matrices $P \succ 0$ and $Y \succeq 0, Y \in \mathbb{R}^{n_w \times n_w}$, satisfy $\operatorname{Ric}(P, F) \prec 0$ and

$$\psi(F, P, Y) := \begin{bmatrix} Y & B^T_{0F} P \\ P B_{0F} & P \end{bmatrix} \succeq 0. \tag{2.7}$$

Note that (2.7) is equivalent to $Y \succeq B^T_{0F} P B_{0F} \succeq 0$ and can be stated also as follows:

$$\begin{bmatrix} Y & B^T_0 P \\ P B_0 & P \end{bmatrix} + \begin{bmatrix} 0 \\ P B_2 \end{bmatrix} F \begin{bmatrix} D_{20} & 0 \end{bmatrix} + \begin{bmatrix} D^T_{20} \\ 0 \end{bmatrix} F^T \begin{bmatrix} 0 & B^T_2 P \end{bmatrix} \succeq 0. \tag{2.8}$$

Thus, for given $P \in \mathcal{S}^+_{n_x}$, the matrix inequality (2.8) is an LMI in Y and F.

By this discussion, it is immediate that $||T_{zw_1}||_\infty < \gamma$ and $||T_{zw_0}||^2_{\mathcal{H}_2} < C_\gamma(P, F)$ if and only if $D_{0F} \equiv 0$. Hence, the Riccati inequality $\operatorname{Ric}(P, F) \prec 0$ leads to an $\mathcal{H}_\infty$ norm bound γ and an $\mathcal{H}_2$ cost upper bound $C_\gamma(P, F)$. If $D_{0F} \neq 0$, then $||T_{zw_0}||^2_{\mathcal{H}_2} = \infty$. In this case, we do not interpret the auxiliary cost function $C_\gamma(P,F)$ as an upper bound on the $\mathcal{H}_2$ cost, but we can interpret it as a robust performance measure similar to the results of [46, Theorem 4] and [6, Theorem 1]. The results concerning the optimization of the auxiliary static $\mathcal{H}_2/\mathcal{H}_\infty$ performance measure over the set of all stabilizing static controller gains F satisfying $||T_{zw_1}||_\infty < \gamma$ can be obtained from the following characterization. The proof of this result is very similar to [29, Lemma 2.1] and is thus omitted (see also [6, Theorems 1, 2]).

Lemma 2.2. *Consider the stable closed loop system Σ_{cl} and let T_{zw_1} (T_{zw_0}) denote the closed loop transfer matrix from w_1 to z (from w_0 to z). Let $\gamma > 0$ be given and suppose that T_{zw_0} is strictly proper, i.e., $D_{0F} \equiv 0$. Let $Ric(\cdot)$ be defined by* (2.4). *Then there exists a gain $F \in \mathbb{R}^{n_u \times n_y}$ satisfying $||T_{zw_1}||_\infty < \gamma$ if and only if there exists a pair (P,F), $F \in \mathbb{R}^{n_u \times n_y}$, $P \in \mathcal{S}^+_{n_x}$, such that $R_F \succ 0$ and $Ric(P,F) \prec 0$. In this case,*

$$C_\gamma(P,F) = \inf\{\mathrm{Tr}(B_{0F}^T P B_{0F}) \mid (P,F) \textit{ satisfy } R_F \succ 0, Ric(P,F) \prec 0, P \succ 0\}. \tag{2.9}$$

With the result of Lemma 2.2, our goal is to minimize the auxiliary static $\mathcal{H}_2/\mathcal{H}_\infty$ performance measure over all stabilizing static output feedback gains F that enforce the $\mathcal{H}_\infty$ constraint. From the previous discussion, this is equivalent to minimizing $\mathrm{Tr}(Y)$ over all matrices F, P, and Y satisfying $\mathrm{Ric}(P,F) \prec 0$, $R_F \succ 0$, $P \succ 0$, $Y \succeq 0$, and $\psi(F,P,Y) \succeq 0$. Thus, we state the suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem considered in this paper as the following nonconvex optimization problem:

$$\begin{array}{lll} \min\ \mathrm{Tr}(Y), & \text{subject to (s.t.)} & P \succ 0,\ R_F \succ 0,\ \mathrm{Ric}(P,F) \prec 0,\ Y \succeq 0, \\ & & \psi(F,P,Y) \succeq 0. \end{array} \tag{2.10}$$

In the following sections, we will discuss a procedure for finding solutions to this problem. Note, a solution (P^*, F^*, Y^*) of (2.10), if any exists, is suboptimal in the sense that

$$\begin{aligned} C_\gamma^{opt}(X,F) &= \min\{\mathrm{Tr}(B_{0F}^T X B_{0F}) \mid X \succeq 0, R_F \succ 0, \mathrm{Ric}(X,F) = 0\} \\ &\le \mathrm{Tr}(B_{0F^*}^T P^* B_{0F^*}) \le \mathrm{Tr}(Y^*), \end{aligned} \tag{2.11}$$

where $C_\gamma^{opt}(X,F)$ denotes the minimal value of the "optimal" static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem.

Finally, note that the results in the following sections can be easily extended to the following static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem formulation:

$$\begin{array}{ll} \min\ \mathrm{Tr}(Y), & \text{s.t. } P \succ 0, R_F \succ 0, \mathrm{Ric}(P,F) \prec 0, Y \succeq 0, \psi(F,P,Y) \succeq 0, \\ & \qquad \mathrm{Lyap}(P,F) \prec 0, D_{0F} = 0. \end{array} \tag{2.12}$$

Here, the goal is to minimize an upper estimate of the optimal static $\mathcal{H}_2$ performance subject to the $\mathcal{H}_\infty$ constraint.

3. Suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback design; LMI approach.

This paragraph is devoted to the computational design of a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controller. For deriving the LMI-based formulation of the static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem, we need the following result which can be found, for example, in [25].

Lemma 3.1. *Let $\mathcal{B} \in \mathbb{R}^{n\times m}, \mathrm{rank}(\mathcal{B}) = m < n$, $\mathcal{C} \in \mathbb{R}^{r\times n}, \mathrm{rank}(\mathcal{C}) = r < n$, and $\Omega \in \mathbb{R}^{n\times n}$ be given. Then there exists $F \in \mathbb{R}^{m\times r}$ satisfying $\mathcal{B}F\mathcal{C} + (\mathcal{B}F\mathcal{C})^T + \Omega \prec 0$ if and only if $N(\mathcal{B}^T)^T \Omega N(\mathcal{B}^T) \prec 0$ and $N(\mathcal{C})^T \Omega N(\mathcal{C}) \prec 0$ hold, where $N(\mathcal{B}^T) \in$*

$\mathbb{R}^{n\times(n-m)}$, $N(\mathcal{C}) \in \mathbb{R}^{n\times(n-r)}$, *denoting any matrices whose columns form orthonormal bases of the null spaces of* $\mathcal{B}^T$, $\mathcal{C}$, *respectively.*

By involving Lemma 3.1 and the strict bounded real lemma [47], the suboptimal static $\mathcal{H}_\infty$ output feedback problem, i.e., find a static output feedback gain F, if any exists, such that A_F is Hurwitz and $||T_{zw_1}||_\infty < \gamma$, can be transformed to a problem of solving two LMIs coupled through a bilinear matrix equation. Similar results can be found, for example, in [12], [25], and [30].

THEOREM 3.2 (existence of static $\mathcal{H}_\infty$ controllers). *Let* $\mathcal{F}_s \neq \emptyset$, $\gamma > 0$ *be given and consider the closed loop system* Σ_{cl} *with* $w_0 = 0$. *Then the following are equivalent.*

(i) *There exists a static output feedback gain* $F \in \mathbb{R}^{n_u\times n_y}$ *such that* A_F *is a Hurwitz matrix and* $||T_{zw_1}||_\infty < \gamma$.

(ii) *There exists a pair* (F, P), $F \in \mathbb{R}^{n_u\times n_y}$, $P \in \mathcal{S}^+_{n_x}$, *satisfying*

$$\mathcal{B}F\mathcal{C} + (\mathcal{B}F\mathcal{C})^T + \Omega \prec 0, \tag{3.1}$$

$$\Omega = \begin{bmatrix} A^TP + PA & PB_1 & C_1^T \\ B_1^TP & -\gamma I_{n_w} & D_{11}^T \\ C_1 & D_{11} & -\gamma I_{n_z} \end{bmatrix}, \mathcal{B} = \begin{bmatrix} PB_2 \\ 0 \\ D_{12} \end{bmatrix}, \mathcal{C} = [C_2 \;\; D_{21} \;\; 0].$$

(iii) *There exist matrices* $P \in \mathcal{S}^+_{n_x}$ *and* $Q \in \mathcal{S}^+_{n_x}$ *satisfying* $PQ = I$ *and*

$$N_Q^T \begin{bmatrix} AQ + QA^T + \gamma^{-1}B_1B_1^T & (C_1Q + \gamma^{-1}D_{11}B_1^T)^T \\ (C_1Q + \gamma^{-1}D_{11}B_1^T) & \gamma^{-1}D_{11}D_{11}^T - \gamma I_{n_z} \end{bmatrix} N_Q \prec 0, \tag{3.2}$$

$$N_P^T \begin{bmatrix} A^TP + PA + \gamma^{-1}C_1^TC_1 & PB_1 + \gamma^{-1}C_1^TD_{11} \\ (PB_1 + \gamma^{-1}C_1^TD_{11})^T & \gamma^{-1}D_{11}^TD_{11} - \gamma I_{n_w} \end{bmatrix} N_P \prec 0, \tag{3.3}$$

where $N_Q := N([B_2^T \;\; D_{12}^T])$ *and* $N_P := N([C_2 \;\; D_{21}])$, *denoting any matrices whose columns form orthonormal bases of the null spaces of* $[B_2^T \;\; D_{12}^T]$ *and* $[C_2 \;\; D_{21}]$, *respectively.*

Proof. Combining the strict bounded real lemma [47], Theorem 3.1, and using a Schur complement argument yields the desired result with $Q = P^{-1}$. □

The inequalities (3.2) and (3.3) are LMIs in Q and P, respectively, and are therefore convex. But finding $P \succ 0$ and $Q \succ 0$ satisfying $PQ = I$, (3.2), and (3.3) together is a difficult task since

$$\Phi_{\mathcal{H}_\infty}(P, Q, \gamma) := \{(P, Q) \in \mathcal{S}^+_{n_x} \mid (P, Q) \text{ satisfying } PQ = I, (3.2), (3.3)\} \tag{3.4}$$

is a nonconvex set, i.e., elements in $\Phi_{\mathcal{H}_\infty}(P, Q, \gamma)$ must be inverse to each other. A numerical algorithm for determining a pair $(P, Q) \in \Phi_{\mathcal{H}_\infty}(P, Q, \gamma)$ can be found, for example, in Leibfritz [31].

Similarly as in Theorem 3.2, we can eliminate the matrix variable F from (2.8) by using [25, Theorem 1], which is a generalization of Lemma 3.1. In particular, we have the following lemma.

LEMMA 3.3. *Let* $Y \in \mathbb{R}^{n_w\times n_w}$, $Y \succeq 0$, $P \in \mathcal{S}^+_{n_x}$, *and* $F \in \mathbb{R}^{n_u\times n_y}$. *Moreover, let* $B_0 \in \mathbb{R}^{n_x\times n_w}$, $B_2 \in \mathbb{R}^{n_x\times n_u}$, *and* $D_{20} \in \mathbb{R}^{n_y\times n_w}$ *be given. Suppose* $[0 \;\; PB_2]^T \in \mathbb{R}^{(n_w+n_x)\times n_u}$, *rank*$([0 \;\; PB_2]^T) \leq n_u$, *and* $[D_{20} \;\; 0] \in \mathbb{R}^{n_y\times(n_w+n_x)}$, *rank*$([D_{20} \;\; 0]) \leq n_y$. *Then the following statements are equivalent.*

(i) *There exists a triple* (F, P, Y) *satisfying* (2.8).

(ii) *There exist matrices* $Y \in \mathbb{R}^{n_w \times n_w}$, $Y \succeq 0$, $P \in \mathcal{S}^+_{n_x}$, *and* $Q \in \mathcal{S}^+_{n_x}$ *satisfying* $PQ = I$ *and*

$$\tilde{\mathcal{Q}}(Q,Y) = N([0\ \ B_2^T])^T \begin{bmatrix} Y & B_0^T \\ B_0 & Q \end{bmatrix} N([0\ \ B_2^T]) \succeq 0, \tag{3.5}$$

$$\tilde{\mathcal{P}}(P,Y) = N([D_{20}\ \ 0])^T \begin{bmatrix} Y & B_0^T P \\ PB_0 & P \end{bmatrix} N([D_{20}\ \ 0]) \succeq 0, \tag{3.6}$$

where $N([0\ \ B_2^T])$ *and* $N([D_{20}\ \ 0])$ *denote any matrices whose columns form orthonormal bases of the null spaces of* $[0\ \ B_2^T]$ *and* $[D_{20}\ \ 0]$, *respectively.*

Proof. Observing that

$$N\left(\begin{bmatrix} 0 \\ PB_2 \end{bmatrix}^T\right) = \begin{bmatrix} I_{n_w} & 0 \\ 0 & P^{-1} \end{bmatrix} N\left(\begin{bmatrix} 0 \\ B_2 \end{bmatrix}^T\right)$$

and using [25, Theorem 1], we obtain the desired result with $Q = P^{-1}$. □

The set

$$\begin{aligned} \tilde{\Phi}(P,Q,Y) := \{(P,Q) \in \mathcal{S}^+_{n_x}, Y \in \mathcal{S}_{n_w} \mid (P,Q,Y) \\ \text{satisfying } Y \succeq 0, PQ = I, (3.5), (3.6)\} \end{aligned} \tag{3.7}$$

is not convex due to the coupling condition $PQ = I$.

Obviously, a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ controller in the sense of the previous section exists if and only if condition (ii) of Theorem 3.2 and (2.8) hold for the same gain matrix F. Assuming that there exist matrices $F, P \in \mathcal{S}^+_{n_x}, Y \in \mathcal{S}_{n_w}$ satisfying (3.1) and (2.8), there exist matrices $(P,Q) \in \mathcal{S}^+_{n_x}$ and $Y \in \mathcal{S}_{n_w}$ satisfying condition (iii) of Theorem 3.2 and condition (ii) of Lemma 3.3. Thus, the static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem can be necessarily transformed to the following bilinear optimization problem:

$$\min\ \mathrm{Tr}(Y), \quad \text{s.t.} \quad (P,Q,Y) \in \Phi_{\mathcal{H}_\infty}(P,Q,\gamma) \cap \tilde{\Phi}(P,Q,Y). \tag{3.8}$$

Now suppose that there exists a solution triple (P^*, Q^*, Y^*) of (3.8) (not necessarily unique); then there exists a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controller if and only if there exists an F satisfying (3.1) and (2.8) for $P = P^*$ and $Y = Y^*$. Note, for given (P^*, Q^*, Y^*), this is an LMI feasibility problem in F. Therefore, the suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ problem is solvable if and only if the bilinear optimization problem (3.8) has a solution and the corresponding LMI feasibility problem in F is nonempty. Hence, these observations lead to the following necessary and sufficient conditions for the existence of static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controllers.

THEOREM 3.4. *Let* $\mathcal{F}_s \neq \emptyset$, $\gamma > 0$ *be given, and consider the closed loop system* Σ_{cl}. *Then the following statements are equivalent.*

(i) *There exists a triple* (F,P,Y), $F \in \mathbb{R}^{n_u \times n_y}$, $P \in \mathcal{S}^+_{n_x}$, $Y \in \mathcal{S}_{n_w}$, *and* $Y \succeq 0$ *satisfying* (3.1) *and* (2.8).

(ii) *There exist matrices* $P \in \mathcal{S}^+_{n_x}$, $Q \in \mathcal{S}^+_{n_x}$, $Y \in \mathcal{S}_{n_w}$, *and* $Y \succeq 0$ *satisfying* $PQ = I$, (3.2), (3.3), (3.5), (3.6), *and, for all such fixed* (P,Y), *there is a matrix* $F \in \mathbb{R}^{n_u \times n_y}$ *satisfying* (3.1) *and* (2.8).

Due to this result, the suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ problem can be solved by first finding a solution of (3.8) which is also a necessary condition for the solvability of (2.10). Secondly, if (3.8) has a solution, the static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain, if any exists, can be obtained by solving the corresponding LMI feasibility problem in F. Finally, all gains F which can be reconstructed in this way will be called suboptimal due to the relation $\mathrm{Tr}(Y^*) \geq C_\gamma(P^*, F) \geq C_\gamma^{opt}(X, F)$.

3.1. Bilinear LMI-based algorithm. In this subsection we describe a bilinear, multiobjective LMI-based algorithm for finding a triple (P,Q,Y) satisfying approximatively (3.8). As defined by (3.4) and (3.7), the sets $\Phi_{\mathcal{H}_\infty}(P,Q,\gamma)$ and $\tilde{\Phi}(P,Q,Y)$ are not convex and not closed. To make these sets convex, the coupling constraint $PQ = I$ can be weakened to the following well-known semidefinite programming (SDP)-relaxation:

$$P \succeq Q^{-1} \succ 0 \quad \Longleftrightarrow \quad M(P,Q) := \begin{bmatrix} P & I \\ I & Q \end{bmatrix} \succeq 0. \tag{3.9}$$

Replacing $PQ = I$ in the nonconvex sets $\Phi_{\mathcal{H}_\infty}(P,Q,\gamma)$ and $\tilde{\Phi}(P,Q,Y)$ by $M(P,Q) \succeq 0$ yields a convex approximation of these sets. Moreover, for computational purposes, we prefer to have closed sets. Introducing a positive scalar $\beta > 0$ and replacing the closed loop matrix A_F by $A + \beta I + B_2FC_2$ in Theorem 3.2, we can rewrite the existence conditions of Theorem 3.2 to the following bilinear matrix feasibility problem (cf. Leibfritz [31]).

$$\text{Find } (P,Q) \succ 0, \text{ such that } PQ = I, \quad \mathcal{Q}_\beta(Q,\gamma) \preceq 0, \quad \mathcal{P}_\beta(P,\gamma) \preceq 0, \tag{3.10}$$

where, for given $\beta > 0$ and $\gamma > 0$, we define

$$\mathcal{Q}_\beta(Q,\gamma) = N_Q^T \begin{bmatrix} AQ + QA^T + 2\beta Q + \gamma^{-1}B_1B_1^T & (C_1Q + \gamma^{-1}D_{11}B_1^T)^T \\ C_1Q + \gamma^{-1}D_{11}B_1^T & \gamma^{-1}D_{11}D_{11}^T - \gamma I_{n_z} \end{bmatrix} N_Q, \tag{3.11}$$

$$\mathcal{P}_\beta(P,\gamma) = N_P^T \begin{bmatrix} A^TP + PA + 2\beta P + \gamma^{-1}C_1^TC_1 & PB_1 + \gamma^{-1}C_1^TD_{11} \\ (PB_1 + \gamma^{-1}C_1^TD_{11})^T & \gamma^{-1}D_{11}^TD_{11} - \gamma I_{n_w} \end{bmatrix} N_P, \tag{3.12}$$

and N_Q, N_P are defined as in Theorem 3.2. With these definitions, we replace the nonconvex set $\Phi_{\mathcal{H}_\infty}(P,Q,\gamma)$ by the bilinear approximation

$$\begin{aligned}\bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma,\beta) = \{(P,Q) \in \mathcal{S}_{n_x}^+ \mid (P,Q) \text{ satisfying } PQ = I, \mathcal{Q}_\beta(Q,\gamma) \preceq 0,\\ \mathcal{P}_\beta(P,\gamma) \preceq 0\}\end{aligned} \tag{3.13}$$

and redefine the bilinear optimization problem (3.8) as

$$\min \operatorname{Tr}(Y), \quad \text{s.t.} \quad (P,Q,Y) \in \bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma,\beta) \cap \tilde{\Phi}(P,Q,Y). \tag{3.14}$$

Then, replacing $PQ = I$ by $M(P,Q) \succeq 0$ in (3.7) and (3.13) yields the following closed and linear approximations to the nonconvex and open sets $\bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma)$ and $\tilde{\Phi}(P,Q,Y)$, respectively:

$$\mathcal{X}_\beta(P,Q,\gamma) = \{(P,Q) \in \mathcal{S}_{n_x} \mid M(P,Q) \succeq 0, \mathcal{Q}_\beta(Q,\gamma) \preceq 0, \mathcal{P}_\beta(P,\gamma) \preceq 0\} \tag{3.15}$$

and

$$\begin{aligned}\tilde{\mathcal{X}}(P,Q,Y) = \{(P,Q) \in \mathcal{S}_{n_x}, Y \in \mathcal{S}_{n_w} \mid M(P,Q) \succeq 0, Y \succeq 0, \tilde{\mathcal{Q}}(Q,Y) \succeq 0,\\ \tilde{\mathcal{P}}(P,Y) \succeq 0\}.\end{aligned} \tag{3.16}$$

Observe that the condition $PQ = I$, which is equivalent to $\operatorname{Tr}(PQ) = n_x$, is satisfied if and only if $\operatorname{rank}(M(P,Q)) \equiv n_x$, i.e., the n_x smallest eigenvalues of the positive semidefinite $(2n_x \times 2n_x)$ matrix $M(P,Q)$ are equal to zero. In fact, the constraint $PQ = I$ characterizes the boundary of the set

$$\{(P,Q) \in \mathcal{S}_{n_x} \mid P \succeq Q^{-1} \succ 0\} = \{(P,Q) \in \mathcal{S}_{n_x} \mid M(P,Q) \succeq 0\}.$$

Thus, a feasible matrix triple to the following nonconvex bilinear matrix feasibility problem

$$\begin{gathered} \text{Find } \ (P,Q) \succ 0, Y \succeq 0 \text{ such that} \\ PQ = I,\ \mathcal{Q}_\beta(Q,\gamma) \preceq 0,\ \mathcal{P}_\beta(P,\gamma) \preceq 0,\ \tilde{\mathcal{Q}}(Q,Y) \succeq 0,\ \tilde{\mathcal{P}}(P,Y) \succeq 0 \end{gathered} \tag{3.17}$$

can be obtained by searching for boundary points of the linear and closed set $\mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$, defined by (3.15) and (3.16), respectively. This suggests the following nonconvex bilinear matrix inequality minimization problem:

$$\min\ \mathrm{Tr}(PQ), \quad \text{s.t.} \quad (P,Q,Y) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y). \tag{3.18}$$

Note that there exists a feasible triple (P,Q,Y) satisfying (3.17) if and only if the optimal value of (3.18) is equal to n_x. Since we are interested in a solution triple (P,Q,Y) of the nonconvex problem (3.14), this observation motivates us to define the following bilinear, multiobjective programming problem:

$$\min\ \mathrm{Tr}(PQ) + \mathrm{Tr}(Y), \quad \text{s.t.} \quad (P,Q,Y) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y). \tag{3.19}$$

This problem combines the objective functionals of (3.14) and (3.18). Obviously, minimizing $\mathrm{Tr}(PQ)$ enforces a solution of (3.19) to be close to or on the boundary of the feasible set of (3.19), while minimizing $\mathrm{Tr}(Y)$ drives a solution of (3.19) to be suboptimal for the corresponding static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem.

If the triple (P^*,Q^*,Y^*) is a boundary solution of (3.19) satisfying $P^*Q^* = I$, then we know that (P^*,Q^*,Y^*) is contained in the feasible set of (3.14) and also satisfies

$$n_x + \mathrm{Tr}(Y^*) = \min\{\mathrm{Tr}(PQ) + \mathrm{Tr}(Y) \mid (P,Q,Y) \in \bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma,\beta) \cap \tilde{\Phi}(P,Q,Y)\}.$$

Thus, the optimal value of (3.14) fulfills

$$\begin{aligned} \min\{\mathrm{Tr}(Y) \mid (P,Q,Y) \in \bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma,\beta) \cap \tilde{\Phi}(P,Q,Y)\} - n_x &\leq \mathrm{Tr}(Y^*) - n_x \\ = \ \min\{\mathrm{Tr}(PQ) + \mathrm{Tr}(Y) \mid (P,Q,Y) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)\} &- n_x. \end{aligned}$$

Hence, an optimal solution of (3.19) yields an upper bound to the optimal value of (3.14) and at least a suboptimal solution of (3.14) if and only if the minimal triple (P^*,Q^*,Y^*) of (3.19) satisfies $P^*Q^* = I$. Finally, if and only if $P^*Q^* = I$, then the corresponding suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain can be reconstructed from (P^*,Q^*,Y^*) if and only if the following LMI feasibility problem in F is nonempty.

$$\begin{aligned} &\text{Find } F \in \mathbb{R}^{n_u \times n_y}, \text{ such that} \quad \mathcal{B}F\mathcal{C} + (\mathcal{B}F\mathcal{C})^T + \bar{\Omega} \preceq 0 \quad \text{and} \\ &\begin{bmatrix} Y^* & B_0^T P^* \\ P^* B_0 & P^* \end{bmatrix} + \begin{bmatrix} 0 \\ P^* B_2 \end{bmatrix} F \begin{bmatrix} D_{20} & 0 \end{bmatrix} \\ &\qquad + \begin{bmatrix} D_{20}^T \\ 0 \end{bmatrix} F^T \begin{bmatrix} 0 & B_2^T P^* \end{bmatrix} \succeq 0, \end{aligned} \tag{3.20}$$

where

$$\bar{\Omega} = \begin{bmatrix} A^TP^* + P^*A + 2\beta P^* & P^*B_1 & C_1^T \\ B_1^TP^* & -\gamma I_{n_w} & D_{11}^T \\ C_1 & D_{11} & -\gamma I_{n_z} \end{bmatrix}, \mathcal{B} = \begin{bmatrix} P^*B_2 \\ 0 \\ D_{12} \end{bmatrix}, \mathcal{C} = [C_2 \;\; D_{21} \;\; 0].$$

In this case, the discussion in section 3 implies the existence of a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ controller. It also guarantees that A_F is Hurwitz, $||T_{zw_1}||_\infty < \gamma$, and $\mathrm{Tr}(Y^*) \geq \mathrm{Tr}(B_{0F}^TP^*B_{0F}) \geq C_\gamma^{opt}(X,F)$. Thus, such an F is indeed a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain.

In what follows we explain a numerical procedure for determining an optimal solution of the bilinear, multiobjective programming problem (3.19). This problem is not convex since the functional $\mathrm{Tr}(PQ)$ is, in general, not convex, but it is bilinear. Therefore, in problem (3.19) we minimize a combination of a bilinear and linear matrix functional over a closed convex set. To solve such a problem, a sequential linearization programming approach as proposed by [1] for general nonconvex bilinear programming problems can be used. In particular, the idea of this approach is very simple. Instead of solving the nonconvex problem (3.19) directly, we linearize the bilinear part of the objective functional. Then we minimize successively the resulting (linearized) LMI constrained semidefinite programming problems.

Algorithm 1 (SLPMM).

For given $\beta > 0$, let $\mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y) \neq \emptyset$.

```
(0) Determine $(P^0, Q^0, Y^0) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$.
    For $k = 0, 1, 2, \ldots$ do
      (1) Determine $(U^k, V^k, Z^k)$ as the unique solution of
          (3.21)  $\min \mathrm{Tr}(PQ^k + P^kQ) + \mathrm{Tr}(Y)$,
                  s.t. $(P,Q,Y) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$.
      (2) If $\mathrm{Tr}(U^kQ^k + P^kV^k) + \mathrm{Tr}(Z^k) = 2\mathrm{Tr}(P^kQ^k) + \mathrm{Tr}(Y^k)$, $\longrightarrow$ Stop.
      (3) Compute $\alpha \in [0,1]$ by solving
          (3.22)  $\min_{\alpha\in[0,1]} \mathrm{Tr}((P^k + \alpha(U^k - P^k))(Q^k + \alpha(V^k - Q^k)))$
                      $+ \mathrm{Tr}(Y^k + \alpha(Z^k - Y^k))$.
      (4) Set $P^{k+1} = (1-\alpha)P^k + \alpha U^k$, $Q^{k+1} = (1-\alpha)Q^k + \alpha V^k$, and $Y^{k+1} =$
          $(1-\alpha)Y^k + \alpha Z^k$.
```

This algorithm is similar to the SLPMM procedure proposed by Leibfritz [31] for the design of stabilizing static $\mathcal{H}_2$ and suboptimal static $\mathcal{H}_\infty$ output feedback controllers. The initialization step of Algorithm 1 is an LMI feasibility problem, and (3.21) is an SDP problem with a linear objective functional under LMI constraints. There are many algorithms available for solving such kinds of problems. For example, interior point methods developed for SDPs can be used (cf. [13]). Algorithm 1 terminates if the first order necessary minimum principle is satisfied at a (local) minimum of (3.19) [34, section 6.1]. For example, define

$$\mathcal{J}(P,Q,Y) := \mathrm{Tr}(PQ) + \mathrm{Tr}(Y). \tag{3.23}$$

Note that $\mathcal{J}(P,Q,Y)$ is continuous and differentiable on the cone of symmetric matrices. Hence,

$$\mathcal{J}'(P,Q,Y)(U,V,Z) = \mathrm{Tr}(UQ + PV) + \mathrm{Tr}(Z),$$
$$\mathcal{J}''(P,Q,Y)(U,V,Z)(H,G,W) = \mathrm{Tr}(HV + UG)$$

for any $U, V, H, G \in \mathcal{S}_{n_x}$ and $Z, W \in \mathcal{S}_{n_w}$. If (P^*, Q^*, Y^*) denotes a local minimum of $\mathcal{J}$ over the closed convex set $\mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$, then for any feasible direction $(P-P^*, Q-Q^*, Y-Y^*) = (\delta P, \delta Q, \delta Y)$ at (P^*, Q^*, Y^*) we obtain

$$\begin{aligned}\mathcal{J}'(P^*,Q^*,Y^*)(\delta P,\delta Q,\delta Y) = \operatorname{Tr}(PQ^*+P^*Q)+\operatorname{Tr}(Y)\\ - 2\operatorname{Tr}(P^*Q^*)-\operatorname{Tr}(Y^*)\geq 0\end{aligned} \tag{3.24}$$

for all $(P,Q,Y) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$ [34, section 6.1, Proposition 1]. Numerically, for a sufficiently small scalar $\varepsilon > 0$, we terminate the algorithm if

$$\tau_k := \operatorname{Tr}(U^kQ^k + P^kV^k) + \operatorname{Tr}(Z^k) - 2\operatorname{Tr}(P^kQ^k) - \operatorname{Tr}(Y^k) \geq -\varepsilon, \quad k \geq 0. \tag{3.25}$$

Moreover, if $(P^k, Q^k, Y^k) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$ and $(U^k, V^k, Z^k) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$ do not satisfy (3.24), we determine a step size parameter $\alpha \in [0,1]$ by minimizing (3.22) with respect to α. Finally, the new iterates $(P^{k+1}, Q^{k+1}, Y^{k+1})$ are convex combinations of (P^k, Q^k, Y^k) and (U^k, V^k, Z^k), respectively, and therefore, they are also contained in $\mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$. Note that the points $(U^k - P^k, V^k - Q^k, Z^k - Y^k)$, $k \geq 0$, are descent directions for $\mathcal{J}$ at (P^k, Q^k, Y^k) unless $\tau_k \geq 0$ for some $k \geq 0$.

For proofing the convergence of Algorithm 1, we need the following result. Suppose for fixed $\beta > 0$ that there exist matrices $(P^0, Q^0, Y^0) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$; then we can define the following level set:

$$\begin{aligned}\Gamma(P^0,Q^0,Y^0) := \{(P,Q,Y) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y) \mid \mathcal{J}(P,Q,Y)\\ \leq \mathcal{J}(P^0,Q^0,Y^0)\}.\end{aligned} \tag{3.26}$$

For this level set, we conclude the following lemma.

LEMMA 3.5. *Let $\beta > 0$ be given, $\mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$ be nonempty, and let $(P^0, Q^0, Y^0) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$ be given. Then the level set $\Gamma(P^0, Q^0, Y^0)$ is compact.*

Proof. Using [24, Theorem 7.4.10], the closeness of $\mathcal{X}_\beta(P,Q,\gamma)$, $\tilde{\mathcal{X}}(P,Q,Y)$, and the definition of $\Gamma(P^0, Q^0, Y^0)$ shows the desired result. □

With the compactness of the level set $\Gamma(P^0, Q^0, Y^0)$, it is straightforward to show the existence of an optimal solution of (3.19) in this level set. In particular, we have the following lemma.

LEMMA 3.6. *Let $\mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y) \neq \emptyset$ and $(P^0, Q^0, Y^0) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$ be given. Then there exists an optimal solution (P^*, Q^*, Y^*) of* (3.19) *in the level set $\Gamma(P^0, Q^0, Y^0)$.*

Proof. By the compactness of $\Gamma(P^0, Q^0, Y^0)$, the continuity of $\mathcal{J}$, and the theorem of Bolzano–Weierstrass, the result follows immediately. □

Thus, Lemma 3.6 ensures the existence of a solution of the multiobjective programming problem (3.19) at least in the compact level set $\Gamma(P^0, Q^0, Y^0)$.

The following lemma is needed in the proof of the Theorem 3.8 given below.

LEMMA 3.7. *Let $a > 0$, $b < 0$, and $c \geq 1$ be given. Then there exists $\rho_* \geq 0$ such that the optimal solution of* $\min_{\alpha \in [0,1]} \, (c + \alpha\, b + (\alpha^2/2)\, a)$ *satisfies*

$$0 < \alpha_* = -\frac{b + \rho_*}{a} \leq 1.$$

Moreover, $\rho_* = 0$ if and only if $a \geq -b$.

Proof. The result can be proven by using the convexity of the objective function and the Karush–Kuhn–Tucker theorem. □

The next result states the basic convergence properties of Algorithm 1. It also justifies the fact that we need only the existence of a solution of (3.19) in $\Gamma(P^0, Q^0, Y^0)$. Before stating these properties, we define the boundary of the set $\mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$, given by

$$\partial\mathcal{X}(P, Q, Y) := \{(P, Q, Y) \in \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y) \mid PQ = I\}. \tag{3.27}$$

Obviously,

$$\partial\mathcal{X}(P, Q, Y) = \bar{\Phi}_{\mathcal{H}_\infty}(P, Q, \gamma, \beta) \cap \tilde{\Phi}(P, Q, Y) \subseteq \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y).$$

Hence, the level set $\Gamma(P^0, Q^0, Y^0)$ also contains all (P, Q, Y) on the boundary of $\mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$. Therefore, similarly as in Lemma 3.6, we can conclude that there also exists an optimal solution of the nonconvex optimization problem (3.14) in the level set $\Gamma(P^0, Q^0, Y^0)$.

THEOREM 3.8. *Let $\beta > 0$ and $(P^0, Q^0, Y^0) \in \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$ be given. Furthermore, assume that $\{(P^k, Q^k, Y^k)\} \subset \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$ is generated by Algorithm* 1. *Then the sequence $\{(P^k, Q^k, Y^k)\}$ is well defined and for all $k \geq 0$, we have*

$$n_x + \mathrm{Tr}(Y^*) \leq n_x + \mathrm{Tr}(Y^{k+1}) \leq \mathcal{J}(P^{k+1}, Q^{k+1}, Y^{k+1}) < \mathcal{J}(P^k, Q^k, Y^k) \tag{3.28}$$

unless $\mathcal{J}'(\cdot)(\cdot) = 0$, where (P^*, Q^*, Y^*) solves (3.19). *Thus, $\{\mathcal{J}(P^k, Q^k, Y^k)\}$ converges to $\hat{\mathcal{J}} \geq n_x + \mathrm{Tr}(Y^*)$ and for all $k \geq 0$, $(P^k, Q^k, Y^k) \in \Gamma(P^0, Q^0, Y^0)$, i.e., $\{(P^k, Q^k, Y^k)\}$ is bounded. Finally, $\mathcal{J}(P^k, Q^k, Y^k) = n_x + \mathrm{Tr}(Y^k)$ if and only if $(P^k, Q^k, Y^k) \in \partial\mathcal{X}(P, Q, Y)$ and $\mathcal{J}'(\cdot)(\cdot) = 0$ for some k.*

Proof. For all $k \geq 0$, we define the following abbreviations:

$$S^k = (P^k, Q^k, Y^k), \quad \delta S^k = (U^k - P^k, V^k - Q^k, Z^k - Y^k), \quad \text{and} \quad T^k = (U^k, V^k, Z^k).$$

If Algorithm 1 terminates at $S^k \in \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$, $k \geq 0$, then

$$\mathcal{J}'(S^k)(S^k) = \mathcal{J}'(S^k)(T^k) = \min_{S \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)} \mathcal{J}'(S^k)(S)$$

and $\mathcal{J}'(P^k, Q^k, Y^k)(P - P^k, Q - Q^k, Y - Y^k) \geq 0$ for all $(P, Q, Y) \in \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$. Thus, if the algorithm does not terminate in step (2), then

$$\mathcal{J}'(S^k)(\delta S^k) = \mathrm{Tr}(U^k Q^k + P^k V^k) + \mathrm{Tr}(Z^k) - 2\mathrm{Tr}(P^k Q^k) - \mathrm{Tr}(Y^k) < 0, \tag{3.29}$$

i.e., δS^k is a descent direction. Using $S^k \in \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$, $T^k \in \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$, and the convexity of $\mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$, steps (3) and (4) of Algorithm 1 imply that

$$(P^{k+1}, Q^{k+1}, Y^{k+1}) \in \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$$

for all $\alpha \in [0, 1]$. Hence, the sequence $\{S^k\} = \{(P^k, Q^k, Y^k)\}$ is well defined by Algorithm 1.

To show the strictly decreasing property of $\{\mathcal{J}(P^k, Q^k, Y^k)\}$, note that for $\alpha \in [0, 1]$, the construction of the algorithm implies

$$\begin{aligned}
\mathcal{J}(S^{k+1}) &= \mathcal{J}(S^k) + \alpha_*\, \mathcal{J}'(S^k)(\delta S^k) + \frac{\alpha_*^2}{2}\mathcal{J}''(S^k)(\delta S^k)(\delta S^k) \\
&\leq \mathcal{J}(S^k) + \alpha\, \mathcal{J}'(S^k)(\delta S^k) + \frac{\alpha^2}{2}\mathcal{J}''(S^k)(\delta S^k)(\delta S^k) \\
&\leq \mathcal{J}(S^k) + \mathcal{J}'(S^k)(\delta S^k) + \frac{1}{2}\mathcal{J}''(S^k)(\delta S^k)(\delta S^k),
\end{aligned} \tag{3.30}$$

where $\alpha_* \in [0, 1]$ denotes the optimal value of (3.22).

Assuming $\mathcal{J}''(S^k)(\delta S^k)(\delta S^k) = 0$ and using (3.29), from (3.30) we obtain $\mathcal{J}(S^{k+1}) < \mathcal{J}(S^k)$ for all $k \geq 0$ unless $\mathcal{J}'(S^k)(\delta S^k) \equiv 0$. Moreover, if $\mathcal{J}''(S^k)(\delta S^k)(\delta S^k) < 0$, then we can conclude that $\mathcal{J}(S^{k+1}) < \mathcal{J}(S^k)$ for all $k \geq 0$. Finally, if $\mathcal{J}''(S^k)(\delta S^k)(\delta S^k) > 0$, then the problem is convex. Defining $a = \mathcal{J}''(\cdot)(\cdot)(\cdot)$, $b = \mathcal{J}'(\cdot)(\cdot)$, and $c = \mathcal{J}(\cdot)$, Lemma 3.7 implies the existence of $\rho_* \geq 0$ such that the solution α_* of (3.22) satisfies

$$0 < \alpha_* = -\frac{\mathcal{J}'(S^k)(\delta S^k) + \rho_*}{\mathcal{J}''(S^k)(\delta S^k)(\delta S^k)} \leq 1.$$

But this implies

$$\frac{1}{2}\alpha_*^2\, \mathcal{J}''(S^k)(\delta S^k)(\delta S^k) = -\frac{1}{2}\alpha_*\, \mathcal{J}'(S^k)(\delta S^k) - \frac{1}{2}\alpha_*\rho_*. \tag{3.31}$$

Using $\alpha_* \in (0, 1]$, $\rho_* \geq 0$, (3.29), (3.30), and (3.31), we obtain $\mathcal{J}(S^{k+1}) < \mathcal{J}(S^k)$ unless $\mathcal{J}'(S^k)(\delta S^k) \equiv 0$ for all $k \geq 0$. Hence, $\{\mathcal{J}(P^k, Q^k, Y^k)\}$ is a strictly decreasing sequence unless $\mathcal{J}'(S^k)(\delta S^k) \equiv 0$. Since $\mathrm{Tr}(Y^k) \geq 0$ and $\mathrm{Tr}(P^k Q^k) \geq n_x$ for all $k \geq 0$, the sequence $\{\mathcal{J}(P^k, Q^k, Y^k)\}$ is also bounded below by $n_x + \mathrm{Tr}(Y^*)$. Therefore, it converges to a limit $\hat{\mathcal{J}} \geq n_x + \mathrm{Tr}(Y^*)$. Moreover, by the definition of the compact level set $\Gamma(P^0, Q^0, Y^0)$, we know that $\{(P^k, Q^k, Y^k)\} \subset \Gamma(P^0, Q^0, Y^0)$, and thus that the generated sequence $\{(P^k, Q^k, Y^k)\}$ is bounded. Finally, assume that $(P^k, Q^k, Y^k) \in \partial\mathcal{X}(P, Q, Y)$ and $\mathcal{J}'(S^k)(\delta S^k) = 0$ for some k. Note that this is fulfilled if and only if $P^k Q^k = I$ and the minimal point T^k of (3.21) satisfy $U^k = P^k$, $V^k = Q^k$, and $Z^k = Y^k$ for some k. Thus, $\mathcal{J}(P^k, Q^k, Y^k) = n_x + \mathrm{Tr}(Y^k) \geq n_x + \mathrm{Tr}(Y^*)$ if and only if $(P^k, Q^k, Y^k) \in \partial\mathcal{X}(P, Q, Y)$ and $\mathcal{J}'(S^k)(\delta S^k) = 0$ for some k. □

Theorem 3.8 ensures that $\{\mathcal{J}(P^k, Q^k, Y^k)\}$ is a strictly decreasing sequence which is bounded below by $n_x + \mathrm{Tr}(Y^*)$ if and only if $(P^k, Q^k, Y^k) \notin \partial\mathcal{X}(P, Q, Y)$ and $\mathcal{J}'(S^k)(\delta S^k) < 0$. On the other hand, $\mathcal{J}(P^k, Q^k, Y^k) = n_x + \mathrm{Tr}(Y^k) \geq n_x + \mathrm{Tr}(Y^*)$ if and only if $(P^k, Q^k, Y^k) \in \partial\mathcal{X}(P, Q, Y)$ and $\mathcal{J}'(S^k)(\delta S^k) = 0$ for some $k \geq 0$. Then we know that $P^k Q^k = I$ and $U^k = P^k$, $V^k = Q^k$, $Z^k = Y^k$ for some k. Hence, in this case, the SLPMM algorithm terminates in step (2) at a point satisfying the coupling condition $PQ = I$.

The SLPMM algorithm may be interpreted as a modified version of the cone complementarity algorithm, expect that in each iteration we must also compute a step size parameter. But this further computational work is essential for the strictly decreasing property of the SLPMM. Indeed, the novel aspect of the SLPMM algorithm is that it always generates a strictly decreasing sequence $\{\mathcal{J}(P^k, Q^k, Y^k)\}$. Moreover, this approach guarantees the boundedness of the iterates (P^k, Q^k, Y^k) for all $k \geq 0$. In contrast to this, the cone complementarity algorithm, the XY-centering algorithm, and other related computational methods in the literature do not share these properties.

The strictness in the inequality (3.28) and the boundedness of $\{(P^k, Q^k, Y^k)\}$ is essential to prove the global convergence of the SLPMM algorithm. The following theorem states the global convergence of our method under rather weak assumptions compared to the existing algorithms in the literature.

Theorem 3.9. *Let $\beta > 0$ and $(P^0, Q^0, Y^0) \in \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$ be given. Furthermore, assume that $\{(P^k, Q^k, Y^k)\} \subset \mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$ is generated by Algorithm* 1. *Then the following hold.*

(i) *The algorithm terminates at some (P^k, Q^k, Y^k) satisfying* (3.24), *or every accumulation point $(\tilde{P}, \tilde{Q}, \tilde{Y})$ of $\{(P^k, Q^k, Y^k)\}$ is stationary, i.e., $(\tilde{P}, \tilde{Q}, \tilde{Y})$ satisfy* (3.24).

(ii) *Every accumulation point* $(\tilde{P},\tilde{Q},\tilde{Y})$ *of* $\{(P^k,Q^k,Y^k)\}$ *satisfies the nonconvex bilinear matrix feasibility problem* (3.17) *and solves the nonconvex optimization problem* (3.14) *if and only if the boundary set* $\partial\mathcal{X}(P,Q,Y)$ *defined in* (3.27) *is nonempty.*

Proof. (i) This result follows immediately by making straightforward modifications to the proof of Leibfritz [31, Theorem 3.7].

(ii) Note that $\partial\mathcal{X}(P,Q,Y)=\bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma,\beta)\cap\tilde{\Phi}(P,Q,Y)\subset\Gamma(P^0,Q^0,Y^0)$ and that $\Gamma(P^0,Q^0,Y^0)$ is compact. Theorem 3.8 implies $\lim_{k\to\infty}\mathcal{J}(P^k,Q^k,Y^k)=\hat{\mathcal{J}}\geq n_x+\mathrm{Tr}(Y^*)=:\mathcal{J}^*$. Since $(P^k,Q^k,Y^k)\in\Gamma(P^0,Q^0,Y^0)$, we also know that the functions $\mathcal{J}(P,Q,Y)=\mathrm{Tr}(PQ)+\mathrm{Tr}(Y)$ and $\tilde{\mathcal{J}}(P,Q,Y)=\mathrm{Tr}(Y)$ attain the (global) minimum values on the compact level set $\Gamma(P^0,Q^0,Y^0)$; i.e., $\mathcal{J}^*=\mathcal{J}(P^*,Q^*,Y^*)\equiv n_x+\mathrm{Tr}(Y^*)$ and $\tilde{\mathcal{J}}^*=\tilde{\mathcal{J}}(P^*,Q^*,Y^*)\equiv\mathrm{Tr}(Y^*)$, respectively.

Suppose $\hat{\mathcal{J}}>\mathcal{J}^*=n_x+\mathrm{Tr}(Y^*)$. From the compactness of $\Gamma(P^0,Q^0,Y^0)\neq\emptyset$ and the continuity of $\mathcal{J}$, it follows that for every $\varepsilon>0$, there exist a $\delta>0$ and a vicinity $U_\delta(P^*,Q^*,Y^*)$ such that $U_\delta(P^*,Q^*,Y^*)\cap\Gamma(P^0,Q^0,Y^0)\neq\emptyset$ and

$$0\leq\mathcal{J}(P,Q,Y)-\mathcal{J}^*<\varepsilon$$

for all $(P,Q,Y)\in U_\delta(P^*,Q^*,Y^*)\cap\Gamma(P^0,Q^0,Y^0)$. Choosing $\varepsilon:=\hat{\mathcal{J}}-n_x-\mathrm{Tr}(Y^*)>0$ implies

$$\mathcal{J}(P,Q,Y)<\varepsilon+n_x+\mathrm{Tr}(Y^*)=\hat{\mathcal{J}}.$$

But this implies the existence of an integer $\hat{k}$ such that for all $k\geq\hat{k}$ and $(P^k,Q^k,Y^k)\in U_\delta(P^*,Q^*,Y^*)\cap\Gamma(P^0,Q^0,Y^0)$ we have $\mathcal{J}(P^k,Q^k,Y^k)<\hat{\mathcal{J}}$, which contradicts that the sequence $\{\mathcal{J}(P^k,Q^k,Y^k)\}$ converges monotonically to $\hat{\mathcal{J}}$. Hence $\hat{\mathcal{J}}=\mathcal{J}^*=n_x+\mathrm{Tr}(Y^*)$; i.e.,

$$\lim_{k\to\infty}\mathcal{J}(P^k,Q^k,Y^k)=\mathcal{J}^*\iff\lim_{k\to\infty}(\mathcal{J}(P^k,Q^k,Y^k)-n_x)=\tilde{\mathcal{J}}^*. \tag{3.32}$$

Since $\{(P^k,Q^k,Y^k)\}\subset\Gamma(P^0,Q^0,Y^0)$ and $\Gamma(P^0,Q^0,Y^0)$ is compact, we conclude the existence of a convergent subsequence of $\{(P^k,Q^k,Y^k)\}$; i.e.,

$$\lim_{j\to\infty}(P^{k_j},Q^{k_j},Y^{k_j})=(\tilde{P},\tilde{Q},\tilde{Y})\in\Gamma(P^0,Q^0,Y^0).$$

Suppose that this accumulation point is not globally optimal. Then $\mathcal{J}(\tilde{P},\tilde{Q},\tilde{Y})>\mathcal{J}^*=n_x+\mathrm{Tr}(Y^*)>\tilde{\mathcal{J}}^*$ and the sequence $\{\mathcal{J}(P^k,Q^k,Y^k)-n_x-\mathrm{Tr}(Y^*)\}$ do not tend to zero. This contradicts (3.32). Hence, $\partial\mathcal{X}(P,Q,Y)\neq\emptyset$ if and only if every accumulation point satisfies (3.17) and solves

$$\begin{aligned}\min\{\mathrm{Tr}(PQ)+\mathrm{Tr}(Y)\mid(P,Q,Y)\in\bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma,\beta)\cap\tilde{\Phi}(P,Q,Y)\}&=n_x+\mathrm{Tr}(Y^*)\\ =n_x+\min\{\mathrm{Tr}(Y)\mid(P,Q,Y)\in\bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma,\beta)\cap\tilde{\Phi}(P,Q,Y)\},&\end{aligned}$$

which in turn is equivalent to the solvability of (3.14). □

From Theorem 3.8, we know that the sequence $\{\mathcal{J}(P^k,Q^k,Y^k)\}$ is strictly decreasing unless $\mathcal{J}'(\cdot)(\cdot)=0$ and is bounded below by $n_x+\mathrm{Tr}(Y^*)$, and hence converges. If in the limit $\mathrm{Tr}(PQ)=n_x$, then $\partial\mathcal{X}(P,Q,Y)\neq\emptyset$. In this case, Theorem 3.9 guarantees the existence of an accumulation point which satisfies the nonconvex bilinear matrix feasibility problem (3.17). Moreover, this point is also a solution of the nonconvex optimization problem (3.14), and therefore, in the limit the necessary condition for the existence of a static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain is fulfilled.

Note that for reconstructing such a gain, it is not necessary that the computed solution of the multiobjective problem (3.19) is the global solution of this problem. It suffices that this solution is contained in the boundary set $\partial\mathcal{X}(P,Q,Y)$ defined by (3.27), i.e., it satisfies $PQ = I$. Then, this solution triple is at least an upper bound to the global optimal value of (3.14), and the corresponding gain, if any exists, is at least suboptimal. Thus, if Algorithm 1 terminates with a boundary solution of (3.19), i.e., $\mathcal{J}(P^k, Q^k, Y^k) = n_x + \mathrm{Tr}(Y^k)$ for some $k \geq 0$, then Theorem 3.8 implies $(P^k, Q^k, Y^k) \in \partial\mathcal{X}(P,Q,Y)$ for some k. Moreover, $(P^k, Q^k, Y^k) \in \partial\mathcal{X}(P,Q,Y)$ satisfies

$$\min\{\mathrm{Tr}(Y) \mid (P,Q,Y) \in \bar{\Phi}_{\mathcal{H}_\infty}(P,Q,\gamma,\beta) \cap \tilde{\Phi}(P,Q,Y)\} \leq \mathrm{Tr}(Y^k).$$

Then, using the results of the previous section, a corresponding suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controller exists if and only if the LMI feasibility problem (3.20) in F is nonempty. On the other hand, it may occur that Algorithm 1 terminates at a triple (P^k, Q^k, Y^k) satisfying the first order necessary minimum principle, but $(P^k, Q^k, Y^k) \notin \partial\mathcal{X}(P,Q,Y)$. In this case, we know that $\mathrm{Tr}(P^kQ^k) > n_x$ and a static gain F can not be reconstructed from this triple. But we can reconstruct a reduced dynamic output feedback control law of order $n_c \leq n_x - \max\{n_u, n_y\}$ from this triple by using a well-known system augmentation technique. For more details, we refer the reader, for example, to Leibfritz [31].

4. Numerical examples. In this section, several examples are given for test purposes in order to test the SLPMM approach. We present examples for the design of suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controllers.

The SLPMM algorithm as well as the reconstruction of the controller gain F from the corresponding LMIs have been implemented making use of MATLAB 5.0 facilities. Particularly, for determining a feasible solution $(P^0, Q^0, Y^0) \in \mathcal{X}_\beta(P,Q,\gamma) \cap \tilde{\mathcal{X}}(P,Q,Y)$ in Algorithm 1, step (0), we have used the LMI control toolbox [14] routine FEASP, which finds a solution to a given system of LMIs, if any exists. Moreover, for solving the linearized minimization problem (3.21) of Algorithm 1, step (1), which is a semidefinite programming problem, we have taken the LMI control toolbox procedure MINCX. This solver is an implementation of Nesterov and Nemirovski's projective method for minimizing a linear objective function under LMI constraints as described in [13]. For MINCX we have adjusted the desired relative accuracy on the optimal value to 10^{-12}.

We terminated Algorithm 1 if, for a sufficiently small scalar $\varepsilon > 0$, the condition (3.25) was fulfilled. Moreover for the computation of a step size $\alpha \in [0,1]$ we have used the MATLAB function FMIN.

Finally, we have taken the LMI control toolbox function FEASP applied to the LMI feasibility problem (3.20) for reconstructing the static output feedback controller gain F.

The following data are given in the tables: the iteration counter k of the SLPMM algorithm; if $k = 0$, j denotes the total number of iterations for finding a feasible initial point (P^0, Q^0, Y^0) by FEASP; else j is the total number of iterations for solving the semidefinite programming problem (3.21) by MINCX with relative accuracy of 10^{-12}; the objective function value $\mathcal{J}(P^k, Q^k, Y^k) = \mathrm{Tr}(P^kQ^k) + \mathrm{Tr}(Y^k)$ of the corresponding multiobjective optimization problem (3.19); the function values $\mathrm{Tr}(P^kQ^k)$ and $\mathrm{Tr}(Y^k)$; and τ_k, which indicates if the first order necessary minimum principle is approximatively satisfied.

TABLE 4.1

Convergence behavior of the SLPMM algorithm for the VTOL helicopter model: static $\mathcal{H}_2/\mathcal{H}_\infty$ case.

k	j	$\mathcal{J}(P^k, Q^k, Y^k)$	$\mathrm{Tr}(P^kQ^k)$	$\mathrm{Tr}(Y^k)$	τ_k
0	6	1209.96168	1156.3660285	53.5956538	—
1	53	4.21991536	4.0001534809	0.21976188	$-2.058e+03$
2	45	4.21973458	4.0000000000	0.21973458	$-1.808e-04$
3	18	4.21973458	4.0000000000	0.21973458	$0.000e+00$

For all test examples we have chosen the data matrices B_0, D_{10}, D_{11}, and D_{20} as follows:

$$B_0 = B_1, \quad D_{10} = D_{11} = 0, \quad D_{20} = 0.$$

This choice guarantees $D_{0F} = 0$. Therefore, if a static $\mathcal{H}_2/\mathcal{H}_\infty$ controller gain F exists for the corresponding test problem, it is an upper estimate of $||T_{zw_0}||^2_{\mathcal{H}_2}$ that enforces $||T_{zw_1}||_\infty < \gamma$.

Example 1. A state space model of the longitudinal motion of a VTOL helicopter is considered as in [28]. The dynamic equations of the helicopter model are linearized around a nominal solution where the given dynamic equation is computed for typical loading and flight conditions of the VTOL helicopter at a certain airspeed [41]. The linearization results in a fourth order linear time-invariant state equation with two control and two unknown input components. The data matrices of the linearized model are given by

$$A = \begin{bmatrix} -0.0366 & 0.0271 & 0.0188 & -0.4555 \\ 0.0482 & -1.0100 & 0.0024 & -4.0208 \\ 0.1002 & 0.3681 & -0.7070 & 1.4200 \\ 0 & 0 & 1.0000 & 0 \end{bmatrix}, \; B_1 = \begin{bmatrix} 0.0468 & 0 \\ 0.0457 & 0.0099 \\ 0.0437 & 0.0011 \\ -0.0218 & 0 \end{bmatrix},$$

$$B_2 = \begin{bmatrix} 0.4422 & 0.1761 \\ 3.5446 & -7.5922 \\ -5.5200 & 4.4900 \\ 0 & 0 \end{bmatrix},$$

$$C_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \; C_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}, \; D_{12} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$D_{21} = \begin{bmatrix} 0.00039 & 0.00174 \end{bmatrix}.$$

The goal is to design a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controller Σ_C by the SLPMM Algorithm 1 according to the discussion of the previous sections. In the run of Algorithm 1 we have chosen $\beta = 0.01$, $\gamma = 0.423722$, and $\varepsilon = 10^{-10}$. Table 4.1 demonstrates the convergence behavior of the SLPMM. It illustrates numerically that $\{\mathcal{J}(P^k, Q^k, Y^k)\}$ is a strictly decreasing sequence if $(P^k, Q^k, Y^k) \notin \partial\mathcal{X}(P, Q, Y)$ and $\tau_k < 0$, which converges to $n_x + \mathrm{Tr}(Y^*) = 4 + 0.21973458$. Thus, we have $\partial\mathcal{X}(P, Q, Y) \neq \emptyset$, and equality holds if and only if $(P^k, Q^k, Y^k) \in \partial\mathcal{X}(P, Q, Y)$ and $\tau_k = 0$ for some k according to Theorem 3.8. Moreover, Theorem 3.9 guarantees the existence of an accumulation point of $\{P^k, Q^k, Y^k)\}$ which also solves the nonconvex optimization problem (3.14). Therefore, a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controller exists if and only if the LMI feasibility problem (3.20) in F is nonempty. The resulting static gain matrix which satisfies (3.20) and the closed loop eigenvalues are

$$F = \begin{bmatrix} -0.978987 & 19.43483 \end{bmatrix}^T, \lambda(A_F) = \begin{bmatrix} -152.04 & -0.0989 & -0.2990 \pm 0.949i \end{bmatrix}.$$

Moreover, $||T_{zw_1}||_\infty = 0.2937866 < \gamma$ and $\mathrm{Tr}(B_{0F}^T P B_{0F}) = 0.02869950 \leq \mathrm{Tr}(Y) = 0.21973458$, which shows that F is a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain. To compare this solution with the optimal one, we have calculated $C_\gamma^{opt}(X, F) = 0.0231044$ with an algorithm proposed by Leibfritz [32], [30]. This shows, that (3.19) provides a good suboptimal solution to the problem under consideration. As it can be also verified, the sequence $\{(P^k, Q^k, Y^k) \subset \Gamma(P^0, Q^0, Y^0)$ and is thus bounded. For example, we have

$$\begin{aligned} \{||P^k||\} &= \{82.26, 49.82, 49.82, 49.82\}, \\ \{||Q^k||\} &= \{41.16, 1.159, 1.159, 1.159\}, \\ \{||Y^k||\} &= \{26.86, 0.1099, 0.1099, 0.1099\}, \end{aligned}$$

which demonstrates numerically the boundedness of the generated sequence according to Theorem 3.8. The whole computation needs 2.47 CPU seconds on a SUN Ultra 60.

Finally, for this example, we demonstrate numerically that the extended static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem (2.12) can be also solved by the LMI approach as discussed in section 3. The SLPMM terminates after four outer iterations with an approximate solution (P, Q, Y) of the corresponding bilinear programming problem, i.e., $\mathrm{Tr}(PQ) = 4.0000001$. Then, using this solution, the corresponding LMI feasibility problem in F was found to be nonempty. Thus, a static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain F satisfying (2.12) exists, and the resulting gain matrix is given by $F = [-0.457876 \;\; 17.38107]^T$.

Example 2 *(transport airplane).* This example studies the longitudinal motion of a modern transport airplane under VMIN flight conditions [15]. The linearized state space model yields the following data matrices:

$$A = \begin{bmatrix} -0.06254 & 0.01888 & 0 & -0.56141 & -0.02751 & 0 & 0.06254 & -0.00123 & 0 \\ 0.01089 & -0.99280 & 0.99795 & 0.00097 & -0.07057 & 0 & -0.01089 & 0.06449 & 0 \\ 0.07743 & 1.67540 & -1.31111 & -0.00030 & -4.25030 & 0 & -0.07743 & -0.10883 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -20.0000 & 20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -30 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.88206 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.88206 & 0.00882 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.00882 & -0.88206 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1.3282 & 0 & 0 & 0 \\ 0 & 1.62671 & 0 & 0 \\ 0 & -68.75283 & 0 & 0 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 30 \\ 0 \\ 0 \\ 0 \end{bmatrix}, C_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} -0.00519 \\ 0.47604 \\ 0.00098 \\ -0.00031 \\ 0.03378 \\ 0 \\ 0.00519 \\ -0.03086 \\ 0 \end{bmatrix}^T,$$

$$C_2^T = \begin{bmatrix} -0.00519 & 0 \\ 0.47604 & 0 \\ 0.00098 & 0 \\ -0.00031 & 0 \\ 0.03378 & 0 \\ 0 & 0 \\ 0.00519 & 0 \\ -0.03086 & 0 \\ 0 & 0 \end{bmatrix}, D_{21}^T = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix},$$

and $D_{12} = [1/\sqrt{2}]$. The goal is the design of a robust static control law. Defining $\beta = 10^{-2}$, $\gamma = 0.152032$ and $\varepsilon = 10^{-12}$, Table 4.2 illustrates the convergence behavior of the SLPMM for this example. According to Theorem 3.8, it shows numerically

Table 4.2
Convergence behavior of the SLPMM algorithm for the transport airplane: static $\mathcal{H}_2/\mathcal{H}_\infty$ case.

k	j	$\mathcal{J}(P^k, Q^k, Y^k)$	$\mathrm{Tr}(P^k Q^k)$	$\mathrm{Tr}(Y^k)$	τ_k
0	18	$1.766e+14$	$1.766e+14$	$7.622e+07$	—
1	58	131.706754	83.966174369	47.74057934352	$-3.533e+14$
2	49	9.80419099	9.4602075306	0.343983455585	$-1.689e+02$
3	56	9.39445809	9.0504993630	0.343958730681	$-2.843e-01$
4	86	9.34395876	9.0000000338	0.343958722207	$-4.318e-02$
5	24	9.34395876	9.0000000338	0.343958722207	$0.000e+00$

the strictly decreasing property of the objective function values of (3.19) and the nonemptiness of the boundary of the set $\mathcal{X}_\beta(P, Q, \gamma) \cap \tilde{\mathcal{X}}(P, Q, Y)$, i.e., in the limit we achieve $\mathrm{Tr}(PQ) = n_x$. Thus, $\partial\mathcal{X}(P, Q, Y) \neq \emptyset$ and Theorem 3.9 ensures the existence of an accumulation point of $\{P^k, Q^k, Y^k)\}$ which also solves the nonconvex optimization problem (3.14). Therefore, a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controller exists if and only if the LMI feasibility problem (3.20) in F is nonempty. Solving the LMI feasibility problem (3.20) results in the feasible gain

$$F = \begin{bmatrix} 2.286293 & 0.001023 \end{bmatrix}.$$

The real parts of the closed loop poles range between -0.02686 and -33.4213. Moreover, $||T_{zw_1}||_\infty = 0.08088935$ and $\mathrm{Tr}(B_{0F}^T P B_{0F}) = 0.05135721 \leq \mathrm{Tr}(Y) = 0.34395872$. Once again, comparing this solution with the optimal one, we have calculated $C_\gamma^{opt}(X, F) = 0.0502692$ with the algorithm of [32] for solving the optimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback problem (2.11). Finally, the whole computation needs 53.53 CPU seconds. Again, this example demonstrates numerically the theoretical properties of our algorithm and shows that this approach can be used for the design of a suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback gain.

Example 3 *(Euler–Bernoulli beam).* This example consists of a simple supported Euler–Bernoulli beam as discussed in [22] (see also [18]). Following [22], the state space model is given by Σ_P, with matrices as follows:

$$A = \mathrm{diag}\left\{ \begin{bmatrix} 0 & 1 \\ -r^4 & -0.02\, r^2 \end{bmatrix}, \ r = 1, \dots, 5 \right\},$$

$$B_2^T = C_2 = \begin{bmatrix} 0 & 0.9877 & 0 & -0.309 & 0 & -0.891 & 0 & 0.5878 & 0 & 0.7071 \end{bmatrix}, B_1 = \begin{bmatrix} B_2 & 0_{10\times 1} \end{bmatrix},$$

$$C_1 = \begin{bmatrix} 0.809 & 0 & -0.9511 & 0 & 0.309 & 0 & 0.5878 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$D_{12} = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}, D_{21} = \begin{bmatrix} 0 & 1.9 \end{bmatrix}.$$

Choosing $\beta = 0.01$, $\gamma = 3.59251$, and $\varepsilon = 10^{-10}$, Algorithm 1 terminates after 28.59 CPU seconds, and (3.20) provides the static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controller gain $F = -0.6221049$ with $\mathrm{Tr}(B_{0F}^T P B_{0F}) = 0.5742053 \leq \mathrm{Tr}(Y) = 4.4202$ and $||T_{zw_1}||_\infty = 2.179593$. Moreover, the closed loop poles are

$$\lambda(A_F) = \begin{bmatrix} -0.395 \pm 24.9i & -0.257 \pm 15.9i & -0.326 \pm 8.98i & -0.305 \pm 0.95 & -0.059 \pm 3.99i \end{bmatrix}.$$

Using the algorithm proposed by [30], [32] yields the optimal cost value $C_\gamma^{opt}(X, F) = 0.185282$ of (2.11) with $F_{opt} = -1.036985$ and $||T_{zw_1}||_\infty = 2.14176$.

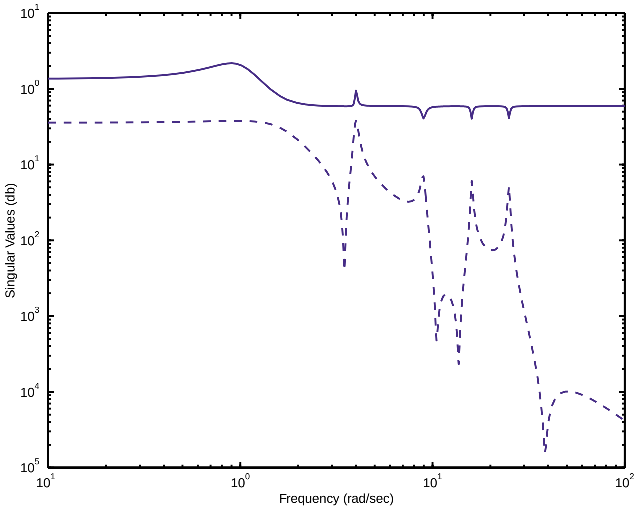


FIG. 4.1. *Singular value plot.*

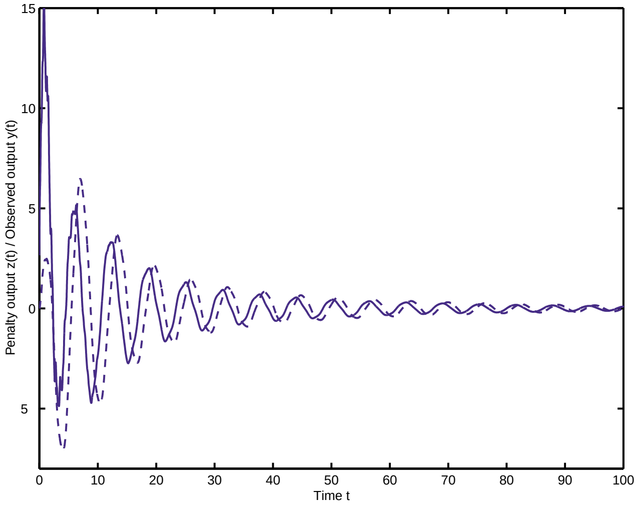


FIG. 4.2. *Phase portraits of the regulated output z (dashed line) and the observed output y (solid line) for the Euler–Bernoulli beam under worst case disturbances: Controlled case.*

The SLPMM algorithm has generated the sequences

$$\{\mathcal{J}(P^k, Q^k, Y^k)\}_{k=0}^4 = \{229.5443, 14.4202522, 14.4202000, 14.4202000, 14.4202000\},$$
$$\{\mathrm{Tr}(P^k Q^k)\}_{k=0}^4 = \{180.32, 10.000029, 10.0000000001, 10.0000000000, 10.0000000000\},$$
$$\{\tau_k\}_{k=1}^4 = \{-3.281 \cdot 10^2, -5.224 \cdot 10^{-5}, -1.278 \cdot 10^{-10}, 0.000\},$$

which underline the theoretical results stated in Theorems 3.8 and 3.9.

Figure 4.1 shows the singular value response of the corresponding closed loop system for the problem under consideration. In Figure 4.2, the phase portraits for the regulated output variable z (dashed line) and the observed output variable y

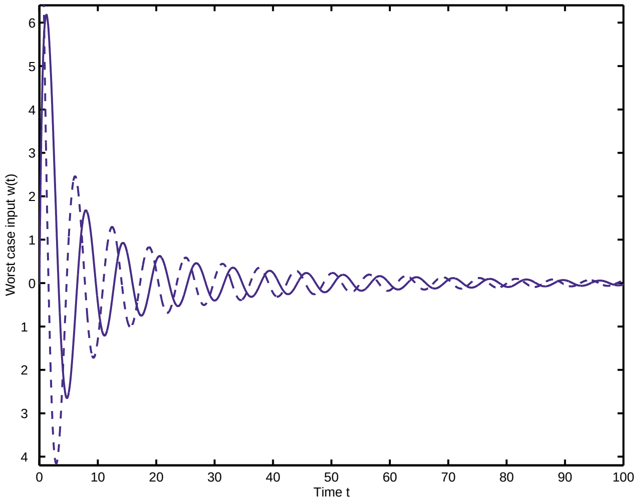


FIG. 4.3. *Worst case disturbances.*

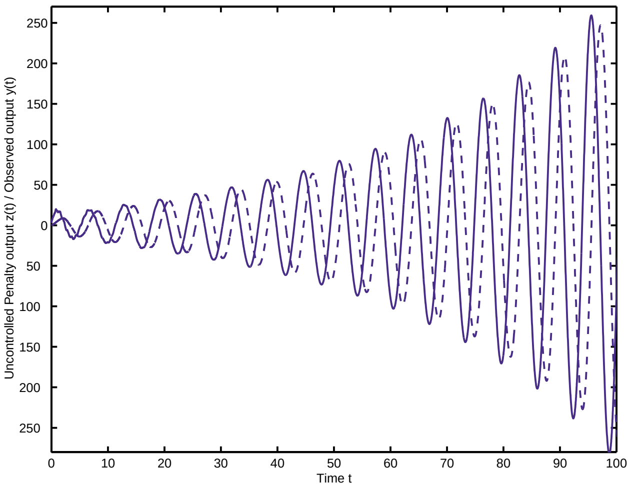


FIG. 4.4. *Phase portraits of the regulated output z (dashed line) and the observed output y (solid line) for the Euler–Bernoulli beam under worst case disturbances: Uncontrolled case.*

(solid line) are displayed for the computed suboptimal static $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback controller under the corresponding worst case inputs w, as illustrated in Figure 4.3. These plots demonstrate that the computed controller gain yields an asymptotically stable closed loop system and satisfies the robustness constraint, even if the worst-case input affects the system. In contrast to this, the uncontrolled system outputs, as shown in Figure 4.4, oscillate away from the stable equilibrium of the system. Especially, the worst case inputs support this behavior and drive the system to be unstable.

Example 4 *(reduced order control).* In this example we consider a three–mass-spring system and illustrate our approach for the design of a reduced order compensator if the SLPMM algorithm terminates at a point (P, Q, Y) satisfying $\mathrm{Tr}(PQ) > n_x$. The system under consideration consists of three unit masses connected by two linear springs with stiffness constant κ. There we have assumed that only the position of the third mass is measured and a control force acts on the first mass. The linear state space realization of this system is given by Σ_P with the following data matrices:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -\kappa & 0 & \kappa & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \kappa & 0 & -2\kappa & 0 & \kappa & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & \kappa & 0 & -\kappa & 0 \end{bmatrix}, B_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, C_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}^T, D_{12} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$C_1 = [I_6 \;\; 0]^T$, $D_{21} = 0$, and the state components x_i, x_{i+1}, $i = 1, 2, 3$, denote the position and the velocity of the mass i, respectively. It is well known that this linear system is not stabilizable by a static output feedback gain, i.e., $\mathcal{F}_s = \emptyset$, but it is stabilizable by a reduced order controller of order $n_c = 3$, where

$$\Sigma_{CD} \;\{\; \dot{x}_c(t) = A_c x_c(t) + B_c y(t), \quad u(t) = C_c x_c(t) + D_c y(t)$$

denotes the compensator of order $n_c \leq n_x$ with $A_c \in \mathbb{R}^{n_c \times n_c}$, $B_c \in \mathbb{R}^{n_c \times n_y}$, $C_c \in \mathbb{R}^{n_u \times n_c}$, $D_c \in \mathbb{R}^{n_u \times n_y}$, and $x_c(t) \in \mathbb{R}^{n_c}$. Using Algorithm 1 with $\beta = 0.1$, $\gamma = 50.0869$, and $\varepsilon = 10^{-8}$, we can compute a reduced order compensator of the form Σ_{CD} for this example as follows. Since $\mathcal{F}_s = \emptyset$, the SLPMM algorithm provides a solution triple (P, Q, Y) of (3.19) satisfying $\mathrm{Tr}(PQ) > n_x$, i.e., $(P, Q, Y) \notin \partial\mathcal{X}(P, Q, Y)$. In particular, the algorithm has generated the following sequences:

$$\begin{aligned} \{\mathcal{J}(P^k, Q^k, Y^k)\}_{k=0}^5 &= \{10192.8935, 48.3013702, 38.2048006, 37.5019180, 36.7597188, \\ &\quad\; 36.7596674\}, \\ \{\mathrm{Tr}(P^k Q^k)\}_{k=0}^5 &= \{9786.7919, 39.2714948, 30.9417654, 30.8152693, 30.7274670, \\ &\quad\; 30.7274665\}, \\ \{\mathrm{Tr}(Y^k)\}_{k=0}^5 &= \{406.1016, 9.0298754, 7.2630352, 6.6866487, 6.0322518, \\ &\quad\; 6.0322009\}, \\ \{\tau_k\}_{k=1}^5 &= \{-1.84 \cdot 10^4, -9.50, -1.24 \cdot 10^{-2}, -9.12 \cdot 10^{-3}, -3.19 \cdot 10^{-9}\}. \end{aligned}$$

Obviously, $\{\mathcal{J}(P^k, Q^k, Y^k)\}$ is a strictly decreasing sequence which tends to $\mathrm{Tr}(P^5 Q^5)$ $+ \mathrm{Tr}(Y^5) = 30.7274665 + 6.0322009 > n_x + \mathrm{Tr}(Y^*)$, and Algorithm 1 terminates after six iterations satisfying approximatively the first order necessary condition. Since $\partial\mathcal{X}(P, Q, Y) = \emptyset$, we know that there exists no static gain for the problem under consideration. But it is possible to reconstruct a reduced order compensator from the computed solution triple (P, Q, Y) of (3.19). For example, first we compute the eigenvalues of $P - Q^{-1} \succeq 0$. If there are m eigenvalues of $P - Q^{-1}$ which are less than or equal to $\tilde{\varepsilon} > 0$, $\tilde{\varepsilon} \ll 1$, then there exists an n_cth order dynamic output feedback control law of the form Σ_{CD} such that the eigenvalues of the augmented closed loop system matrix $\hat{A}_{\hat{F}} = \hat{A} + \hat{B}_2 \hat{F} \hat{C}_2$ have negative real parts [8, Theorem 3.1], where $n_c = n_x - m$ and the augmented system matrices are defined by [12]:

$$\begin{aligned} &\hat{A} = \begin{bmatrix} A & 0 \\ 0 & 0_{n_c} \end{bmatrix}, \; \hat{B}_1 = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \; \hat{B}_2 = \begin{bmatrix} 0 & B_2 \\ I_{n_c} & 0 \end{bmatrix}, \; \hat{C}_1 = [C_1 \;\; 0], \\ &\hat{C}_2 = \begin{bmatrix} 0 & I_{n_c} \\ C_2 & 0 \end{bmatrix}, \; \hat{D}_{12} = [0 \; D_{12}], \; \hat{D}_{21} = \begin{bmatrix} 0 \\ D_{21} \end{bmatrix}, \; \hat{F} = \begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix}, \end{aligned} \tag{4.1}$$

$\hat{B}_0 = \hat{B}_1$, $\hat{D}_{20} = [0\ 0_{n_y \times n_w}]^T$, and $\hat{D}_{10} = \hat{D}_{11} = D_{11}$. Second, we decompose the positive semidefinite matrix $P - Q^{-1}$ as $UU^T + E$ by a singular value decomposition, where $U \in \mathbb{R}^{n_x \times n_c}$, $E \in \mathbb{R}^{n_x \times n_x}$, $||E|| \leq \tilde{\varepsilon}$, and we define

$$\hat{P} = \left[\begin{array}{cc} P & U \\ U^T & I_{n_c} \end{array} \right], \quad \hat{Y} = Y. \tag{4.2}$$

By [36, Lemma 7.5] we have $\hat{P} \succ 0$. Then, replacing in (3.20) the system matrices by their augmented counterparts, the corresponding LMI feasibility problem in $\hat{F}$ has a solution. Choosing $\tilde{\varepsilon} = 10^{-10}$ yields $m = 3$, $n_c = 3$, and the resulting third order compensator gain

$$\hat{F} = \left[\begin{array}{cc} A_c & B_c \\ C_c & D_c \end{array} \right] = \left[\begin{array}{rrr|r} -0.23619 & -0.59416 & 1.85859 & -2.31612 \\ 0.52718 & -0.14806 & 1.38309 & -1.58075 \\ -0.51506 & -0.91840 & -0.64521 & 2.59614 \\ \hline -0.05706 & -0.01872 & 0.55757 & -1.17127 \end{array} \right].$$

The real parts of closed loop eigenvalues of $\hat{A}_{\hat{F}}$ ranges between -0.00963 and -0.02874. Moreover, $\mathrm{Tr}(\hat{B}_{0\hat{F}}^T \hat{P} \hat{B}_{0\hat{F}}) = 6.032201$ and $||T_{zw_1}||_\infty = 46.46845 < \gamma$, which shows numerically that $\hat{F}$ is a suboptimal third order $\mathcal{H}_2/\mathcal{H}_\infty$ output feedback compensator gain. Summing up, this example demonstrates that we always can construct from the computed solution of Algorithm 1 at least a reduced order compensator gain which satisfies the design criteria. Finally, the whole computation time needs 9.62 CPU seconds on a SUN Ultra 60.

REFERENCES

[1] K. P. Bennett and O. L. Mangasarian, *Bilinear separation of two sets in n–space,* Comput. Optim. Appl., 2 (1993), pp. 207–227.

[2] D. S. Bernstein and W. M. Haddad, *LQG control with an $\mathcal{H}_\infty$ performance bound: A Riccati equation approach,* IEEE Trans. Automat. Control, 34 (1989), pp. 293–305.

[3] S. P. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[4] M. Chilali and P. Gahinet, *$\mathcal{H}_\infty$ design with pole placement constraints: An LMI approach*, IEEE Trans. Automat. Control, 41 (1996), pp. 358–367.

[5] J. Doyle, A. Packard, and K. Zhou, *Review of LFTs, LMIs and μ*, in Proceedings of the 30th Conference on Decision and Control, Brighton, England, 1991, pp. 1227–1232.

[6] J. Doyle, K. Zhou, K. Glover, and B. Bodenheimer, *Mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ performance objectives* II: *Optimal control*, IEEE Trans. Automat. Control, 39 (1994), pp. 1575–1587.

[7] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[8] L. El Ghaoui and P. Gahinet, *Rank minimization under LMI constraints: A framework for output feedback problems*, in Proceedings of the European Control Conference, Groningen, The Netherlands, 1993, pp. 1176–1179.

[9] L. El Ghaoui, F. Oustry, and M. AitRami, *A cone complementarity linearization algorithm for static output feedback and related problems*, IEEE Trans. Automat. Control, 42 (1997), pp. 1171–1176.

[10] B. Francis, *A Course in $\mathcal{H}_\infty$ Control Theory,* Lecture Notes in Control and Inform. Sci. 88, Springer-Verlag, New York, London, Paris, 1987.

[11] P. Gahinet, *Explicit controller formulas for LMI-based $\mathcal{H}_\infty$ synthesis*, Automatica J. IFAC, 32 (1996), pp. 1007–1014.

[12] P. Gahinet and P. Apkarian, *A linear matix inequality approach to $\mathcal{H}_\infty$ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[13] P. Gahinet and A. Nemirovski, *The projective method for solving linear matrix inequalities*, Math. Programming, 77 (1997), pp. 163–190.

[14] P. Gahinet, A. Nemirovski, A. J. Laub, and M. Chilali, *LMI Control Toolbox; For Use with MATLAB*, The Math Works Inc., Natick, MA, 1995.

[15] D. Gangsaas, K. R. Bruce, J. D. Blight, and U.-L. Ly, *Application of modern synthesis to aircraft control: Three case studies*, IEEE Trans. Automat. Control, 31 (1986), pp. 995–1014.

[16] J. C. Geromel, C. C. de Souza, and R. E. Skelton, *LMI numerical solution for output feedback stabilization*, in Proceedings of the American Control Conference, Baltimore, MD, 1994, pp. 40–44.

[17] J. C. Geromel, C. C. de Souza, and R. E. Skelton, *Static output feedback controllers: Stability and convexity*, IEEE Trans. Automat. Control, 43 (1998), pp. 120–125.

[18] J. C. Geromel and P. B. Gapski, *Synthesis of positive real $\mathcal{H}_2$ controllers*, IEEE Trans. Automat. Control, 42 (1997), pp. 988–992.

[19] J. C. Geromel, P. L. D. Peres, and S. R. Souza, *A convex approach to the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem for discrete-time uncertain systems*, SIAM J. Control Optim., 33 (1995), pp. 1816–1833.

[20] A. Giusto, A. Trofino-Neto, and E. B. Castelan, *$\mathcal{H}_\infty$ and $\mathcal{H}_2$ design techniques for a class of prefilters*, IEEE Trans. Automat. Control, 41 (1996), pp. 864–871.

[21] K. M. Grigoriadis and R. E. Skelton, *Low order control design for LMI problems using alternating projection methods*, Automatica J. IFAC, 32 (1996), pp. 1117–1125.

[22] W. M. Haddad, D. S. Bernstein, and Y. W. Wang, *Dissipative $\mathcal{H}_2/\mathcal{H}_\infty$ controller synthesis*, IEEE Trans. Automat. Control, 49 (1994), pp. 827–831.

[23] W. M. Haddad, V. Kapila, and D. S. Bernstein, *Robust $\mathcal{H}_\infty$ stabilization via parameterized lyapunov bounds*, IEEE Trans. Automat. Control, 42 (1997), pp. 241–248.

[24] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[25] T. Iwasaki and R. E. Skelton, *All controllers for the general $\mathcal{H}_\infty$ control problem: LMI existence conditions and state space formulas*, Automatica J. IFAC, 30 (1994), pp. 1307–1317.

[26] T. Iwasaki and R. E. Skelton, *The XY-centering algorithm for the dual LMI problem: A new approach to fixed-order control design*, Internat. J. Control, 62 (1995), pp. 1257–1272.

[27] I. Kaminer, P. P. Khargonekar, and M. A. Rotea, *Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control for discrete time systems via convex optimization*, Automatica J. IFAC, 29 (1993), pp. 57–70.

[28] L. H. Keel, S. P. Bhattacharyya, and J. W. Howze, *Robust control with structured perturbations*, IEEE Trans. Automat. Control, 33 (1988), pp. 68–77.

[29] P. P. Khargonekar and M. A. Rotea, *Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control: A convex optimization approach*, IEEE Trans. Automat. Control, 36 (1991), pp. 824–837.

[30] F. Leibfritz, *Static Output Feedback Design Problems*, Shaker Verlag, Aachen, Germany, 1998.

[31] F. Leibfritz, *Computational Design of Stabilizing Static Output Feedback Controllers*, Technical Report Trierer Forschungsberichte Mathematik/Informatik 99–01, Universität Trier, Trier, Germany, 1999.

[32] F. Leibfritz, *Static Output Feedback Design by Using a Newton-SQP Interior Point Method*, Technical Report Trierer Forschungsberichte Mathematik/Informatik 99–03, Universität Trier, Trier, Germany, 1999.

[33] D. J. N. Limebeer, B. D. O. Anderson, and B. Hendel, *A Nash game approach to mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control*, IEEE Trans. Automat. Control, 39 (1994), pp. 69–82.

[34] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Menlo Park, London, Sydney, 1973.

[35] D. Mustafa, *Combined $\mathcal{H}_\infty$/LQG control via the optimal projection equations: On minimizing the LQG cost bound*, Internat. J. Robust Nonlinear Control, 1 (1991), pp. 99–109.

[36] A. Packard, K. Zhou, P. Pandey, and G. Becker, *A collection of robust control problems leading to LMIs*, in Proceedings of the 30th Conference on Decision and Control, Brighton, England, 1991, pp. 1245–1250.

[37] M. A. Rotea, *The generalized $\mathcal{H}_2$ control problem*, Automatica J. IFAC, 29 (1993), pp. 373–385.

[38] M. A. Rotea and P. P. Khargonekar, *$\mathcal{H}_2$-optimal control with an $\mathcal{H}_\infty$-constraint: The state feedback case*, Automatica J. IFAC, 27 (1991), pp. 307–316.

[39] M. A. Rotea and P. P. Khargonekar, *Generalized $\mathcal{H}_2/\mathcal{H}_\infty$ control*, in Robust Control Theory, IMA Vol. Math. Appl. 66, B. A. Francis and P. P. Khargonekar, eds., Springer-Verlag, New York, 1995, pp. 81–103.

[40] A. Saberi, B. M. Chen, S. Peddapullaiah, and U.-L. Ly, *Simultaneous $\mathcal{H}_2/\mathcal{H}_\infty$ optimal control: The state feedback case*, Automatica J. IFAC, 29 (1993), pp. 1611–1614.

[41] S. N. Singh and A. R. Coelho, *Nonlinear control of mismatched uncertain linear systems and application to control of aircraft*, Journal of Dynamic Systems, Measurement and Control, 106 (1984), pp. 203–210.

[42] R. E. Skelton, J. Stoustrup, and T. Iwasaki, *The $\mathcal{H}_\infty$ control problem using static output feedback*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 449–455.

[43] V. L. Syrmos, C. T. Abdallah, P. Dorato, and K. Grigoriadis, *Static output feedback—a survey*, Automatica J. IFAC, 33 (1997), pp. 125–137.

[44] I. Yaesh and U. Shaked, *Minimum entropy static output feedback control with an $\mathcal{H}_\infty$ norm performance bound*, IEEE Trans. Automat. Control, 42 (1997), pp. 853–858.

[45] H.-H. Yeh, S. S. Banda, and B.-C. Chang, *Necessary and sufficient conditions for mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ optimal control*, IEEE Trans. Automat. Control, 37 (1992), pp. 355–358.

[46] K. Zhou, K. Glover, B. Bodenheimer, and J. Doyle, *Mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ performance objectives* I: *Robust performance analysis*, IEEE Trans. Automat. Control, 39 (1994), pp. 1564–1574.

[47] K. Zhou and P. P. Khargonekar, *An algebraic Riccati equation approach to $\mathcal{H}_\infty$ optimization*, Systems Control Lett., 11 (1988), pp. 85–91.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1736–1747

RIESZ BASIS APPROACH TO THE STABILIZATION OF A FLEXIBLE BEAM WITH A TIP MASS*

BAO-ZHU GUO†

Abstract. Using an abstract condition of Riesz basis generation of discrete operators in the Hilbert spaces, we show, in this paper, that a sequence of generalized eigenfunctions of an Euler–Bernoulli beam equation with a tip mass under boundary linear feedback control forms a Riesz basis for the state Hilbert space. In the meanwhile, an asymptotic expression of eigenvalues and the exponential stability are readily obtained. The main results of [*SIAM J. Control Optim.*, 36 (1998), pp. 1962–1986] are concluded as a special case, and the additional conditions imposed there are removed.

Key words. beam equation, boundary control, stability, eigenvalues, Riesz basis

AMS subject classifications. 93C20, 93D15, 35B35, 35P10

PII. S0363012999354880

1. Introduction.

When a tip mass is attached to the free end, the vibration of a flexible beam that is clamped at one end and controlled at the free end can be described by the following Euler–Bernoulli beam equation (Conrad and Morgül, 1998):

$$(1)\qquad \begin{cases} y_{tt}(x,t) + y_{xxxx}(x,t) = 0, \quad 0 < x < 1, \quad t > 0, \\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = 0, \quad t \geq 0, \\ -y_{xxx}(1,t) + my_{tt}(1,t) = u(t), \quad t \geq 0, \end{cases}$$

where y is the amplitude of the vibration, m is the tip mass, and u is the boundary control force applied at the free end. In order to achieve uniform stability for this system, one has to employ "higher" derivative controllers. The following linear feedback control law is proposed in Conrad and Morgül (1998):

$$u(t) = -\alpha y_t(1,t) + \beta y_{xxxt}(1,t), \quad t \geq 0,$$

where α and β are real constants. The closed-loop system then becomes

$$(2)\qquad \begin{cases} y_{tt}(x,t) + y_{xxxx}(x,t) = 0, \\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = 0, \\ y_{xxx}(1,t) = \alpha y_t(1,t) + my_{tt}(1,t) - \beta y_{xxxt}(1,t). \end{cases}$$

The energy multiplier method is used in Conrad and Morgül (1998) to show that system (2) is exponentially stable for any $\alpha, \beta > 0$. It is further proved for a special case where $m = \alpha\beta$ that a set of generalized eigenfunctions of system (2) forms a Riesz basis for the state Hilbert space, usually referred to as the Riesz basis (generation) property, and that the spectrum-determined growth condition holds, both for almost

*Received by the editors April 16, 1999; accepted for publication (in revised form) September 27, 2000; published electronically February 28, 2001. This research was supported by the National Key Project of China and the National Natural Science Foundation of China. This work was done while the author was working in the Department of Electrical and Electronic Engineering at the University of Pretoria, South Africa.

http://www.siam.org/journals/sicon/39-6/35488.html

†Institute of Systems Science, Academy of Mathematics and System Sciences, Academia Sinica, Beijing 100080, China (bzguo@iss03.iss.ac.cn).

every $\alpha > 0$. Systems with the Riesz basis property are usually referred to as Riesz spectral systems (Xu and Sallet, 1996).

Verification of the Riesz basis property is a very important problem both theoretically and practically. Usually, the property will lead to the establishment of such results as the spectrum-determined growth condition and the exponential stability of the system. However, such verification is usually difficult because the associated system operator is non-self-adjoint. For one-dimensional string equations with general variable coefficients under linear boundary feedback control, successful treatments have been made for the basis property in last two decades; we refer to Cox and Zuazua (1994), Shubov (1996, 1997), and the references therein. The case of a string equation with a tip mass was investigated in Morgül, Conrad, and Rao (1994). An abstract treatment of general Riesz spectral system with one rank perturbation can be found in Sun (1981), Rebarber (1989), and Xu and Sallet (1996), to name just a few. In Rao (1997) and, recently, Li et al. (1999), the beam equations with "low order" perturbations were considered. Since the model of serially connected Euler–Bernoulli beams under joint linear feedback control was proposed in Chen et al. (1987), many efforts have been made to study the asymptotic distribution of the eigenvalues (Chen et al., 1989) and the exponential stability (Rebarber, 1995). However, the spectrum-determined growth condition had not been reported until Conrad (1990), where a cantilevered beam equation was shown to have the Riesz basis property for small feedback gain, and hence the spectrum-determined growth condition is concluded for this special case. The general case for this cantilevered beam equation was resolved partly by Conrad and Morgül (1998).

In these works mentioned above, the verification of Riesz basis generation relies upon Bari's theorem (see, for example, Gohberg and Krein, 1969): if $\{\phi_n\}_1^\infty$ is a Riesz basis for a Hilbert space H, and $\{\psi_n\}_1^\infty$, an ω-linearly independent sequence in H, is quadratically close to $\{\phi_n\}_1^\infty$ in the sense that

$$\sum_{n=1}^{\infty} \parallel \phi_n - \psi_n \parallel^2 < \infty,$$

then $\{\psi_n\}_1^\infty$ is also a Riesz basis itself for H. In order to use Bari's theorem, the following steps are required:

(i) to estimate "high" eigenfrequencies by asymptotic analysis technique;

(ii) to find a sequence of generalized eigenvectors $\{\psi_n\}_{N+1}^\infty$ (where N is a large integer) such that $\{\psi_n\}_{N+1}^\infty$ is quadratically close to a given Riesz basis $\{\phi_n\}_1^\infty$: $\sum_{n=N+1}^{\infty} \parallel \phi_n - \psi_n \parallel^2 < \infty$; and

(iii) to show that the number of linearly independent "low" eigenvectors is exactly N, or, more generally, as in Rao (1997) and Shubov (1996), to show that the root subspace of the system is complete in the state space.

While steps (i) and (ii) are relatively easy, step (iii) has been very difficult, in general, so far. Toward easing this difficulty, Guo (to appear) recently establishes an abstract condition under which steps (i) and (ii) automatically imply step (iii) for discrete operators in general Hilbert spaces. This greatly simplifies the verification of the Riesz basis property in applications. In this paper, we shall use this result (a simplified proof is presented in the appendix of the present paper) to show that a sequence of generalized eigenfunctions of system (2) forms a Riesz basis for the state Hilbert space for any real parameters $\alpha, \beta \neq 0$ and m. This covers the main results of Conrad and Morgül (1998) as a special case and removes the additional conditions imposed there. The exponential stability of the system is then readily established from an asymptotic expression of the eigenvalues, which is also obtained in the process of verification of the Riesz basis generation property.

The paper is organized as follows. In sections 2 and 3, some asymptotic expressions of eigenvalues and eigenfunctions are presented. Section 4 is devoted to the Riesz basis generation. Concluding remarks are given in section 5. Finally, in the appendix, we present a much simplified proof of the abstract result obtained in Guo (to appear) about the Riesz basis property of discrete operators in general Hilbert spaces.

2. Asymptotic expressions of eigenvalues and eigenfunctions. Throughout the paper, we always assume that $\beta \neq 0$. As in Conrad and Morgül (1998), the state Hilbert space for system (2) is $\mathbf{H} = H_E^2(0,1) \times L^2(0,1) \times \mathbb{C}$, where $H_E^2(0,1) = \{f \in H^2(0,1) \mid f(0) = f'(0) = 0\}$, with the inner product induced norm defined as

$$\| (f,g,\eta) \|^2 = \int_0^1 [| f''(x) |^2 + | g(x) |^2] dx + K \mid \eta \mid^2,$$

where $K > 0$ is any constant. Equation (2) can be written as an evolutionary equation in $\mathbf{H}$:

$$\frac{dY(t)}{dt} = \mathcal{A}Y(t), \tag{3}$$

where $Y(t) = (y(\cdot,t), y_t(\cdot,t), -y_{xxx}(1,t) + m\beta^{-1}y_t(1,t))$ and the operator $\mathcal{A} : D(\mathcal{A})(\subset \mathbf{H}) \to \mathbf{H}$ is defined as follows:

$$\begin{cases} \mathcal{A}(f,g,\eta) = (g, -f^{(4)}, -\beta^{-1}\eta - \beta^{-1}(\alpha - m\beta^{-1})g(1)) \quad \forall (f,g,\eta) \in D(\mathcal{A}), \\ D(\mathcal{A}) = \{(f,g,\eta) \in (H^4 \cap H_E^2) \times H_E^2 \times \mathbb{C}, f''(1) = 0, \eta = -f'''(1) + m\beta^{-1}g(1)\}. \end{cases}$$

Now, we present the following lemma on the spectrum of the operator $\mathcal{A}$.

LEMMA 2.1. *$\mathcal{A}^{-1}$ exists and is compact on $\mathbf{H}$. Hence the spectrum $\sigma(\mathcal{A})$ of $\mathcal{A}$ consists of isolated eigenvalues only: $\sigma(\mathcal{A}) = \sigma_p(\mathcal{A})$, where $\sigma_p(\mathcal{A})$ denotes the set of eigenvalues of $\mathcal{A}$. Moreover, each $\lambda = i\tau^2 \in \sigma(\mathcal{A}), \lambda \neq -\beta^{-1}$, is geometrically simple and satisfies the characteristic equation*

$$\tau(i\tau^2 + \beta^{-1})(1 + \cosh\tau\cos\tau) + (m\tau^2 - \alpha i)(\sinh\tau\cos\tau - \cosh\tau\sin\tau) = 0. \tag{4}$$

An eigenfunction (f,g,η) corresponding to $\lambda = i\tau^2 \in \sigma(\mathcal{A})(\lambda \neq -\beta^{-1})$ is given by

$$\begin{cases} f(x) & = \sinh\tau(1-x) - \sin\tau(1-x) - (\sinh\tau\cos\tau x + \sinh\tau x\cos\tau) \\ & \quad +(\cosh\tau x\sin\tau + \cosh\tau\sin\tau x), \\ g & = \lambda f, \\ \eta & = -\dfrac{\lambda\beta^{-1}}{\lambda + \beta^{-1}}(\alpha - m\beta^{-1})f(1). \end{cases} \tag{5}$$

Proof. A simple calculation shows that

$$\mathcal{A}^{-1}(f,g,\eta) = (f_1, g_1, \eta_1) \quad \forall (f,g,\eta) \in \mathbf{H},$$

$$f_1 = \int_x^1 \frac{(x-\tau)^3}{6} g(\tau)d\tau + \int_0^1 \left(\frac{\tau^3}{6} - x\frac{\tau^2}{2}\right) g(\tau)d\tau + \frac{(x-1)^3 - 3x + 1}{6}[\beta\eta + \alpha f(1)],$$

$$g_1 = f, \eta_1 = -\beta\eta - (\alpha - m\beta^{-1})f(1).$$

Since $f_1^{(4)} = -g, g_1 = f, |\eta_1| \leq |\beta||\eta| + |\alpha - m\beta^{-1}|\|f\|_{H^2}$, it follows that

$$\|\mathcal{A}^{-1}(f,g,\eta)\|_{H^4\times H^2\times\mathbb{C}} \leq M\|(f,g,\eta)\|_{\mathbf{H}}$$

for some constant $M > 0$. By the Sobolev embedding theorem, $\mathcal{A}^{-1}$ is compact on $\mathbf{H}$. This is the first part.

Second, for any $\lambda \in \sigma_p(\mathcal{A}), \lambda \neq -\beta^{-1}$, solving eigenvalue problem

$$\mathcal{A}(f,g,\eta) = (g, -f^{(4)}, -\beta^{-1}\eta - \beta^{-1}(\alpha - m\beta^{-1})g(1)) = \lambda(f,g,\eta),$$

one gets

$$g = \lambda f, \eta = -\frac{\lambda\beta^{-1}}{\lambda + \beta^{-1}}(\alpha - m\beta^{-1})f(1),$$

where f satisfies

$$\left\{\begin{array}{l} f^{(4)}(x) + \lambda^2 f(x) = 0, \\ f(0) = f'(0) = f''(1) = 0, \\ (\lambda + \beta^{-1})f'''(1) = \beta^{-1}\lambda(\alpha + m\lambda)f(1). \end{array}\right. \tag{6}$$

If (6) has two linearly independent solutions f_1, f_2, then there are constants c, d ($|c| + |d| \neq 0$) such that $cf_1(1) + df_2(1) = 0$. It follows from (6) that $f = cf_1 + df_2$ satisfies

$$\left\{\begin{array}{l} f^{(4)}(x) + \lambda^2 f(x) = 0, \\ f(0) = f'(0) = f(1) = f''(1) = f'''(1) = 0. \end{array}\right.$$

A simple calculation shows that the above equation has only zero solution. Hence $cf_1 + df_2 \equiv 0$. This contradicts the assumption that f_1, f_2 are linearly independent. Therefore, each $\lambda = i\tau^2 \in \sigma(\mathcal{A}), \lambda \neq -\beta^{-1}$, is geometrically simple.

Now, let $\lambda = i\tau^2$. By the first equation in (6) and the conditions $f(0) = f'(0) = 0$, we have

$$f(x) = c_1(\cosh\tau x - \cos\tau x) + c_2(\sinh\tau x - \sin\tau x),$$

where c_1 and c_2 are constants. Since $f''(1) = 0$, we can set

$$c_1 = \sinh\tau + \sin\tau, c_2 = -\cosh\tau - \cos\tau.$$

Obviously c_1, c_2 can not be zero simultaneously. Hence

$$\begin{aligned} f(x) &= \sinh\tau(1-x) - \sin\tau(1-x) - (\sinh\tau\cos\tau x + \sinh\tau x\cos\tau) \\ &\quad +(\cosh\tau x\sin\tau + \cosh\tau\sin\tau x), \end{aligned}$$

which satisfies

$$\left\{\begin{array}{l} f^{(4)}(x) + \lambda^2 f(x) = 0, \\ f(0) = f'(0) = f''(1) = 0 \end{array}\right. \tag{7}$$

for all $\lambda = i\tau^2$.

Finally, from the last condition $(\lambda + \beta^{-1})f'''(1) = \beta^{-1}\lambda(\alpha + m\lambda)f(1)$, one can obtain (4) (see also Conrad and Morgül, 1998). The proof is complete. □

Lemma 2.2. *There is a family of eigenvalues* $\{\lambda_n, \overline{\lambda}_n\}, \lambda_n = i\tau_n^2$ *of* $\mathcal{A}$ *satisfying*

$$\lambda = \lambda_n = i\tau_n^2 = -2m + i(k\pi)^2 + \mathcal{O}(n^{-1}), \tag{8}$$

where $k = n - 1/2$ *and* n *is a sufficiently large positive integer. An eigenfunction* (f_n, g_n, η_n) *of* $\mathcal{A}$ *corresponding to* λ_n *satisfies*

$$\begin{aligned} F_n(x) &= 2\tau_n^{-2}e^{-\tau_n}\begin{pmatrix} f_n''(x) \\ g_n(x) \\ \eta_n \end{pmatrix}^T \\ &= \begin{pmatrix} e^{-k\pi x} + (-1)^n e^{-k\pi(1-x)} + (\cos k\pi x - \sin k\pi x) \\ i[e^{-k\pi x} + (-1)^n e^{-k\pi(1-x)} - (\cos k\pi x - \sin k\pi x)] \\ 0 \end{pmatrix}^T + \mathcal{O}(n^{-1}), \end{aligned} \tag{9}$$

which holds uniformly for $x \in [0,1]$. *Consequently,*

$$\| F_n(x) \|^2_{L^2\times L^2\times\mathbb{C}} = \| 2\tau_n^{-2}e^{-\tau_n}(f_n, g_n, \eta_n) \|^2_{\mathbf{H}} \to 2 \text{ as } n \to \infty. \tag{10}$$

Proof. Let $k = n - 1/2$, with n as a sufficiently large positive integer. Noting that as $n \to \infty, \|e^{-k\pi x}\|^2_{L^2} \to 0, \|e^{-k\pi(1-x)}\|^2_{L^2} \to 0, \| \cos k\pi x - \sin k\pi x\|^2_{L^2} \to 1$, we can conclude (10) from (9) immediately. So only (9) should be verified. First, in a small neighborhood of $k\pi$, the following estimates are valid uniformly for all $n > 0$:

$$2e^{-\tau}\sinh\tau = 1 + \mathcal{O}(e^{-2|\tau|}), 2e^{-\tau}\cosh\tau = 1 + \mathcal{O}(e^{-2|\tau|}), \sin\tau = \mathcal{O}(1), \cos\tau = \mathcal{O}(1).$$

Second, multiplying $-2i\tau^{-3}e^{-\tau}$ on both sides of (4) yields

$$\cos\tau = \mathcal{O}(| \tau |^{-1}) \text{ or } \cos\tau = \frac{m}{\tau}i(\cos\tau - \sin\tau) + \mathcal{O}(| \tau |^{-2}), \tag{11}$$

which is valid uniformly in a small neighborhood of $k\pi$ for all $n > 0$. Since $\cos k\pi = 0$, we can apply Rouche's theorem to the functions $f(\tau) = \cos\tau$ and $g(\tau) = -\mathcal{O}(| \tau |^{-1})$ in a small neighborhood of $k\pi$ to find a solution to the first equation of (11) to be

$$\tau = \tau_n = k\pi + \mathcal{O}(n^{-1}). \tag{12}$$

Note that

$$\begin{cases} e^{-\tau_n y} = e^{-k\pi y} + \mathcal{O}(n^{-1}), \\ \sin\tau_n x = \sin k\pi x + \mathcal{O}(n^{-1}), \cos\tau_n x = \cos k\pi x + \mathcal{O}(n^{-1}), \end{cases} \tag{13}$$

which holds uniformly for bounded $y > 0$ and $x \in [0,1]$. Upon substituting (12) into the second equation of (11), the term $\mathcal{O}(n^{-1})$ in the expression (12) satisfies

$$-\sin k\pi\mathcal{O}(n^{-1}) = -\frac{mi}{k\pi}\sin k\pi + \mathcal{O}(n^{-2}),$$

so

$$\mathcal{O}(n^{-1}) = \frac{mi}{k\pi} + \mathcal{O}(n^{-2}).$$

This, together with (12), gives

$$\tau_n = k\pi + \frac{mi}{k\pi} + \mathcal{O}(n^{-2}).$$

Then (8) readily follows.

Now let $\tau = \tau_n$ and $(f_n, g_n, \eta_n) = (f, g, \eta)$ be defined by (5). Since

$$\begin{aligned}\tau^{-2} f_n''(x) &= \sinh \tau(1-x) + \sin \tau(1-x) + (\sinh \tau \cos \tau x - \sinh \tau x \cos \tau)\\ &\quad +(\cosh \tau x \sin \tau - \cosh \tau \sin \tau x),\end{aligned}$$

it follows from (13) that

$$\begin{aligned}2\tau^{-2}e^{-\tau} f_n''(x) &= e^{-\tau x} + \cos \tau x - e^{-\tau(1-x)} \cos \tau + e^{-\tau(1-x)} \sin \tau - \sin \tau x + \mathcal{O}(e^{-k\pi})\\ &= e^{-k\pi x} + e^{-k\pi(1-x)} \sin k\pi + \cos k\pi x - \sin k\pi x + \mathcal{O}(n^{-1})\\ &= e^{-k\pi x} + (-1)^n e^{-k\pi(1-x)} + (\cos k\pi x - \sin k\pi x) + \mathcal{O}(n^{-1}),\\ 2\tau^{-2}e^{-\tau} g_n(x) &= ie^{-\tau} f_n(x) = i[e^{-\tau x} - \cos \tau x - e^{-\tau(1-x)} \cos \tau + e^{-\tau(1-x)} \sin \tau\\ &\quad + \sin \tau x] + \mathcal{O}(e^{-k\pi})\\ &= i[e^{-k\pi x} + (-1)^n e^{-k\pi(1-x)} - (\cos k\pi x - \sin k\pi x)] + \mathcal{O}(n^{-1}),\\ 2\tau^{-2}e^{-\tau}\eta_n &= -2\tau^{-2}e^{-\tau} \tfrac{\lambda\beta^{-1}}{\lambda+\beta^{-1}}(\alpha - m\beta^{-1}) f_n(1) = \mathcal{O}(n^{-2}).\end{aligned}$$

The above estimates are valid uniformly for $x \in [0, 1]$. (9) is established. □

It should be pointed out that we are not sure at this stage that (8) is an asymptotic expression for all eigenvalues of $\mathcal{A}$. This will be cleared up after the verification of the Riesz basis generation in section 4.

3. Results of an auxiliary system. In this section, we consider an auxiliary system which is composed of a conservative system and an ordinary differential equation coupled. This system will produce a reference Riesz basis of $\mathbf{H}$, required in Theorem 6.3 in the appendix in verification of basis generation. The principle of constructing this system is based on an observation of the characteristic equation (4) that the "dominant" equation of (4) is $i\tau^3(1+\cosh \tau \cos \tau) = 0$, which can be obtained by letting $\alpha = m = \beta^{-1} = 0$. In this case, system (2) becomes

$$\begin{cases} y_{tt} + y_{xxxx} = 0,\\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = y_{xxxt}(1,t) = 0.\end{cases}$$

Naturally, we consider the well-posed conservative system

$$\begin{cases} y_{tt} + y_{xxxx} = 0,\\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = y_{xxx}(1,t) = 0,\end{cases}$$

which has the same nonzero eigenvalues as that of the system above. In order to get a state space the same as that of the system (2), i.e., $\mathbf{H}$, we complete the conservative system with another ordinary differential equation; then the auxiliary system is obtained, which is described by the following equation:

$$\begin{cases} y_{tt} + y_{xxxx} = 0,\\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = y_{xxx}(1,t) = 0,\\ \dot{\eta}(t) = 0.\end{cases}$$

Alternatively, we can describe the auxiliary system in the form of an evolutionary equation in $\mathbf{H}$,

$$\frac{dY(t)}{dt} = \mathcal{A}_0 Y(t), \tag{14}$$

where the operator $\mathcal{A}_0 : D(\mathcal{A}_0)(\subset \mathbf{H}) \to \mathbf{H}$ is defined as follows:

$$\begin{cases} \mathcal{A}_0(f,g,\eta) = (g, -f^{(4)}, 0) \quad \forall (f,g,\eta) \in D(\mathcal{A}_0), \\ D(\mathcal{A}_0) = \{(f,g,\eta) \in (H^4 \cap H_E^2) \times H_E^2 \times \mathbb{C}, f''(1) = f'''(1) = 0\}. \end{cases}$$

It is easy to show that $1 \in \rho(\mathcal{A}_0), (I - \mathcal{A}_0)^{-1}$ is compact on $\mathcal{H}$ and $\mathcal{A}_0^* = -\mathcal{A}_0$. That is, $\mathcal{A}_0$ is a skew-adjoint operator with compact resolvent on $\mathbf{H}$ (and hence $i\mathcal{A}_0$ is self-adjoint with compact resolvent). It then follows from a well-known result in functional analysis that (i) there is a sequence of normalized eigenfunctions of $\mathcal{A}_0$ which forms an orthonormal basis of $\mathbf{H}$; (ii) for each eigenvalue of $\mathcal{A}_0$, its geometric multiplicity equals its algebraic multiplicity; and (iii) all eigenvalues of $\mathcal{A}_0$ lie on the imaginary axis. (ii) is actually a consequence of (i). (iii) comes directly from the skew-adjointness of $\mathcal{A}_0$. These are advantages of the construction of $\mathcal{A}_0$.

All the analysis of the operator $\mathcal{A}$ in the preceding section is true for the operator $\mathcal{A}_0$. In particular, each $\mu \in \sigma(\mathcal{A}_0), \mu \neq 0$, is geometrically simple and hence algebraically simple. And the characteristic equation for $\mu = i\omega^2 (\neq 0) \in \sigma(\mathcal{A}_0)$ is

(15) $$1 + \cosh\omega \cos\omega = 0.$$

Since all eigenvalues of $\mathcal{A}_0$ lie on the imaginary axis, we need consider only the positive solutions to (15) in order to find all nonzero eigenvalues of $\mathcal{A}_0$.

For $\omega > 0$, writing (15) as $\cos\omega = \mathcal{O}(e^{-\omega})$, we can get the positive solutions of (15) being

(16) $$\omega = \omega_n = k\pi + \mathcal{O}(e^{-k\pi}),$$

where $k = n - 1/2$ for all sufficiently large positive integers n.

Therefore, the spectrum of $\mathcal{A}_0$ consists of all pairs $\{\mu_n, \overline{\mu}_n\}$ together with possibly another finite set, where $\mu_n = i\omega_n^2$ with ω_n given in (16). This is unlike $\mathcal{A}$; $\mu_n = i\omega_n^2 = i(k\pi)^2 + \mathcal{O}(k\pi e^{-k\pi})$ is now indeed an asymptotic expression for all eigenvalues of $\mathcal{A}_0$.

Now, letting $\alpha = m = \beta^{-1} = 0$ and $\tau_n = \omega_n$, in Lemma 2.2, we get an eigenvector (u_n, v_n, ν_n) of $\mathcal{A}_0$ corresponding to $\mu_n = i\omega_n^2 (\neq 0)$ given below:

(17) $$\begin{cases} u_n(x) & = \sinh\omega_n(1-x) - \sin\omega_n(1-x) - (\sinh\omega_n \cos\omega_n x + \sinh\omega_n x \cos\omega_n) \\ & \quad +(\cosh\omega_n x \sin\omega_n + \cosh\omega_n \sin\omega_n x), \\ v_n & = \mu_n u_n, \\ \nu_n & = 0. \end{cases}$$

Clearly, the asymptotic expression (12) is also valid for ω_n defined in (16). Noting that only the expression (12) is used in the proof of Lemma 2.2, we have the following counterpart of Lemma 2.2 for $\mathcal{A}_0$. Similar results were also obtained in Lancaster and Shkalikov (1994).

LEMMA 3.1. *The spectrum of $\mathcal{A}_0$ consists of all $\{\mu_n, \overline{\mu}_n\}$ but possibly a finite number of the other eigenvalues, where $\mu_n = i\omega_n^2, \omega_n$ is determined by* (16). *And the eigenvalues μ_n ($\overline{\mu}_n$) are algebraically simple for all large n. In addition, an eigenfunction (u_n, v_n, ν_n) of $\mathcal{A}_0$ corresponding to μ_n satisfies*

(18) $$\begin{aligned} G_n(x) & = 2\omega_n^{-2} e^{-\omega_n} \begin{pmatrix} u_n''(x) \\ v_n(x) \\ \nu_n \end{pmatrix}^T \\ & = \begin{pmatrix} e^{-k\pi x} + (-1)^n e^{-k\pi(1-x)} + (\cos k\pi x - \sin k\pi x) \\ i[e^{-k\pi x} + (-1)^n e^{-k\pi(1-x)} - (\cos k\pi x - \sin k\pi x)] \\ 0 \end{pmatrix}^T + \mathcal{O}(n^{-1}), \end{aligned}$$

which holds uniformly for all $x \in [0,1]$.

4. Riesz basis generation. In this section, we shall apply Theorem 6.3 in the appendix to get the basis property of $\mathcal{A}$. To do this, we need a reference Riesz basis of $\mathbf{H}$ first. This is accomplished by collecting the eigenfunctions of $\mathcal{A}_0$. As we can conclude from Lemma 3.1 that a "maximal" set (see appendix) of ω-linearly independent eigenfunctions of $\mathcal{A}_0$ consists of all (u_n, v_n, ν_n) defined by (17) but a finite number of the other eigenfunctions, we may assume, without loss of generality, that such a set is

$$\{2\omega_n^{-2}e^{-\omega_n}(u_n, v_n, \nu_n)\}_1^\infty \cup \{\text{their conjugates}\}.$$

Since $\mathcal{A}_0$ is skew-adjoint, the set $\{2\omega_n^{-2}e^{-\omega_n}(u_n, v_n, \nu_n)\}_1^\infty \cup \{\text{their conjugates}\}$ forms an orthogonal basis of $\mathbf{H}$. Because they are approximately normalized (that is, they are upper and lower bounded) according to (10), the set is indeed a Riesz basis of $\mathbf{H}$ by a well-known fact that all approximately normalized Riesz bases in a separate Hilbert space are equivalent.

From (9) and (18), it follows that there is a large positive integer N such that

$$\begin{aligned}
&\sum_{n>N}^{\infty} \parallel 2\tau_n^{-2}e^{-\tau_n}(f_n, g_n, \eta_n) - 2\omega_n^{-2}e^{-\omega_n}(u_n, v_n, \nu_n) \parallel_{\mathbf{H}}^2 \\
&= \sum_{n>N}^{\infty} \parallel F_n - G_n \parallel_{L^2\times L^2\times\mathbb{C}}^2 = \sum_{n>N}^{\infty} \mathcal{O}(n^{-2}) < \infty.
\end{aligned} \tag{19}$$

The same is true for their conjugates. Note that all $\lambda_n = i\tau_n^2$ are different for large n; we can now apply Theorem 6.3 in the appendix to obtain the main results of the present paper.

THEOREM 4.1. *Let the operator $\mathcal{A}$ be defined as in* (3). *Then*

(i) *there is a sequence of generalized eigenfunctions of $\mathcal{A}$ which forms a Riesz basis for the state space* $\mathbf{H}$*;*

(ii) *all of the eigenvalues of $\mathcal{A}$ have the asymptotic expression* (8)*; and*

(iii) *all $\lambda \in \sigma(\mathcal{A})$ with sufficiently large modulus are algebraically simple.*

Therefore, $\mathcal{A}$ generates a C_0-group on $\mathbf{H}$ for any real constants m, α, and β. Moreover, for the semigroup $e^{\mathcal{A}t}$ generated by $\mathcal{A}$, the spectrum-determined growth condition holds. And the growth rate of $e^{\mathcal{A}t}$ is not less than $-2m$.

The stability result for the system (2) is given in the following corollary.

COROLLARY 4.2. *The semigroup $e^{\mathcal{A}t}$ is exponentially stable for any $m, \alpha, \beta > 0$.*

Proof. Taking the inner product of $\mathbf{H}$ as in the beginning of the section 2 with $K = \beta^2/(m + \alpha\beta)$, it is calculated in Conrad and Morgül (1998) that

$$\mathrm{Re}\langle \mathcal{A}Y, Y\rangle = -\frac{K}{\beta} \mid f'''(1) \mid^2 - \frac{Km\alpha}{\beta^2} \mid g(1) \mid^2 \leq 0 \quad \forall\ Y = (f, g, \eta) \in D(\mathcal{A}).$$

That is, $\mathcal{A}$ is dissipative and hence no eigenvalues of $\mathcal{A}$ lie on the open right half complex plane. Now, if $\mathcal{A}Y = \lambda Y$, $Y = (f, g, \eta)$, and $\mathrm{Re}\lambda = 0$, then $f'''(1) = g(1) = 0$. It follows from (6) that

$$\begin{cases} f^{(4)}(x) + \lambda^2 f(x) = 0, \\ f(0) = f'(0) = f''(1) = 0, \\ f'''(1) = f(1) = 0. \end{cases}$$

As it is indicated in the proof of Lemma 2.1, the above equation has a zero solution only. Hence $f \equiv 0$ and so $g = \eta = 0$ by (5). Therefore,

$$\mathrm{Re}\lambda < 0 \quad \forall \lambda \in \sigma(\mathcal{A}). \tag{20}$$

Finally, since $\mathcal{A}$ is of compact resolvent, there are only finitely many eigenvalues of $\mathcal{A}$ in any bounded region of the complex plane, which, together with Theorem 4.1 (ii), shows that there is a constant $\omega > 0$ such that

$$S(\mathcal{A}) = \sup_{\lambda \in \sigma(\mathcal{A})} \mathrm{Re}\lambda < -\omega. \tag{21}$$

The exponential stability then follows from the spectrum-determined growth condition. The proof is complete. □

Before ending the section, we indicate that Theorem 4.1 can be used to obtain the basis property and the spectrum-determined growth condition of a beam equation without tip mass under linear boundary feedback control. Let $\mathcal{A}_1$ be defined by setting $m = \theta\beta$, where θ is real, in the definition of the operator $\mathcal{A}$; that is,

$$\begin{cases} \mathcal{A}_1(f, g, \eta) = (g, -f^{(4)}, -\beta^{-1}\eta - \beta^{-1}(\alpha - \beta^{-1})g(1)) \quad \forall (f, g, \eta) \in D(\mathcal{A}_1), \\ D(\mathcal{A}_1) = \{(f, g, \eta) \in (H^4 \cap H_E^2) \times H_E^2 \times \mathbb{C}, f''(1) = 0, \eta = -f'''(1) + \theta g(1)\}. \end{cases}$$

Then Theorem 4.1 holds true for operator $\mathcal{A}_1$ for any reals β^{-1}, α, and θ. Furthermore, let $\mathcal{A}_2$ be defined by setting $\alpha = \beta^{-1} = 0$ in the definition of $\mathcal{A}_1$. We have

$$\begin{cases} \mathcal{A}_2(f, g, \eta) = (g, -f^{(4)}, 0) \quad \forall (f, g, \eta) \in D(\mathcal{A}_2), \\ D(\mathcal{A}_2) = \{(f, g, \eta) \in (H^4 \cap H_E^2) \times H_E^2 \times \mathbb{C}, f''(1) = 0, \eta = -f'''(1) + \theta g(1)\}. \end{cases}$$

And Theorem 4.1 is also true for operator $\mathcal{A}_2$ for any real θ. However,

$$\frac{dY}{dt} = \mathcal{A}_2 Y(t)$$

is equivalent to

$$\begin{cases} y_{tt}(x,t) + y_{xxxx}(x,t) = 0, \quad 0 < x < 1, \quad t > 0, \\ y(0,t) = y_x(0,t) = y_{xx}(1,t) = 0, \quad t \geq 0, \\ y_{xxx}(1,t) = \theta y_t(1,t), \quad t \geq 0. \end{cases} \tag{22}$$

That is, system (22) is also a Riesz spectral system. This system is just the cantilevered beam equation considered in Conrad (1990), Conrad and Morgül (1998), and Guo (to appear).

Remark 4.3. The conclusion of Corollary 4.2 was proved in Conrad and Morgül (1998) by energy multiplier method. Theorem 4.1 (i) was shown there in the case of $m = \alpha\beta$ for almost every $\alpha > 0$ with other additional conditions. Theorem 4.1 (iii) was also shown there by complex analysis. The Riesz basis property for (22) was obtained there for almost every $\theta > 0$.

5. Concluding remarks. In this paper, an abstract condition for Riesz basis generation of discrete operators in Hilbert spaces is used to show that a sequence of generalized eigenfunctions of an Euler–Bernoulli beam equation with a tip mass under boundary linear feedback control forms a Riesz basis for the state Hilbert space. The stability of the system is also established. This paper greatly improves the work of Conrad and Morgül (1998), where the same results are obtained but for a very special case where $m = \alpha\beta$. Besides these results, the contributions of this paper lie in providing a very simple method which enables us (a) to obtain the asymptotic expressions of eigenvalues and eigenfunctions; (b) to avoid the usual treatment for "low" eigenfrequencies in applying Bari's theorem; and (c) to study potential applications to other problems of beam equations (see Guo and Chan, 2001).

6. Appendix. Abstract result on Riesz basis property. In this appendix, we present the abstract result together with a simplified proof about Riesz basis generation for discrete operators in the Hilbert spaces. This result is crucial to the establishment of the main results of the present paper.

Let us recall that for a closed linear operator A in a Hilbert space H, a nonzero $x \in H$ is called a generalized eigenvector of A, corresponding to an eigenvalue λ (with finite algebraic multiplicity) of A, if there is a positive integer n such that $(\lambda - A)^n x = 0$. Let $\overline{sp}(A)$, the so-called root subspace of A, be the closed subspace spanned by all generalized eigenvectors of A. The following theorem gives a simple characterization of the completeness of $\overline{sp}(A)$; that is, $\overline{sp}(A) = H$.

LEMMA 6.1. *Let A be a densely defined discrete operator (that is, there is a $\lambda \in \rho(A)$ such that $R(\lambda, A) = (\lambda - A)^{-1}$ is compact) in a Hilbert space H. Then $\overline{sp}(A) = H$ if and only if the codimension of $\overline{sp}(A)$ in H is finite.*

Proof. It is well known that the adjoint operator A^* of a densely defined discrete operator A is also a discrete operator. It follows from Lemma 5 on p. 2355 of Dunford and Schwartz (1971) that the following orthogonal decomposition holds:

$$H = \sigma_\infty(A^*) \oplus \overline{sp}(A),$$

where $\sigma_\infty(A^*) = \{x | E(\lambda)x = 0,\ \forall\lambda \in \sigma(A^*)\}, E(\lambda)$ is the eigen-projector of A^* corresponding to λ. Hence $\overline{sp}(A) = H$ if and only if $\sigma_\infty(A^*) = \{0\}$. On the other hand, Lemma 5 on p. 2295 of Dunford and Schwartz (1971) suggests that $\sigma_\infty(A^*)$ is either $\{0\}$ or infinite dimensional. Therefore the codimension of $\overline{sp}(A)$ is finite if and only if $\sigma_\infty(A^*) = \{0\}$. The proof is complete. □

LEMMA 6.2. *Let $\{\phi_n\}_1^\infty$ be a Riesz basis in a Hilbert space H. Let $\{\psi_n\}_{N+1}^\infty$ ($N \geq 0$) be another sequence in H. If*

$$\sum_{n=N+1}^{\infty} \| \phi_n - \psi_n \|^2 < \infty,$$

then there exists an $M \geq N$ such that $\{\phi_n\}_1^M \cup \{\psi_n\}_{M+1}^\infty$ is a Riesz basis of H. In particular, $\{\psi_n\}_{M+1}^\infty$ is ω-linearly independent.

Proof. The proof can follow from Corollary 11.4 on page 374 of Singer (1970).

THEOREM 6.3. *Let A be a densely defined discrete operator in a Hilbert space H. Let $\{\phi_n\}_1^\infty$ be a Riesz basis of H. If there are an integer $N \geq 0$ and a sequence of generalized eigenvectors $\{\psi_n\}_{N+1}^\infty$ of A such that*

$$\sum_{N+1}^{\infty} \|\phi_n - \psi_n\|^2 < \infty,$$

then the following hold.

(i) *There are a constant $M > N$ and generalized eigenvectors $\{\psi_{n0}\}_1^M$ of $\mathcal{A}$ such that $\{\psi_{n0}\}_1^M \cup \{\psi_n\}_{M+1}^\infty$ forms a Riesz basis of H.*

(ii) *Let $\{\psi_{n0}\}_1^M \cup \{\psi_n\}_{M+1}^\infty$ correspond to eigenvalues $\{\sigma_n\}_1^\infty$ of A. Then $\sigma(A) = \{\sigma_n\}_1^\infty$, where σ_n is counted according to its algebraic multiplicity.*

(iii) *If there is an* $M_0 > 0$ *such that* $\sigma_n \neq \sigma_m$ *for all* $m, n > M_0$, *then there is an* $N_0 > M_0$ *such that all* σ_n *are algebraically simple if* $n > N_0$. □

Proof. (ii) and (iii) are consequences of (i). Only proof of (i) is needed. By Lemma 6.2, there is an $M > N$ such that $\{\psi_n\}_{M+1}^{\infty}$ is ω-linearly independent. Let $\{\psi_\alpha\}$ be an arbitrary set such that $\{\psi_n\}_{M+1}^{\infty} \cup \{\psi_\alpha\}$ is a "maximal" ω-linearly independent set of generalized eigenvectors of A; that is, $\{\psi_n\}_{M+1}^{\infty} \cup \{\psi_\alpha\}$ is a ω-linearly independent subset of the set of the generalized eigenvectors of A, and for any generalized eigenvector ψ of A, the extended set $\{\psi_n\}_{M+1}^{\infty} \cup \{\psi_\alpha\} \cup \{\psi\}$ must not be ω-linearly independent anymore. Therefore, $\{\psi_n\}_{M+1}^{\infty} \cup \{\psi_\alpha\}$ spans the root subspace $\overline{sp}(A)$. By the assumption and Bari's theorem, the number of such ψ_α's does not exceed M. Let $\{\psi_\alpha\} = \{\psi_{n0}\}_1^L, L \leq M$. It follows from Theorem 3.2 of Rao (1997) that $\{\psi_{n0}\}_1^L \cup \{\psi_n\}_{M+1}^{\infty}$ forms a Riesz basis of $\overline{sp}(A)$.

On the other hand, by the assumption and Bari's theorem, the number of linearly independent elements in the orthogonal complement of $\overline{sp}(A)$ in H cannot exceed M, and hence the codimension of $\overline{sp}(A)$ is finite. Then from Lemma 6.1, $\overline{sp}(A) = H$.

Therefore, $\{\psi_{n0}\}_1^L \cup \{\psi_n\}_{M+1}^{\infty}$ forms a Riesz basis for the entire space H.

Since a "proper" subset of a basis can not be a basis, it follows from Bari's theorem and the assumption that $L = M$. This is (i). The proof is complete. □

Acknowledgments. The author would like to thank the anonymous referees for their valuable comments and suggestions.

REFERENCES

G. Chen, M. C. Delfour, A. M. Krall, and G. Payre (1987), *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25, pp. 526–546.

G. Chen, S. G. Krantz, D. L. Russell, C. E. Wayne, H. H. West, and M. P. Coleman (1989), *Analysis, designs, and behavior of dissipative joints for coupled beams*, SIAM J. Appl. Math., 49, pp. 1665–1693.

F. Conrad (1990), *Stabilization of beams by pointwise feedback control*, SIAM J. Control Optim., 28, pp. 423–437.

F. Conrad and O. Mörgül (1998), *On the stabilization of a flexible beam with a tip mass*, SIAM J. Control Optim., 36, pp. 1962–1986.

S. Cox and E. Zuazua (1994), *The rate at which energy decays in a damped string*, Comm. Partial Differential Equations, 19, pp. 213–243.

N. Dunford and J. T. Schwartz (1971), *Linear Operators, Part* III, Wiley-Interscience, New York.

B. Z. Guo, *The Riesz basis property of discrete operators and application to an Euler-Bernoulli beam equation with boundary linear feedback control*, IMA J. Math. Control Inform., to appear.

B. Z. Guo and K. Y. Chan, *Riesz basis generation, eigenvalues distribution, and exponential stability for an Euler-Bernoulli beam with joint feedback control*, Rev. Mat. Complut., 14 (2001), pp. 1–24.

I. C. Gohberg and M. G. Krein (1969), *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI.

P. Lancaster and A. Shkalikov (1994), *Damped vibrations of beams and related spectral problems*, Canad. Appl. Math. Quart., 2, pp. 45–90.

S. Li, J. Yu, Z. Liang, and G. Zhu (1999), *Stabilization of high eigenfrequencies of a beam equation with generalized viscous damping*, SIAM J. Control Optim., 37, pp. 1767–1779.

O. Morgül, F. Conrad, and B. P. Rao (1994), *On the stabilization of a cable with a tip mass*, IEEE Trans. Automat. Control, 39, pp. 2140–2145.

B. P. Rao (1997), *Optimal energy decay rate in a damped Rayleigh beam*, in Optimization Methods in Partial Differential Equations (South Hadley, MA, 1996), Contemp. Math. 209, S. Cox and I. Lasiecka, eds., AMS, Providence, RI, pp. 221–229.

R. Rebarber (1989), *Spectral assignability for distributed parameter systems with unbounded scalar control*, SIAM J. Control Optim., 27, pp. 148–169.

R. Rebarber (1995), *Exponential stability of coupled beams with dissipative joints: A frequency domain approach*, SIAM J. Control Optim., 33, pp. 1–28.

M. A. Shubov (1996), *Basis property of eigenfunctions of nonselfadjoint operator pencils generated by the equation of nonhomogenerous damped string*, Integral Equations Operator Theory, 25, pp. 289–328.

M. A. Shubov (1997), *Spectral operators generated by damped hyperbolic equations*, Integral Equations Operator Theory, 28, pp. 358–372.

I. Singer (1970), *Bases in Banach Spaces* I, Springer-Verlag, New York, Berlin.

S. H. Sun (1981), *On spectrum distribution of completely controllable linear systems*, SIAM J. Control Optim., 19, pp. 730–743 (translated by L. F. Ho).

C.-Z. Xu and G. Sallet (1996), *On spectrum and Riesz basis assignment of infinite-dimensional linear systems by bounded linear feedbacks*, SIAM J. Control Optim., 34, pp. 521–541.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1748–1755

ACHIEVING ARBITRARILY LARGE DECAY IN THE DAMPED WAVE EQUATION*

CARLOS CASTRO[†] AND STEVEN J. COX[‡]

Abstract. We exhibit a sequence of viscous dampings for the fixed string that yields arbitrarily fast attenuation of any and all initial disturbances. The limit case produces extinction of all solutions in finite time.

Key words. damped wave equation, decay rate

AMS subject classifications. 35P15, 35L05

PII. S0363012900370971

1. Introduction.

The displacement u of a string of unit length, fixed at its ends, and in the presence of viscous damping $2a$, satisfies

$$\begin{aligned} &u_{tt}(x,t) - u_{xx}(x,t) + 2a(x)u_t(x,t) = 0, \quad 0 < x < 1, \quad 0 < t, \\ &u(0,t) = u(1,t) = 0, \quad 0 < t, \end{aligned} \tag{1.1}$$

upon being set in motion by the initial disturbance

$$u(x,0) = u_0(x), \quad u_t(x,0) = v_0(x). \tag{1.2}$$

If $a \in L^\infty(0,1)$ is nonnegative and strictly positive on some subinterval, then the energy

$$E(t) \equiv \int_0^1 u_x^2(x,t) + u_t^2(x,t)\,dx$$

is known to obey $E(t) \le CE(0)e^{2\omega t}$ for some finite $C > 0$ and $\omega < 0$, independent of the initial disturbance. The smallest such ω,

$$\omega(a) \equiv \inf\{\quad \begin{aligned} &\omega : \exists C(\omega) > 0 \text{ such that (s.t.) } E(t) \le CE(0)e^{2\omega t} \\ &\text{for every finite energy solution of (1.1)}\}, \end{aligned}$$

is referred to as the *decay rate* associated with a. If a is to be introduced in order to absorb an initial disturbance, then one naturally wishes to strike upon that a with the least possible (most negative) decay rate. The mathematical attraction here lies in the often noted fact that, with respect to damping, "more is not better." More precisely, for constant a, the decay rate is not a decreasing function of a. Rather, for small a, ω decreases until a reaches π, after which ω rapidly increases to 0. Our aim is to show that there exist nonconstant a that circumvent this phenomena of overdamping and hence that more indeed can be better.

*Received by the editors April 13, 2000; accepted for publication (in revised form) August 28, 2000; published electronically April 12, 2001.

http://www.siam.org/journals/sicon/39-6/37097.html

[†]Departamento de Matemática e Informática Aplicadas a la Ingeniría Civil, ETSI Caminos Canales y Puertos, Universidad Politécnica de Madrid, 28040 Madrid, Spain (ccastro@dumbo.caminos.upm.es). The research of this author was partially supported by project PB96-0663 of the DGES (Spain).

[‡]Department of Computational and Applied Mathematics, Rice University, P.O. Box 1892, Houston, TX 77005 (cox@caam.rice.edu).

Cox and Zuazua [3, Thm. 6.5] have shown that $a \mapsto \omega(a)$ attains its finite minimum over $\{a \in BV(0,1) : T(a) \leq M\}$, where $T(a)$ denotes the total variation of a. We show here that the total variation constraint is not superfluous. More precisely, we establish the following theorem.

THEOREM 1.1. *If $a_n(x) = 1/(x + 1/n)$, then $\omega(a_n) \to -\infty$ and $T(a_n) \to \infty$ as $n \to \infty$.*

This result is a dramatic improvement over previous attempts to minimize $a \mapsto \omega(a)$. In particular, Cox and Overton [2], based on the study of ω near $a = \pi$, conjectured that $\omega(\pi)$ may indeed be the minimum. Later on, Freitas [4] suggested a negative answer to the conjecture by numerical evidence based on the clever choices $a(x) = \pi(1 + \cos(2\pi x))/2$.

When $n \to \infty$, $a_n(x) \to 1/x \notin BV(0,1)$. In this case the solutions are extinct in finite time.

THEOREM 1.2. *If $a(x) = 1/x$, for any initial data $(u_0, u_1) \in H_0^1(0,1) \times L^2(0,1)$ the solution $u(x,t)$ of (1.1)–(1.2) satisfies*

$$u(x,t) = u_t(x,t) = 0 \qquad \forall t > 2. \tag{1.3}$$

The remainder of the paper is organized as follows. In section 2 we equate the decay rate with the spectral abscissa of the associated damped wave operator and express its shooting function in the case that $a(x) = 1/(x + c)$. In section 3, via simple calculus, we show that all zeros of the shooting function travel arbitrarily far to the left as c is made small. Finally, in section 4 we prove the result for the limit case $a(x) = 1/x$.

2. The shooting function.

For a of finite total variation Benaddi and Rao [1] have shown that $\omega(a)$ coincides with the *spectral abscissa*

$$\mu(a) = \sup\{\Re\lambda : \lambda \in \sigma(a)\},$$

where $\sigma(a)$ denotes the spectrum of the associated damped wave operator

$$\mathcal{A}(a) = \begin{pmatrix} 0 & I \\ d^2/dx^2 & -2a \end{pmatrix}, \qquad D(\mathcal{A}(a)) = (H^2(0,1) \cap H_0^1(0,1)) \times H_0^1(0,1).$$

If $V = [y, z]$ is an eigenvector of $\mathcal{A}(a)$ with eigenvalue λ, then $z = \lambda y$ and $y'' - 2az = \lambda z$, or

$$y'' - \lambda^2 y - 2a\lambda y = 0, \tag{2.1}$$

subject to

$$y(0) = y(1) = 0. \tag{2.2}$$

We adopt the shooting method for the study of (2.1)–(2.2). That is, we denote by $x \mapsto y_2(x, \lambda)$ the function that satisfies (2.1) and the initial conditions

$$y_2(0,\lambda) = 0 \quad \text{and} \quad y_2'(0,\lambda) = 1.$$

The eigenvalues are then simply the roots of $\lambda \mapsto y_2(1,\lambda)$. This gives the exact spectrum when a is constant (in which case y_2 may be obtained in closed form) and asymptotic results in general. In establishing such asymptotic results, Cox and Zuazua

[3, equation 5.3] observed that y_2 could also be simply expressed when the damping was of the particular form

$$a(x) = \frac{1}{x+c} \tag{2.3}$$

for any $c > 0$. In this case,

$$y_2(x,\lambda) = c(x+c)e^{\lambda x} \int_0^x \frac{e^{-2\lambda s}}{(s+c)^2}\, ds.$$

Hence the eigenvalues of $\mathcal{A}(a)$ are the zeros of the *shooting function* $\lambda \mapsto F(\lambda, c)$, where

$$F(\lambda, c) \equiv \int_0^1 \frac{e^{-2\lambda x}}{(x+c)^2}\, dx.$$

One first notices that $F(\cdot, c)$ has no real zeros, and hence the choice (2.3) is not thwarted by overdamping, even for small c. Second, we note that as c becomes small the integrand in F becomes large unless the real part of λ itself becomes quite (negatively) large. We devote the next section to a precise statement of the latter observation and the ensuing proof of Theorem 1.1.

3. Calculus lemmas.

We split F into its real and imaginary parts

$$F(\lambda, c) = \int_0^1 \frac{e^{\alpha x}\cos(\beta x)}{(x+c)^2}\, dx + i\int_0^1 \frac{e^{\alpha x}\sin(\beta x)}{(x+c)^2}\, dx,$$

where $\alpha = -2\Re\lambda$ and $\beta = -2\Im\lambda$, and we attack them separately.

LEMMA 3.1. *If $f \in L^\infty(0,1)$ is nonnegative and decreasing, then*

$$\int_0^1 f(x)\sin(\beta x)\, dx > 0 \qquad \forall\, \beta > 0. \tag{3.1}$$

Proof. For $\beta \le 2\pi$ the result is obvious. For larger β, let K be the greatest integer for which $2\pi K < \beta$, and write

$$\int_0^1 f(x)\sin(\beta x)\, dx = \int_0^{2\pi K/\beta} f(x)\sin(\beta x)\, dx + \int_{2\pi K/\beta}^1 f(x)\sin(\beta x)\, dx \equiv I_1 + I_2.$$

The first integral

$$\begin{aligned} I_1 &= \sum_{k=0}^{K-1} \int_{2\pi k/\beta}^{2\pi(k+1)/\beta} f(x)\sin(\beta x)\, dx \\ &= \sum_{k=0}^{K-1} \left[\int_{2\pi k/\beta}^{\pi(2k+1)/\beta} f(x)\sin(\beta x)\, dx + \int_{\pi(2k+1)/\beta}^{2\pi(k+1)/\beta} f(x)\sin(\beta x)\, dx \right] \\ &= \sum_{k=0}^{K-1} \int_{2\pi k/\beta}^{\pi(2k+1)/\beta} (f(x) - f(x+\pi/\beta))\sin(\beta x)\, dx > 0 \end{aligned}$$

because each integrand is positive.

Concerning I_2, if $(1-2\pi K/\beta) < \pi/\beta$, then the integrand is positive and so $I_2 > 0$. On the other hand, if $(1-2\pi K/\beta) \geq \pi/\beta$, we find

$$\begin{aligned} I_2 &= \int_{2\pi K/\beta}^{\pi(2K+1)/\beta} f(x)\sin(\beta x)\,dx + \int_{\pi(2K+1)/\beta}^{1} f(x)\sin(\beta x)\,dx \\ &= \int_{2\pi K/\beta}^{\pi(2K+1)/\beta} f(x)\sin(\beta x)\,dx - \int_{2\pi K/\beta}^{1-\pi/\beta} f(x+\pi/\beta)\sin(\beta x)\,dx \\ &= \int_{2\pi K/\beta}^{1-\pi/\beta} \{f(x)-f(x+\pi/\beta)\}\sin(\beta x)\,dx + \int_{1-\pi/\beta}^{\pi(2K+1)/\beta} f(x)\sin(\beta x)\,dx > 0. \qquad \square \end{aligned}$$

LEMMA 3.2. *If $\beta > 0$ and $g \in L^\infty(0,1)$ is nonnegative and strictly convex, then*

$$\int_0^{\pi(2J+1)/\beta} g(x)\sin(\beta x)\,dx > 0 \qquad \forall\, J \in \mathbb{N},\ s.t.\ \pi(2J+1) \leq \beta. \tag{3.2}$$

Proof. As g is strictly convex, either (i) g is decreasing, (ii) g is increasing, or (iii) there exists $x_0 \in (0,1)$ such that g is decreasing on $[0,x_0)$ and increasing on $(x_0,1]$.

In case (i) the result follows from Lemma 3.1. In case (ii)

$$\begin{aligned} \int_0^{\pi(2J+1)/\beta} g(x)\sin(\beta x)\,dx &= \int_0^{\pi(2J+1)/\beta} g(\pi(2J+1)/\beta - x)\sin(\pi(2J+1)-\beta x)\,dx \\ &= \int_0^{\pi(2J+1)/\beta} g(\pi(2J+1)/\beta - x)\sin(\beta x)\,dx, \end{aligned}$$

and we can apply again Lemma 3.1, as $f(x) = g(\pi(2J+1)/\beta - x)$ is decreasing. Regarding the third case, we write

$$\begin{aligned} \int_0^{\pi(2J+1)/\beta} g(x)\sin(\beta x)\,dx &= \int_0^{x_0} g(x)\sin(\beta x)\,dx + \int_{x_0}^{\pi(2J+1)/\beta} g(x)\sin(\beta x)\,dx \\ &= \int_0^{x_0} g(x)\sin(\beta x)\,dx + \int_0^{\pi(2J+1)/\beta - x_0} g(\pi(2J+1)/\beta - x)\sin(\beta x)\,dx, \end{aligned}$$

and we note that as g and $x \mapsto g(\pi(2J+1)/\beta - x)$ are decreasing on $(0,x_0)$ and $(0,\pi(2J+1)/\beta - x_0)$, respectively, we may apply the previous lemma to each and conclude positivity of the whole. $\square$

LEMMA 3.3. *Given $A > 0$, there exist $B_1 = B_1(A) > 0$ and $C_1 = C_1(A) > 0$ such that if $0 < \alpha \leq A$, $|\beta| \geq B_1$, and $c \leq C_1$, then*

$$\int_0^1 \frac{e^{\alpha x}\sin(\beta x)}{(x+c)^2}\,dx \neq 0. \tag{3.3}$$

Proof. We assume, without loss of generality, that $\beta > 0$. We decompose

$$\int_0^1 \frac{e^{\alpha x}\sin(\beta x)}{(x+c)^2}\,dx = \int_0^1 \frac{\sin(\beta x)}{(x+c)^2}\,dx + \int_0^1 \frac{(e^{\alpha x}-1)\sin(\beta x)}{(x+c)^2}\,dx \equiv I_1 + I_2$$

and estimate I_2 via

$$I_2 = \int_0^{\pi(2J+1)/\beta} \frac{(e^{\alpha x}-1)\sin(\beta x)}{(x+c)^2}\,dx + \int_{\pi(2J+1)/\beta}^1 \frac{(e^{\alpha x}-1)\sin(\beta x)}{(x+c)^2}\,dx, \tag{3.4}$$

where J is the greatest integer such that $\pi(2J+1) \le \beta$. The first integral in (3.4) is positive, via Lemma 3.2, thanks to the convexity of $x \mapsto (e^{\alpha x}-1)(x+c)^{-2}$ on $[0,1]$ when α and c are positive. Concerning the second integral in (3.4) as

$$\left|\int_{\pi(2J+1)/\beta}^{1} \frac{(e^{\alpha x}-1)\sin(\beta x)}{(x+c)^2}\,dx\right| \le \frac{(1-\pi(2J+1)/\beta)(1+e^{\alpha})}{(\pi(2J+1)/\beta+c)^2} \le \frac{2\pi}{\beta}\frac{(1+e^{\alpha})}{(c+1-2\pi/\beta)^2},$$

it follows that

$$I_2 \ge -\frac{2\pi(1+e^{\alpha})}{\beta(c+1-2\pi/\beta)^2}. \tag{3.5}$$

We now estimate I_1. Integrating by parts, we easily obtain

$$\beta\int_0^1 \frac{\sin(\beta x)}{(x+c)^2}\,dx = \frac{1}{c^2} - \frac{\cos(\beta)}{(1+c)^2} - 2\int_0^1 \frac{\cos(\beta x)}{(x+c)^3}\,dx. \tag{3.6}$$

The third term in this expression can be bounded by

$$-2\int_0^1 \frac{\cos(\beta x)}{(x+c)^3}\,dx > -2\int_0^{\pi/2\beta} \frac{\cos(\beta x)}{(x+c)^3}\,dx \tag{3.7}$$

for

$$-2\int_{\pi/2\beta}^1 \frac{\cos(\beta x)}{(x+c)^3}\,dx = 2\int_0^{1-\pi/2\beta} \frac{\sin(\beta x)}{(x+\pi/2\beta+c)^3}\,dx > 0,$$

thanks to Lemma 3.1 and the decreasing nature of $x \mapsto (x+\pi/2\beta+c)^{-3}$. Returning to (3.7), we find

$$-2\int_0^1 \frac{\cos(\beta x)}{(x+c)^3}\,dx > -2\int_0^{\pi/2\beta} \frac{1}{(x+c)^3}\,dx = \frac{1}{(\pi/2\beta+c)^2} - \frac{1}{c^2}.$$

Using this in (3.6) brings

$$\beta\int_0^1 \frac{\sin(\beta x)}{(x+c)^2}\,dx \ge \frac{1}{(\pi/2\beta+c)^2} - \frac{\cos(\beta)}{(1+c)^2} \ge \frac{1}{(\pi/2\beta+c)^2} - \frac{1}{(1+c)^2}.$$

From this and (3.5) we finally deduce

$$\int_0^1 \frac{e^{\alpha x}\sin(\beta x)}{(x+c)^2}\,dx = I_1 + I_2 \ge \frac{1}{\beta}\left[\frac{1}{(\pi/2\beta+c)^2} - \frac{1}{(1+c)^2} - \frac{2\pi(1+e^{\alpha})}{(c+1-2\pi/\beta)^2}\right].$$

Note that we can choose $C(A)$ and $B_1(A)$ such that the term in the brackets is strictly positive when $0 < c < C(A)$, $\beta \ge B_1(A)$, and $0 < \alpha \le A$. □

Lemma 3.4. *Given positive A and B, there exists a constant $C_2 = C_2(A,B) > 0$ such that if $0 < \alpha \le A$, $|\beta| \le B$, and $c \le C_2$, then*

$$\int_0^1 \frac{e^{\alpha x}\cos(\beta x)}{(x+c)^2}\,dx > 0. \tag{3.8}$$

Proof. Assume that $0 \le \alpha \le A$ and $|\beta| \le B$. Integrating by parts, we obtain

$$\begin{aligned}\int_0^1 \frac{e^{\alpha x}\cos(\beta x)}{(x+c)^2}\,dx &= \frac{1}{c} - \frac{e^{\alpha}\cos(\beta)}{(1+c)} - \beta\int_0^1 \frac{e^{\alpha x}\sin(\beta x)}{(x+c)}\,dx + \alpha\int_0^1 \frac{e^{\alpha x}\cos(\beta x)}{(x+c)}\,dx\\ &= \frac{1}{c} - \frac{e^{\alpha}\cos(\beta)}{(1+c)} - \beta\int_0^1 \frac{e^{\alpha x}\sin(\beta x)}{(x+c)}\,dx\\ &\quad + \alpha\int_0^1 \frac{(e^{\alpha x}-1)\cos(\beta x)}{(x+c)}\,dx + \alpha\int_0^1 \frac{\cos(\beta x)}{(x+c)}\,dx\\ &\equiv M_1 + M_2 + M_3 + M_4 + M_5. \end{aligned} \tag{3.9}$$

Note that the terms $|M_2|, |M_3|$, and $|M_4|$ remain uniformly bounded as $c \to 0$ for all $0 \le \alpha \le A$ and $|\beta| \le B$. Concerning M_5, we have

$$M_5 = \alpha\int_0^1 \frac{\cos(\beta x)}{(x+c)}\,dx = \alpha\left[\cos(\beta)\log(1+c) - \log(c) - \beta\int_0^1 \sin(\beta x)\log(x+c)\,dx\right],$$

which is clearly positive if c is small enough. Therefore, due to the M_1 term, we can choose C sufficiently small so that (3.9) is positive for all $c \le C$. □

Let us now deduce Theorem 1.1 from these last two lemmas. For $A > 0$ we choose

$$C(A) \equiv \min\{C_1(A), C_2(A, B_1)\},$$

where B_1 and C_1 are the constants of Lemma 3.3 and C_2 is the constant of Lemma 3.4. Now, if $0 < \alpha \le A$, then either (3.8) or (3.3) holds, depending on the size of β. As a result, if $c < C(A)$, then all zeros of $\lambda \mapsto F(\lambda, c)$ must lie in the half-space $\Re\lambda \le -A/2$. As A was arbitrary, we may indeed, via (2.3), produce arbitrarily large decay. As A approaches ∞, we note that $C(A)$ decreases to 0 and so the total variation of $1/(x+c)$ approaches ∞.

4. The limit case $a(x) = 1/x$. In this section we prove the extinction result stated in Theorem 1.2. For the sake of simplicity we present here a formal argument. The rigorous proof can be achieved in a standard way.

Consider the limit damped wave equation

$$\begin{cases} u_{tt}(x,t) - u_{xx}(x,t) + \frac{2}{x}u_t(x,t) = 0, \quad 0 < x < 1, \quad t > 0,\\ u(0,t) = u(1,t) = 0,\\ u(x,0) = u_0, \quad u_t(x,0) = u_1. \end{cases} \tag{4.1}$$

We introduce the Laplace transform

$$\mathcal{L}\{u\}(\tau) = U(x,\tau) = \int_0^\infty e^{-\tau t}u(x,t)dt. \tag{4.2}$$

When applying the Laplace transform to (4.1), we obtain

$$\begin{cases} \tau^2 U - U_{xx} + \tau\frac{2}{x}U = \tau u_0(x) + u_1(x) + \frac{2}{x}u_0(x),\\ U(0,\tau) = U(1,\tau) = 0. \end{cases} \tag{4.3}$$

Observe that

$$U(x,\tau) = xe^{\tau x} \tag{4.4}$$

is a solution of the homogeneous equation associated to (4.3):

$$\tau^2 U - U_{xx} + \tau \frac{2}{x} U = 0. \tag{4.5}$$

We look for the solution of (4.3) via reduction of order, i.e., we assume that

$$U(x,\tau) = xe^{\tau x} v(x,\tau). \tag{4.6}$$

Then $v(x,\tau)$ must satisfy

$$\left\{ \begin{array}{l} xv_{xx} + v_x\left(2 + 2\tau x\right) = -e^{-\tau x}\left(\tau u_0 + u_1 + \frac{2u_0}{x}\right), \\ v(1,\tau) = 0, \\ \lim_{x\to 0} |v(x,\tau)| < \infty. \end{array} \right. \tag{4.7}$$

Hence

$$\begin{aligned} (v_x x^2 e^{-2\tau(1-x)})_x &= -xe^{-2\tau(1-x)} e^{-\tau x}\left(\tau u_0 + u_1 + \frac{2u_0}{x}\right), \\ v &= \int_x^1 \frac{1}{r^2} e^{2\tau(1-r)} \int_0^r s e^{-\tau(2-s)} \left(\tau u_0(s) + u_1(s) + \frac{2u_0(s)}{s}\right) ds dr, \end{aligned}$$

$$U = xe^{\tau x} \int_x^1 \frac{1}{r^2} e^{2\tau(1-r)} \int_0^r s e^{-\tau(2-s)} \left(\tau u_0(s) + u_1(s) + \frac{2u_0(s)}{s}\right) ds dr. \tag{4.8}$$

In order to invert the Laplace transform, we simplify the term in the form $\tau e^{\tau\alpha}$. Integrating by parts, we have

$$\begin{aligned} \int_0^r s e^{-\tau(2-s)} \tau u_0(s) ds = \int_0^r s u_0(s) \frac{d}{ds} e^{-\tau(2-s)} ds &= s u_0(s) e^{-\tau(2-s)} \Big]_0^r \\ &\quad - \int_0^r e^{-\tau(2-s)} \left(s u_0' + u_0\right) ds \\ &= r u_0(r) e^{-\tau(2-r)} - \int_0^r e^{-\tau(2-s)} \left(s u_0' + u_0\right) ds. \end{aligned} \tag{4.9}$$

Substituting in (4.8), we have

$$\begin{aligned} U &= xe^{\tau x} \int_x^1 \frac{1}{r} e^{2\tau(1-r)} u_0(r) e^{-\tau(2-r)} dr \\ &\quad + xe^{\tau x} \int_x^1 \frac{1}{r^2} e^{2\tau(1-r)} \int_0^r e^{-\tau(2-s)} \left(-s u_0'(s) + u_0(s) + s u_1(s)\right) ds dr. \end{aligned} \tag{4.10}$$

Now we apply the inverse Laplace transform $\mathcal{L}^{-1}$ to obtain u:

$$\begin{aligned} u(x,t) = \mathcal{L}^{-1}\{U\}(x,t) &= x \int_x^1 \frac{1}{r} u_0(r) \mathcal{L}^{-1}\{e^{\tau(x-r)}\} dr \\ &\quad + x \int_x^1 \frac{1}{r^2} \int_0^r \left(-s u_0'(s) + u_0(s) + s u_1(s)\right) \mathcal{L}^{-1}\{e^{\tau(x-2r+s)}\} ds dr \\ &= x \int_x^1 \frac{1}{r} u_0(r) \delta(t - r + x) dr \\ &\quad + x \int_x^1 \frac{1}{r^2} \int_0^r \left(-s u_0'(s) + u_0(s) + s u_1(s)\right) \delta(t - 2r + x + s) ds dr. \end{aligned} \tag{4.11}$$

Here $\delta(x)$ represents the Dirac delta at $x = 0$.

Once we have an explicit formula for the solution of (4.1), we easily prove the theorem.

Assume that $t > 2$. We have

$$t + x > 1,$$

and the first integral in (4.11) is zero due to the fact that the support of $\delta(t - r + x)$ is not in the domain of the integral. Moreover,

$$2r - x - t < 0$$

(because $r - x < 1$ and $r - t < 1 - 2 = -1$), and the second integral in (4.11) is also zero.

Acknowledgment. The first author is grateful to J. J. L. Velazquez for some fruitful discussions and comments related with this work.

REFERENCES

[1] A. Benaddi and B. Rao, *Energy decay rate of wave equations with indefinite damping*, J. Differential Equations, 161 (2000), pp. 337–357.

[2] S. J. Cox and M. L. Overton, *Perturbing the critically damped wave equation*, SIAM J. Appl. Math., 56 (1996), pp. 1353–1362.

[3] S. J. Cox and E. Zuazua, *The rate at which energy decays in a damped string*, Comm. Partial Differential Equations, 19 (1994), pp. 213–243.

[4] P. Freitas, *Optimizing the rate of decay of solutions of the wave equation using genetic algorithms: A counterexample to the constant damping conjecture*, SIAM J. Control Optim., 37 (1999), pp. 376–387.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1756–1778

THE TOPOLOGICAL ASYMPTOTIC FOR PDE SYSTEMS: THE ELASTICITY CASE*

STÉPHANE GARREAU[†], PHILIPPE GUILLAUME[‡], AND MOHAMED MASMOUDI[§]

Abstract. The aim of the topological sensitivity analysis is to obtain an asymptotic expansion of a design functional with respect to the creation of a small hole. In this paper, such an expansion is obtained and analyzed in the context of linear elasticity for general functionals and arbitrary shaped holes by using an adaptation of the adjoint method and a domain truncation technique. The method is general and can be easily adapted to other linear PDEs and other types of boundary conditions.

Key words. topological sensitivity, topological derivative, shape optimization, design sensitivity, elasticity, compliance

AMS subject classifications. 49Q10, 49Q12, 74P05, 74P10, 74P15

PII. S0363012900369538

1. Introduction.

The goal of topological optimization is to find an optimal design with an a priori poor information on the optimal shape of the structure. Unlike the case of classical shape optimization, the topology of the structure may change during the optimization process, as, for example, through the inclusion of holes.

Most of the known results concern structural mechanics. Classical topology optimization involves relaxed formulations and homogenization (see, e.g., [1, 19, 3, 12]). This approach leads us to introduce some microstructures. In the case of compliance (external work) minimization under a volume constraint, a class of laminated materials is exhibited, together with an explicit expression of the optimal material at each point of the structure. Such a method has two drawbacks.

- The optimal solution is not a classical design: it is a quasi-uniform distribution of composite materials. Then some penalization methods must be applied to retrieve a realistic shape.
- Optimization of a criterion like Von Mises stress via homogenization seems to be a difficult task.

For this latter reason, global optimization techniques like genetic algorithms or simulated annealing have been proposed (see, e.g., [22]), but these methods have a high computational cost and can hardly be applied to industrial problems.

Another approach instigated by the work of Schumacher [21] is presented and analyzed in this paper. The shape optimization problem consists in minimizing a functional $j(\Omega) = J(\Omega, u_\Omega)$, where the displacement u_Ω is defined on a variable open and bounded subset Ω of $\mathbb{R}^n$, $n = 2$ or 3. For $\rho > 0$, let $\Omega_\rho = \Omega \backslash \overline{(x_0 + \rho\omega)}$ be the set obtained by removing a small part $x_0 + \rho\omega$ from Ω, where $x_0 \in \Omega$ and $\omega \subset \mathbb{R}^n$ are a fixed open and bounded subset containing the origin. Then, an asymptotic expansion

*Received by the editors March 20, 2000; accepted for publication (in revised form) October 31, 2000; published electronically April 12, 2001.
http://www.siam.org/journals/sicon/39-6/36953.html

[†]CADOE SA, Novacité Alpha, 43 Boulevard du 11 Novembre, F 69603 Villeurbanne Cedex, France (stephane.garreau@cadoe.com).

[‡]Département de Génie Mathématique, INSA Toulouse and CNRS UMR 5640 MIP, Complexe Scientifique de Rangueil, F 31077 Toulouse Cedex 4, France (guillaum@gmm.insa-tlse.fr).

[§]UMR MIG, Université Paul Sabatier and CNRS UMR 5640 MIP, 118 route de Narbonne - F 31062 Toulouse Cedex, France (masmoudi@mip.ups-tlse.fr).

of the function j can be obtained in the following form:

$$j(\Omega_\rho) = j(\Omega) + f(\rho)g(x_0) + o(f(\rho)), \qquad \lim_{\rho\to 0} f(\rho) = 0, \quad f(\rho) > 0. \tag{1.1}$$

The "topological sensitivity" $g(x_0)$ provides an information for creating a small hole located at x_0. Hence the function g can be used like a descent direction in an optimization process. The physical meaning of creating a hole depends on the boundary condition which is imposed on its boundary.

- A homogeneous Neumann boundary condition means that $x_0 + \rho\omega$ represents a perforation, i.e., a lack of material.
- A homogeneous Dirichlet boundary condition means that $x_0 + \rho\omega$ represents a weld or a rivet. This kind of boundary condition is difficult to handle by classical homogenization methods.

Schumacher [21] introduced the topological sensitivity in the case of compliance minimization. Next, Sokolowski and Żochowski [23] generalized it to a class of functionals in the plane stress case with a homogeneous Neumann condition imposed on the boundary of a circular hole. A topological sensitivity framework using an adaptation of the adjoint method and a truncation technique has been introduced in [17] in the case of a homogeneous Dirichlet condition imposed on the boundary of a circular hole (see also [6]). The fundamental property of an adjoint technique is to provide the variation of a function with respect to a parameter by using a solution u_Ω and an adjoint state v_Ω which do not depend on the chosen parameter. Numerically, it means that only two systems must be solved for obtaining the discrete approximation of $g(x)$ for all $x \in \Omega$. In [10], the topological sensitivity has been derived for a large class of problems, functionals, and boundary conditions on a circular hole $x_0 + \rho\omega$. For example, in linear elasticity the first variation of the function j reads

$$\begin{aligned} j(\Omega_\rho) - j(\Omega) = -\frac{\pi(\lambda+2\mu)}{\mu(9\lambda+14\mu)} &\{20\mu\sigma(u_\Omega) : \varepsilon(v_\Omega) \\ &+(3\lambda-2\mu)\mathrm{tr}\,\sigma(u_\Omega)\,\mathrm{tr}\,\varepsilon(v_\Omega)\}\rho^3 \; + o(\rho^3) \end{aligned} \tag{1.2}$$

for a three-dimensional (3D) Neumann boundary condition on the hole. In the case of a 3D Dirichlet boundary condition, the variation becomes

$$j(\Omega_\rho) - j(\Omega) = \frac{12\pi\mu(\lambda+2\mu)}{2\lambda+5\mu}\, u_\Omega . v_\Omega\, \rho + o(\rho). \tag{1.3}$$

In [23, 14], only variations in $O(\rho^n)$, that is, proportional to the volume of the hole, were considered. One can observe here that other behaviors may occur, depending on the boundary condition which is imposed on the hole.

In this paper is presented the mathematical analysis of the topological sensitivity in the case of general functionals and arbitrary shaped holes. For example, optimization of a criterion like Von Mises stress can be handled by this method. The case of arbitrary shaped holes can be interesting for the identification of cracks. The equations which are here considered are those of linear elasticity, with a Neumann boundary condition on the hole. However, the method is general and can be adapted without any significant modification to many other equations like Stokes or Helmholtz equations. For these equations, the natural boundary condition on the hole is a Dirichlet condition for which the analysis is very similar to the one presented here.

First, an adaptation of the adjoint method to the topological context [17] is developed in section 2. Next, the elasticity problem, the truncation method for simulating the creation of a hole, and the main result are presented in section 3. In the case of a circular or spherical hole, explicit expressions of the topological sensitivity are given for both Dirichlet and Neumann boundary conditions and dimension $n = 2$ or 3. Section 4 is devoted to the proof of the main result presented in section 3. Finally, section 5 describes a topology optimization algorithm illustrated by some numerical examples.

2. A generalized adjoint method.

In this section, the adjoint method [4] is adapted to the topology optimization [17, 10]. Let $\mathcal{V}$ be a fixed Hilbert space. For $\rho \geq 0$, let a_ρ be a bilinear, symmetric, uniformly continuous, and coercive form on $\mathcal{V}$, and let l_ρ be a linear and uniformly continuous form on $\mathcal{V}$. That is, there exist constants $\alpha > 0$, $M > 0$, and $L > 0$ independent of ρ such that for all $\rho \geq 0$,

$$\begin{aligned} a_\rho(u,v) &\leq M\|u\|\,\|v\| \qquad \forall u,\, v \in \mathcal{V}, \\ a_\rho(u,u) &\geq \alpha\|u\|^2 \qquad \forall u \in \mathcal{V}, \\ |l_\rho(v)| &\leq L\|v\| \qquad \forall v \in \mathcal{V}. \end{aligned}$$

We assume in this section that there exist a bilinear and continuous form δ_a, a linear and continuous form δ_l, and a real function $f(\rho) > 0$ defined on $\mathbb{R}_+$ such that

$$\lim_{\rho\to 0} f(\rho) = 0, \tag{2.1}$$

$$\|a_\rho - a_0 - f(\rho)\delta_a\|_{\mathcal{L}_2(\mathcal{V})} = o(f(\rho)), \tag{2.2}$$

$$\|l_\rho - l_0 - f(\rho)\delta_l\|_{\mathcal{L}(\mathcal{V})} = o(f(\rho)), \tag{2.3}$$

where $\mathcal{L}(\mathcal{V})$ (respectively, $\mathcal{L}_2(\mathcal{V})$) denotes the space of continuous and linear (respectively, bilinear) forms on $\mathcal{V}$. It will be shown in section 4 that this assumption is fulfilled in the topology optimization context. The same function f is used here for both asymptotics (2.2)–(2.3). It does not exclude the case where $a_\rho - a_0$ and $l_\rho - l_0$ have different behaviors $O(f_1(\rho))$ and $O(f_2(\rho))$, in which case f is chosen to be the "slowest" between f_1 and f_2; that is, $f_i(\rho) = O(f(\rho))$, $i = 1, 2$.

For $\rho \geq 0$, let u_ρ be the solution to the problem: find $u_\rho \in \mathcal{V}$ such that

$$a_\rho(u_\rho, v) = l_\rho(v) \qquad \forall v \in \mathcal{V}. \tag{2.4}$$

LEMMA 2.1. *For $\rho \geq 0$, this problem has a unique solution u_ρ, and*

$$\|u_\rho - u_0\| = O(f(\rho)).$$

Proof. It follows from the coercivity of a_ρ that

$$\alpha\|u_\rho - u_0\|^2 \leq a_\rho(u_\rho - u_0, u_\rho - u_0), \tag{2.5}$$

which implies

$$\begin{aligned} \alpha\|u_\rho - u_0\|^2 &\leq a_\rho(u_\rho, u_\rho - u_0) - a_\rho(u_0, u_\rho - u_0) \\ &= l_\rho(u_\rho - u_0) - a_\rho(u_0, u_\rho - u_0) \\ &= l_0(u_\rho - u_0) + (l_\rho - l_0)(u_\rho - u_0) - a_\rho(u_0, u_\rho - u_0) \\ &= a_0(u_0, u_\rho - u_0) - a_\rho(u_0, u_\rho - u_0) + (l_\rho - l_0)(u_\rho - u_0) \\ &= f(\rho)(\delta_a(u_0, u_\rho - u_0) + \delta_l(u_\rho - u_0)) + (\|u_0\| + 1)\,\|u_\rho - u_0\|\,o(f(\rho)). \qquad \square \end{aligned}$$

Now consider a function $j(\rho) = J_\rho(u_\rho)$, where J_0 is differentiable with respect to u, its derivative being denoted by $DJ(u)$. Moreover, we suppose that there exists a function δ_J defined on $\mathcal{V}$ such that

$$J_\rho(v) - J_0(u) = DJ(u)(v-u) + f(\rho)\delta_J(u) + o(\|v-u\| + f(\rho)). \tag{2.6}$$

This expression looks like a first order derivative and would be, in fact, the first order derivative of the function $\mathcal{J}(s,u)$ defined by $\mathcal{J}(s,u) := J_{f^{-1}(s)}(v) - J_0(u)$ with the change of variable $s = f(\rho)$.

Next, the Lagrangian $\mathcal{L}$ [4] is defined by

$$\mathcal{L}_\rho(u,v) = J_\rho(\,u) + a_\rho(u,v) - l_\rho(v) \qquad \forall u,v \in \mathcal{V}. \tag{2.7}$$

Its variation with respect to ρ is given by

$$\delta_{\mathcal{L}}(u,v) = \delta_J(u) + \delta_a(u,v) - \delta_l(v),$$

and we have

$$\mathcal{L}_\rho(u,v) - \mathcal{L}_0(u,v) = f(\rho)\delta_{\mathcal{L}}(u,v) + o(f(\rho)).$$

THEOREM 2.2. *The function j has the asymptotic expansion*

$$j(\rho) = j(0) + f(\rho)\delta_{\mathcal{L}}(u_0,v_0) + o(f(\rho)), \tag{2.8}$$

where v_0 is the solution to the adjoint problem: find $v_0 \in \mathcal{V}$ such that

$$a_0(w,v_0) = -DJ(u_0)w \qquad \forall w \in \mathcal{V}. \tag{2.9}$$

Proof. For all $v \in \mathcal{V}$, one has

$$j(\rho) = \mathcal{L}_\rho(u_\rho, v).$$

Hence

$$\begin{aligned} j(\rho) - j(0) &= \mathcal{L}_\rho(u_\rho,v) - \mathcal{L}_0(u_0,v),\\ &= a_\rho(u_\rho,v) - a_0(u_0,v) + J_\rho(u_\rho) - J_0(u_0) - l_\rho(v) + l_0(v). \end{aligned}$$

It follows from (2.6) and Lemma 2.1 that

$$J_\rho(\,u_\rho) - J_0(u_0) = DJ(u_0)(u_\rho - u_0) + f(\rho)\delta_J(u_0) + o(f(\rho)).$$

Next, choosing v_0 as the solution to (2.9), we obtain with (2.3)

$$\begin{aligned} j(\rho) - j(0) &= a_\rho(u_\rho,v_0) - a_0(u_0,v_0) + DJ(u_0)(u_\rho - u_0)\\ &\quad + f(\rho)(\delta_J(u_0) - \delta_l(v_0)) + o(f(\rho))\\ &= a_\rho(u_\rho,v_0) - a_0(u_\rho,v_0) + a_0(u_\rho - u_0, v_0) + DJ(u_0)(u_\rho - u_0)\\ &\quad + f(\rho)(\delta_J(u_0) - \delta_l(v_0)) + o(f(\rho))\\ &= a_\rho(u_\rho,v_0) - a_0(u_\rho,v_0) + f(\rho)(\delta_J(u_0) - \delta_l(v_0)) + o(f(\rho)). \end{aligned}$$

Then, it follows from (2.1), (2.2), and Lemma 2.1 (with $||u_\rho||$ bounded) that

$$\begin{aligned} j(\rho) - j(0) &= f(\rho)\delta_a(u_\rho,v_0) + f(\rho)(\delta_J(u_0) - \delta_l(v_0)) + o(f(\rho))\\ &= f(\rho)(\delta_a(u_0,v_0) + \delta_a(u_\rho - u_0, v_0)) + f(\rho)(\delta_J(u_0) - \delta_l(v_0)) + o(f(\rho))\\ &= f(\rho)\delta_{\mathcal{L}}(u_0,v_0) + o(f(\rho)). \qquad \square \end{aligned}$$

3. The elasticity problem.

Let Ω be an open and bounded subset of $\mathbb{R}^n$, $n = 2$ or 3. The linear elasticity problem [7, 13] is the following: find u_Ω such that

$$\left\{\begin{array}{rcll} -\mathrm{div}\,\sigma(u_\Omega) & = & 0 & \text{in } \Omega, \\ \sigma(u_\Omega)\mathbf{n} & = & F & \text{on } \Gamma_N, \\ u_\Omega & = & 0 & \text{on } \Gamma_D, \end{array}\right. \tag{3.1}$$

where the strain tensor ε and the stress tensor σ are given by

$$\begin{aligned} \sigma_{ij}(u) &= \lambda \mathrm{div}\, u\, \delta_{ij} + 2\mu\varepsilon_{ij}(u), \\ \varepsilon_{ij}(u) &= \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \qquad 1 \le i, \quad j \le n. \end{aligned}$$

Here δ_{ij} is the Kronecker symbol, and $\mathbf{n}$ denotes the outward normal to the boundary $\Gamma = \Gamma_D \cup \Gamma_N$, where Γ_D and Γ_N have both a nonnegative Lebesgue measure and $\Gamma_D \cap \Gamma_N = \emptyset$. More general stress tensors of the form $\sigma_{ij}(u) = H_{ij}^{kl}\varepsilon_{kl}(u)$ could be considered without any major modifications. Equations (3.1) describe the displacement u_Ω of an isotropic solid which is loaded by a surface traction F on Γ_N and which is clamped on Γ_D. For simplicity, no volume forces are considered.

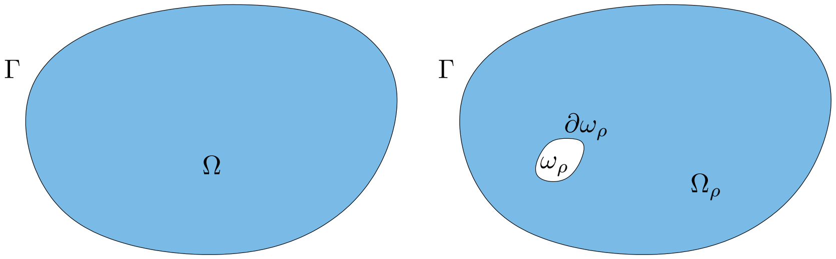


FIG. 3.1. *The initial domain and the same domain after the perforation.*

For a given $x_0 \in \Omega$, consider the perforated open set $\Omega_\rho = \Omega \backslash \overline{\omega_\rho}$, $\omega_\rho = x_0 + \rho\omega$, where $\omega \subset \mathbb{R}^n$ is a fixed open and bounded set containing the origin ($\omega_\rho = \emptyset$ if $\rho = 0$), whose boundary $\partial\omega$ is connected and piecewise of class $\mathcal{C}^1$ (see Figure 3.1). The new displacement u_{Ω_ρ} is the solution to the problem: find u_{Ω_ρ} such that

$$\left\{\begin{array}{rcll} -\mathrm{div}\,\sigma(u_{\Omega_\rho}) & = & 0 & \text{in } \Omega_\rho, \\ \sigma(u_{\Omega_\rho})\mathbf{n} & = & F & \text{on } \Gamma_N, \\ u_{\Omega_\rho} & = & 0 & \text{on } \Gamma_D, \\ \sigma(u_{\Omega_\rho})\mathbf{n} & = & 0 & \text{on } \partial\omega_\rho. \end{array}\right. \tag{3.2}$$

Note that for $\rho = 0$, one has $u_{\Omega_0} = u_\Omega$.

In the context of identification of conductivity imperfections, the asymptotic behavior of a voltage potential $u_{\Omega_\rho} - u_\Omega$ has been studied in [24, 25] for the Laplace equation. In that context, a new variation of u_Ω is computed for each new location of a hole. This method is usually referred to as the direct method. An alternative to this method is to use the adjoint technique. In this case, it is not necessary to compute the variation of u_Ω in order to obtain the variation of a cost function; only an adjoint state must be evaluated, which is independent of the location of the hole.

The displacement u_{Ω_ρ} is defined on the variable domain Ω_ρ; thus it belongs to a functional space which depends on ρ. Hence if we want to derive the asymptotic expansion of a function of the form

$$j(\rho) = \widetilde{J}(u_{\Omega_\rho}), \tag{3.3}$$

we cannot apply directly the tools of section 2, which require a fixed functional space.

In classical shape optimization, this requirement can be satisfied with the help of a domain parameterization technique [18, 16, 11]. This technique involves a fixed domain and a bi-Lipshitz map between this domain and the modified one. In the topology optimization context, such a map does not exist between Ω and Ω_ρ. However, a functional space independent of ρ can be constructed by using a domain truncation technique. Let $R > 0$ be such that the open ball $B(x_0, R)$ is included in Ω. Then the truncated open set Ω_R (see Figure 3.2) is defined by

$$\Omega_R = \Omega \setminus \overline{B(x_0, R)}.$$

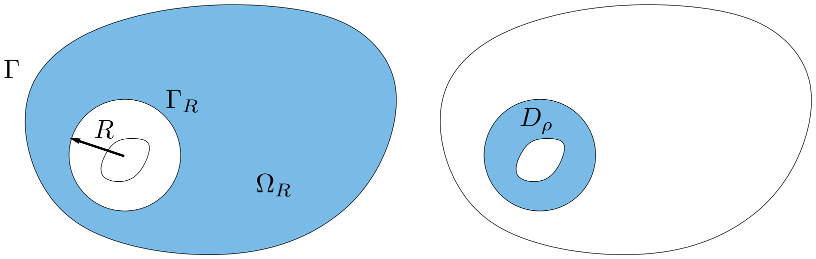


Fig. 3.2. *The truncated domain.*

In section 3.1 are defined

- a Hilbert space $\mathcal{V}_R$ independent of ρ,
- a $\mathcal{V}_R$-elliptic bilinear and continuous form a_ρ, and
- a linear and continuous form l,

such that the solution u_ρ to the equation

$$a_\rho(u_\rho, v) = l(v) \qquad \forall v \in \mathcal{V}_R$$

is equal to the restriction of u_{Ω_ρ} to Ω_R. A bilinear form δ_a satisfying (2.2) will be obtained in section 4, from which the asymptotic expansion of the cost function will be derived by using the framework described in section 2. As no volume forces are considered, we have $\delta_l \equiv 0$. The main result is presented in section 3.2, and the particular case of a spherical hole is detailed in section 3.3.

3.1. Truncation. The open set $B(x_0, R) \setminus \overline{\omega_\rho}$ is denoted by D_ρ (see Figure 3.2). For $\varphi \in H^{1/2}(\Gamma_R)^n$ and $\rho > 0$, let u_ρ^φ be the solution to the problem: find u_ρ^φ such that

$$\left\{ \begin{array}{rcll} -\operatorname{div} \sigma(u_\rho^\varphi) & = & 0 & \text{in } D_\rho, \\ u_\rho^\varphi & = & \varphi & \text{on } \Gamma_R, \\ \sigma(u_\rho^\varphi)\mathbf{n} & = & 0 & \text{on } \partial\omega_\rho, \end{array} \right. \tag{3.4}$$

where Γ_R is the boundary of $B(x_0, R)$. The normal $\mathbf{n}$ is chosen outward to D_ρ on $\partial\omega_\rho$ and Γ_R, regardless of whether D_ρ or Ω_R are considered. For $\rho = 0$, the function u_0^φ is the solution to

$$\left\{ \begin{array}{rcll} -\operatorname{div} \sigma(u_0^\varphi) & = & 0 & \text{in } B(x_0, R), \\ u_0^\varphi & = & \varphi & \text{on } \Gamma_R. \end{array} \right. \tag{3.5}$$

For $\rho \geq 0$, the Dirichlet-to-Neumann operator T_ρ is defined by

$$\begin{array}{rcl} T_\rho : H^{1/2}(\Gamma_R)^n & \longrightarrow & H^{-1/2}(\Gamma_R)^n, \\ \varphi & \longmapsto & T_\rho \varphi = \sigma(u_\rho^\varphi)\mathbf{n}. \end{array}$$

One can observe that the null space of T_ρ consists in the constant functions. Finally, the displacement u_ρ is defined for $\rho \geq 0$ as the solution to the truncated problem: find u_ρ such that

$$\left\{ \begin{array}{rcll} -\operatorname{div} \sigma(u_\rho) & = & 0 & \text{in } \Omega_R, \\ \sigma(u_\rho)\mathbf{n} & = & F & \text{on } \Gamma_N, \\ u_\rho & = & 0 & \text{on } \Gamma_D, \\ \sigma(u_\rho)\mathbf{n} & = & T_\rho u_\rho & \text{on } \Gamma_R. \end{array} \right. \tag{3.6}$$

The variational formulation associated to problem (3.6) is the following: find $u_\rho \in \mathcal{V}_R$ such that

$$a_\rho(u_\rho, v) = l(v) \qquad \forall v \in \mathcal{V}_R,$$

where the functional space $\mathcal{V}_R$, the bilinear form a_ρ, and the linear form l are defined by

$$\begin{aligned} \mathcal{V}_R &= \left\{ u \in H^1(\Omega_R)^n,\ u = 0 \text{ on } \Gamma_D \right\}, \\ a_\rho(u, v) &= \int_{\Omega_R} \sigma(u) : \varepsilon(v)\, dx + \int_{\Gamma_R} T_\rho u.v \, d\gamma(x), \\ l(v) &= \int_{\Gamma_N} F.v \, d\gamma(x). \end{aligned} \tag{3.7}$$

Here $x.y$ denotes the usual dot product of $\mathbb{R}^n$, $\sigma : \varepsilon = \sum_{i,j=1}^n \sigma_{ij}\varepsilon_{ij}$, and $d\gamma(x)$ is the Lebesgue measure on the boundary. Symmetry, continuity, and coercivity of a_ρ follow directly from

$$\int_{\Gamma_R} T_\rho \varphi.\psi \, d\gamma(x) = \int_{D_\rho} \sigma(u_\rho^\varphi) : \varepsilon(u_\rho^\psi)\, dx.$$

The following result is standard in PDE theory.

Proposition 3.1. *Problems* (3.2) *and* (3.6) *have a unique solution. Moreover, the restriction to* Ω_R *of the solution* u_{Ω_ρ} *to problem* (3.2) *is the solution* u_ρ *to problem* (3.6).

We have now at our disposal the fixed Hilbert space $\mathcal{V}_R$ required by section 2. Function (3.3) can be redefined in the following way: for $u \in \mathcal{V}_R$, let $\widetilde{u} \in H^1(\Omega_\rho)^n$ be the extension of u which coincides with u on Ω_R and Γ_R and which satisfies $\operatorname{div} \sigma(\widetilde{u}) = 0$ on D_ρ, $\sigma(\widetilde{u})\mathbf{n} = 0$ on $\partial\omega_\rho$. Then a function J_ρ can be defined on $\mathcal{V}_R$ by

$$J_\rho(u) = \widetilde{J}(\widetilde{u}). \tag{3.8}$$

In particular, it follows from the previous proposition that

$$j(\rho) = \widetilde{J}(u_{\Omega_\rho}) = J_\rho(u_\rho). \tag{3.9}$$

Notice that $J_\rho(u_\rho)$ is independent of the choice of R. For example, if $\widetilde{J}(u_{\Omega_\rho}) = \int_{\Omega_\rho} |u_{\Omega_\rho}|^2 dx$, then we have

$$J_\rho(u) = \int_{\Omega_R} |u|^2 dx + \int_{D_\rho} |u_\rho^\varphi|^2 dx, \quad u \in \mathcal{V}_R \text{ and } \varphi = u \text{ on } \Gamma_R.$$

3.2. The main result. Possibly changing the coordinate system, we can suppose for convenience that $x_0 = 0$. Let v_ω be the solution to the problem

$$\left\{ \begin{array}{rcll} -\operatorname{div}\sigma(v_\omega) & = & 0 & \text{in } \mathbb{R}^n\backslash\overline{\omega}, \\ v_\omega & = & 0 & \text{at } \infty, \\ \sigma(v_\omega)\mathbf{n} & = & \sigma(u_\Omega)(x_0)\mathbf{n} & \text{on } \partial\omega. \end{array} \right. \tag{3.10}$$

In order to be consistent with previous notations, the normal $\mathbf{n}$ is chosen outward to $\mathbb{R}^n\backslash\overline{\omega}$ on $\partial\omega$. This function v_ω can be expressed by a single layer potential on $\partial\omega$ in the following way. Let

$$E(y) = \frac{1}{r}\left(\beta I + \gamma \mathrm{e}_r \mathrm{e}_r^T\right) \quad \text{if } n = 3, \tag{3.11}$$

$$E(y) = \beta \log r\, I + \gamma \mathrm{e}_r \mathrm{e}_r^T \quad \text{if } n = 2,$$

where I is the $n \times n$ identity matrix, $r = ||y||$, $\mathrm{e}_r = y/r$, e_r^T is the transposed vector of e_r, and

$$\beta = \frac{\lambda + 3\mu}{8\pi\mu(\lambda + 2\mu)}, \quad \gamma = \frac{\lambda + \mu}{8\pi\mu(\lambda + 2\mu)} \quad \text{for } n = 3,$$

$$\beta = -\frac{\lambda + 3\mu}{4\pi\mu(\lambda + 2\mu)}, \quad \gamma = \frac{\lambda + \mu}{4\pi\mu(\lambda + 2\mu)} \quad \text{for } n = 2 \text{ (plain strain).}$$

For two-dimensional (2D) plain stress, $\lambda^* = 2\mu\lambda/(\lambda + 2\mu)$ must be substituted for λ. The matrix distribution $E \in \mathcal{D}'(\mathbb{R}^n, \mathbb{R}^{n\times n})$ is a fundamental solution for the elasticity problem in $\mathbb{R}^n$; that is, each column E_j is a solution to

$$-\operatorname{div}\sigma(E_j) = \delta \mathrm{e}_j \quad \text{in } \mathbb{R}^n,$$

where δ is the Dirac distribution and $(\mathrm{e}_j)_{j=1,n}$ is the canonical basis of $\mathbb{R}^n$. Then function v_ω reads

$$v_\omega(y) = \int_{\partial\omega} E(y - x)p(x)\, d\gamma(x), \quad y \in \mathbb{R}^n\backslash\overline{\omega},$$

where $p \in H^{-1/2}(\partial\omega)^n$ is the solution to boundary integral equation [8, 9]

$$\frac{p(y)}{2} + \int_{\partial\omega} \sigma_y(E(y - x)p(x))\mathbf{n}(y)\, d\gamma(x) = \sigma(u_\Omega)(x_0)\mathbf{n}(y) \quad \forall y \in \partial\omega, \tag{3.12}$$

the subscript y in σ_y denoting a differentiation with respect to the variable y. Moreover, due to

$$\int_{\partial\omega} \sigma(u_\Omega)(x_0)\mathbf{n}(x)\ d\gamma(x) = 0,$$

$$\int_{\partial\omega} \sigma_y(E_j(y - x))\mathbf{n}(y)\, d\gamma(y) = \frac{1}{2}\mathrm{e}_j \quad \text{if } x \in \partial\omega, \quad \partial\omega \text{ of class } \mathcal{C}^1 \text{ at } x,$$

multiplying (3.12) by e_i and integrating over $\partial\omega$ yield

$$\begin{aligned} 0 &= \frac{1}{2}\int_{\partial\omega} p_i(y)\, d\gamma(y) + \int_{\partial\omega} p_j(x) \int_{\partial\omega} \sigma_y(E_j(y - x))\mathbf{n}(y)\, d\gamma(y).\mathrm{e}_i\, d\gamma(x) \\ &= \frac{1}{2}\int_{\partial\omega} p_i(y)\, d\gamma(y) + \frac{1}{2}\int_{\partial\omega} p_j(x)\delta_{ij}\, d\gamma(x) \\ &= \int_{\partial\omega} p_i(y)\, d\gamma(y), \end{aligned}$$

where Einstein's summation convention on repeated indices is used throughout this paper. Hence we have

$$\int_{\partial\omega} p(x)\, d\gamma(x) = 0, \tag{3.13}$$

and, using the first order Taylor expansion of E at the point $y \neq x$ for x bounded

$$E(y-x) = E(y) - DE(y)x + O\left(\frac{1}{r^n}\right),$$

we obtain the following asymptotic expansion at infinity of the function v_ω:

$$\begin{aligned} v_\omega(y) &= \int_{\partial\omega} E(y)p(x) \ - (DE(y)x)p(x)\, d\gamma(x) + O\left(\frac{1}{r^n}\right) \\ &= 0 - \int_{\partial\omega} (DE(y)x)p(x)\, d\gamma(x) + O\left(\frac{1}{r^n}\right). \end{aligned}$$

The dominant part $P_\omega(y)$ of v_ω is given by

$$v_\omega(y) = P_\omega(y) + W_\omega(y), \tag{3.14}$$

$$P_\omega(y) = -\int_{\partial\omega} (DE(y)x)p(x)\, d\gamma(x), \quad W_\omega(y) = O\left(\frac{1}{r^n}\right), \tag{3.15}$$

where the function $r^{n-1}P_\omega(y)$ is homogeneous of degree 0; that is, for all $y \neq 0$,

$$P_\omega(y) = \rho^{n-1}P_\omega(\rho y) \quad \forall \rho > 0. \tag{3.16}$$

Moreover, it can easily be checked that

$$-\operatorname{div}\sigma(P_\omega) = 0 \quad \text{in } \mathbb{R}^n\backslash\{0\}. \tag{3.17}$$

Next we consider the solution Q_ω to the problem

$$\left\{ \begin{array}{rcll} -\operatorname{div}\sigma(Q_\omega) & = & 0 & \text{in } D_0, \\ Q_\omega & = & P_\omega & \text{on } \Gamma_R. \end{array} \right. \tag{3.18}$$

The main result is the following, which will be proved in section 4.

THEOREM 3.2. *Let $j(\rho) = J_\rho(u_\rho)$ be a cost function such that for all $v \in \mathcal{V}_R$ and $\rho > 0$,*

$$J_\rho(v) - J_0(u_0) = DJ(u_0)(v - u_0) + \delta_J(u_0)\rho^n + o(\|v - u_0\|_{\mathcal{V}_R} + \rho^n),$$

where $DJ(u_0)$ is continuous and linear on $\mathcal{V}_R$. Let $v_0 \in \mathcal{V}_R$ be the solution to the adjoint equation

$$a_0(w, v_0) = -DJ(u_0)w \quad \forall w \in \mathcal{V}_R, \tag{3.19}$$

and let

$$\delta_a(u_0, v_0) := \int_{\Gamma_R} \sigma(Q_\omega - P_\omega)\mathbf{n}.v_0\, d\gamma(x). \tag{3.20}$$

Then the function j has the following asymptotic expansion:

$$j(\rho) = j(0) + \{\delta_a(u_0, v_0) + \delta_J(u_0)\}\,\rho^n + o(\rho^n). \tag{3.21}$$

The expression $g(x_0) := \delta_a(u_0, v_0) + \delta_J(u_0)$ is called the "topological sensitivity." It is also called the "topological gradient." Moreover, as $g(x_0)$ is independent of ρ, it follows from the uniqueness of an asymptotic expansion that $g(x_0)$ is also independent of R.

Practically, what is computed is the solution u_Ω to (3.1) and the solution v_Ω to

$$\int_\Omega \sigma(w) : \varepsilon(v_\Omega)\, dx = -D\widetilde{J}(u_\Omega)w \quad \forall w \in \mathcal{V}_0. \tag{3.22}$$

As observed in Proposition 3.1, u_0 is the restriction to Ω_R of u_Ω. The same property holds for v_0 and v_Ω. This can easily be seen by observing that for $w \in \mathcal{V}_0$ such that $\operatorname{div} \sigma(\omega) = 0$ in D_0, and denoting by v_R and w_R the restrictions of v_Ω and w to Ω_R, on the one hand, we have

$$\begin{aligned} a_0(w_R, v_R) &= \int_{\Omega_R} \sigma(w_R) : \varepsilon(v_R)\, dx + \int_{\Gamma_R} T_0 w_R . v_R \, d\gamma(x) \\ &= \int_\Omega \sigma(w) : \varepsilon(v_\Omega)\, dx + 0, \end{aligned} \tag{3.23}$$

and on the other hand, due to (3.8), we have $\widetilde{J}(u) = J_0(u_R)$ for all $u \in \mathcal{V}_0$ such that $\operatorname{div} \sigma(u) = 0$ in D_0. Hence

$$D\widetilde{J}(u_\Omega)w = DJ(u_0)w_R. \tag{3.24}$$

Then, gathering (3.23), (3.22), and (3.24), we obtain

$$a_0(w_R, v_R) = -DJ(u_0)w_R \quad \forall w_R \in \mathcal{V}_R,$$

which proves that v_R is the solution to (3.19); that is, v_0 is the restriction to Ω_R of v_Ω.

The basic property of an adjoint technique is also satisfied here, in that the displacement u_Ω (or u_0) and the adjoint state v_Ω (or v_0) do not depend on x_0. Hence only two systems have to be solved in order to compute the topological sensitivity $g(x)$ for all $x \in \Omega$. This plays a crucial role in the efficiency of the optimization algorithm described in section 5.

Thanks to Green's formula and $P_\omega = Q_\omega$ on Γ_R, (3.20)–(3.21) reads also

$$\begin{aligned} j(\rho) = j(0) + \Bigg\{ &\int_{\Gamma_R} \sigma(v_\Omega)\mathbf{n}.P_\omega - \sigma(P_\omega)\mathbf{n}.v_\Omega \, d\gamma(x) \\ &- \int_{D_0} \operatorname{div} \sigma(v_\Omega).Q_\omega \, dx + \delta_J(u_0) \Bigg\} \rho^n + o(\rho^n). \end{aligned} \tag{3.25}$$

In particular, we have the following result, which can be used when the function J does not depend on ρ. The tensor $-A$ is known as the mass matrix; see, e.g., [14].

COROLLARY 3.3. *Let p be defined by* (3.12)*, and*

$$A_{ik}(\sigma(u_\Omega)(x_0)) = \int_{\partial\omega} p_i(x) x_k \, d\gamma(x) \tag{3.26}$$

with $p = (p_i)_{1 \le i \le n}$, $x = (x_k)_{1 \le k \le n}$. Under the assumptions of Theorem 3.2*, if* $\operatorname{div} \sigma(v_\Omega) = 0$ *in* D_0*, then*

$$\delta_a(u_\Omega, v_\Omega) = A(\sigma(u_\Omega)(x_0)) : \varepsilon(v_\Omega)(x_0),$$

and the function j has the following asymptotic expansion:

$$j(\rho) = j(0) + \delta_a(u_\Omega, v_\Omega)\rho^n + o(\rho^n).$$

Moreover, the matrix $A(\sigma(u_\Omega)(x_0)) \in \mathbb{R}^{n\times n}$ is symmetric, is linear with respect to $\sigma(u_\Omega)(x_0)$, and depends only on $\sigma(u_\Omega)(x_0)$ and ω, and the tensor A is symmetric in the following sense:

$$A(\sigma(u_\Omega)(x_0)) : \varepsilon(v_\Omega)(x_0) = A(\sigma(v_\Omega)(x_0)) : \varepsilon(u_\Omega)(x_0).$$

Proof. Due to $\operatorname{div}\sigma(v_\Omega) = 0$, the function v_Ω is analytical around x_0, and through a regularization and localization technique, it can easily be shown that (3.26) becomes

$$j(\rho) = j(0) - \rho^n \left\langle \operatorname{div}\sigma(P_\omega), \varphi v_\Omega \right\rangle_{\mathcal{D}'(D_0),\mathcal{D}(D_0)} + o(\rho^n), \tag{3.27}$$

where $\varphi \in \mathcal{D}(D_0)$ satisfies $\varphi(x) = 1$ on a neighborhood of x_0. For legibility, A_{ik} may stand for $A_{ik}(\sigma(u_\Omega)(x_0))$. It follows from (3.15) that

$$P_\omega(y) = -A_{ik}\partial_k E_i(y).$$

Hence

$$-\operatorname{div}\sigma(P_\omega)(y) = A_{ik}\partial_k \operatorname{div}\sigma(E_i(y)) = -A_{ik}\partial_k\delta \mathrm{e}_i,$$

and

$$\begin{aligned} -\left\langle \operatorname{div}\sigma(P_\omega)(y), \varphi v_\Omega(y)\right\rangle &= \left\langle A_{ik}\delta\mathrm{e}_i, \partial_k(\varphi v_\Omega)(y)\right\rangle \\ &= A_{ik}(\partial_k v_\Omega)_i(x_0) \\ &= A(\sigma(u_\Omega)(x_0)) : Dv_\Omega(x_0). \end{aligned}$$

Function p is the jump of $\sigma(v_\omega)\mathbf{n}$ across $\partial\omega$ if v_ω, defined by (3.10), is extended in ω by $-\operatorname{div}\sigma(v_\omega) = 0$ in ω and v_ω is continuous across $\partial\omega$. Thus

$$A_{ik} = \int_{\partial\omega} [\sigma(v_\omega)\mathbf{n}].(\mathrm{e}_i x_k)\, d\gamma(x),$$

and $A_{ik} = A_{ki}$ follows from Green's formula and $\sigma_{ik}(v_\omega) = \sigma_{ki}(v_\omega)$. Finally, the symmetry of the tensor A is a straightforward consequence of the symmetry of a_ρ for all $\rho \geq 0$. □

3.3. Case of a spherical hole.

Computing the coefficients

$$A_{ik}(\sigma(u_\Omega)(x_0)) = \int_{\partial\omega} p_i(x) x_k\, d\gamma(x)$$

requires solving integral equation (3.12) which is here recalled:

$$\frac{p(y)}{2} + \int_{\partial\omega} \sigma_y(E(y-x)p(x))\mathbf{n}(y)\, d\gamma(x) = \sigma(u_\Omega)(x_0)\mathbf{n}(y) \quad \forall y \in \partial\omega.$$

However, using Saint-Venant's principle, a good approximation of p can be obtained by computing the displacement u_{ext} on $\Omega\backslash\omega$ and the displacement u_{int} on ω with the Dirichlet boundary condition $u_{int} = u_{ext}$ on $\partial\omega$. Then p is approximately equal to the jump $\sigma(u_{ext})\mathbf{n}-\sigma(u_{int})\mathbf{n}$ on $\partial\omega$. When ω is the unit ball $B(0,1)$, then v_ω (and thus P and A) can be computed explicitly with the help of symbolic calculus: with $\sigma_0 := \sigma(u)(0)$, the solution v to

$$\left\{\begin{array}{rcll} -\operatorname{div}\sigma(v) & = & 0 & \text{in } \mathbb{R}^n \setminus \overline{B(0,1)}, \\ \sigma(v)\mathbf{n} & = & \sigma_0\mathbf{n} & \text{on } \partial B(0,1), \\ v & = & 0 & \text{at } \infty, \end{array}\right.$$

is given for $n = 2$ by

$$v_i = \frac{\pi(\mu+\eta)}{2\eta\mu}\left\{4\mu\sigma_0 \colon \varepsilon(E_i) + (\eta - 2\mu)\operatorname{tr}\sigma_0 \operatorname{tr}\varepsilon(E_i) + \frac{\mu+2\eta}{6}\sigma_0 : \varepsilon(\Delta E_i)\right\}, \quad i = 1,2, \tag{3.28}$$

and for $n = 3$ by

$$v_i = \frac{\pi(\lambda+2\mu)}{\mu(9\lambda+14\mu)}\left\{20\mu\sigma_0 \colon \varepsilon(E_i) + (3\lambda - 2\mu)\operatorname{tr}\sigma_0 \operatorname{tr}\varepsilon(E_i) + 2\mu\sigma_0 : \varepsilon(\Delta E_i)\right\}, \quad i = 1,2,3, \tag{3.29}$$

where the constant η is defined by

$$\eta = \left\{\begin{array}{ll} \frac{\mu(3\lambda+2\mu)}{\lambda+2\mu} & \text{plane stress}, \\ \lambda + \mu & \text{plane strain}, \end{array}\right.$$

and $\operatorname{tr}\sigma = \sum_i \sigma_{ii}$ is the usual trace operator.

Also, some similar results can be obtained for a homogeneous Dirichlet boundary condition on $\partial\omega_\rho$. The main difference is the asymptotic behavior at infinity of the solution to the exterior problem corresponding to (3.10):

$$\left\{\begin{array}{rcll} -\operatorname{div}\sigma(v_\omega) & = & 0 & \text{in } \mathbb{R}^n \backslash \overline{\omega}, \\ v_\omega & = & u_\Omega(x_0) & \text{on } \partial\omega, \end{array}\right.$$

that is, $O(1/r)$ for $n = 3$ and $O(\log(r))$ for $n = 2$ instead of $O(1/r^{n-1})$ for the Neumann case (see (3.15)). The solution v_ω to this exterior problem involves E and ΔE, instead of DE and $D(\Delta E)$.

Table 3.1 reports the different expressions of the topological sensitivity for the Dirichlet and Neumann boundary condition in two and three dimensions. For the Neumann boundary condition, they are obtained by simply keeping the principal part P_ω in (3.30) and (3.29) (that is, by removing the terms containing ΔE_i), and by substituting $-v_\Omega$ for E_i. That follows from a direct computation of (3.27) using (3.2), similar to the proof of Corollary 3.3. In particular, we retrieve the expression given by Sokolowski and Żochowski [23] in plane stress elasticity with a Neumann boundary condition.

One can observe that in both cases $\delta_a(u_\Omega, v_\Omega) \neq \sigma(u_\Omega) : \varepsilon(v_\Omega)$; that is, the topological sensitivity is not obtained by considering the limit when $\rho \to 0$ of the classical shape optimization expression of the derivative with respect to $\rho > 0$ (here for the Neumann case):

$$Dj(\rho) = -\int_{\partial\omega_\rho} \sigma(u_\Omega) : \varepsilon(v_\Omega)\, d\gamma(x).$$

Table 3.1
Expressions of the topological sensitivity.

Boundary condition on $\partial\omega_\rho$	$f(\rho)$	$\delta_a(u_\Omega, v_\Omega)$
2D Dirichlet	$\dfrac{1}{\log(\rho)}$	$-\dfrac{4\pi\mu(\mu+\eta)}{2\mu+\eta} u_\Omega . v_\Omega$
3D Dirichlet	ρ	$\dfrac{12\pi\mu(\lambda+2\mu)}{2\lambda+5\mu} u_\Omega . v_\Omega$
2D Neumann	ρ^2	$-\dfrac{\pi(\mu+\eta)}{2\eta\mu}\{4\mu\sigma(u_\Omega):\varepsilon(v_\Omega)+(\eta-2\mu)\mathrm{tr}\,\sigma(u_\Omega)\,\mathrm{tr}\,\varepsilon(v_\Omega)\}$
3D Neumann	ρ^3	$-\dfrac{\pi(\lambda+2\mu)}{\mu(9\lambda+14\mu)}\{20\mu\sigma(u_\Omega):\varepsilon(v_\Omega)+(3\lambda-2\mu)\mathrm{tr}\,\sigma(u_\Omega)\,\mathrm{tr}\,\varepsilon(v_\Omega)\}$

4. Proof of the main result. This section consists in the proof of Theorem 3.2. The variation of the bilinear form a_ρ (see (3.7)) reads

$$a_\rho(u,v) - a_0(u,v) = \int_{\Gamma_R} (T_\rho - T_0)u.v \, d\gamma(x). \tag{4.1}$$

Hence, the problem reduces to the analysis of $(T_\rho - T_0)\varphi$ for $\varphi \in H^{1/2}(\Gamma_R)^n$. More precisely, it will be shown that there exists an operator $\delta_T \in \mathcal{L}(H^{1/2}(\Gamma_R)^n; H^{-1/2}(\Gamma_R)^n)$ such that

$$\|T_\rho - T_0 - \rho^n \delta_T\|_{\mathcal{L}(H^{1/2}(\Gamma_R)^n;\, H^{-1/2}(\Gamma_R)^n)} = O(\rho^{n+1}). \tag{4.2}$$

Then defining δ_a by

$$\delta_a(u,v) = \int_{\Gamma_R} \delta_T u.v \, d\gamma(x), \quad u,\, v \in \mathcal{V}_R,$$

will yield straightforwardly

$$\|a_\rho - a_0 - \rho^n \delta_a\|_{\mathcal{L}_2(\mathcal{V})} = O(\rho^{n+1}).$$

In order to derive (4.2), first we need some definitions and preliminary lemmas.

4.1. Definitions. For convenience, the following norms and seminorms are chosen for the functional spaces which will be used.

- For a bounded, connected, and open subset $\mathcal{O} \subset \mathbb{R}^n$ with a Lipschitz continuous boundary, the Sobolev space $H^1(\mathcal{O})^n$ is equipped with the norm

  $$\|u\|^2_{1,\mathcal{O}} := \int_{\mathcal{O}} \sigma(u) : \varepsilon(u) + u.v \, dx,$$

  which, due to Korn's inequality [20], is equivalent to the usual norm. We will also need the seminorm defined by

  $$|u|^2_{1,\mathcal{O}} := \int_{\mathcal{O}} \sigma(u) : \varepsilon(u) \, dx. \tag{4.3}$$

  This seminorm is equivalent to the previous one on the subspace orthogonal to the constants $\{u \in H^1(\mathcal{O})^n;\ \int_{\mathcal{O}} u \, dx = 0\}$.

- For a given $\rho > 0$, the fractional Sobolev space $H^{1/2}(\Gamma_{R/\rho})^n$ is equipped with the following norm which is equivalent to the usual norm:

  $$\|v\|_{1/2,\Gamma_{R/\rho}} = \inf\left\{\|u\|_{1,C(R/(2\rho),R/\rho)}; \quad u = v \quad \text{on} \quad \Gamma_{R/\rho}\right\},$$

  where $C(r, r') := \{x \in \mathbb{R}^n; \quad r < ||x|| < r'\}$. We will also need the seminorm

  $$|v|_{1/2,\Gamma_{R/\rho}} = \inf\{|u|_{1,C(R/(2\rho),R/\rho)}\,; \quad u = v \quad \text{on} \quad \Gamma_{R/\rho}\}. \tag{4.4}$$

  The introduction of this seminorm is related to the fact that the null space of T_ρ consists in the constant functions.
- The dual space $H^{-1/2}(\Gamma_{R/\rho})^n$ is equipped with the natural norm

  $$\|w\|_{-1/2,\Gamma_{R/\rho}} = \sup\{\langle w, v\rangle_{-1/2,1/2}\,; \quad v \in H^{1/2}(\Gamma_{R/\rho})^n,\ \|v\|_{1/2,\Gamma_{R/\rho}} = 1\}.$$

  It can easily be checked that if $\psi \in H^1(C(R/2, R))^n$ with $\operatorname{div}\sigma(\psi) = 0$ in $C(R/2, R)$, then

  $$\|\sigma(\psi)\mathbf{n}\|_{-1/2,\Gamma_R} \leq c\,|\psi|_{1,C(R/2,R)}\,. \tag{4.5}$$

Here and in what follows, c is a positive constant independent of the data (e.g., of ρ).

4.2. Preliminary lemmas. Recall that $x_0 = 0$. We will use intensively the following change of variable: for a given function u defined on a set $\mathcal{O}$, function $\widetilde{u}$ is defined on $\widetilde{\mathcal{O}} := \mathcal{O}/\rho$ by

$$\widetilde{u}(y) = u(x), \quad y = x/\rho.$$

Unless otherwise specified, the derivation in the operators σ and ε will be considered with respect to the current variable; that is, $\varepsilon_{ij}(u)(x) = (\partial u_i(x)/\partial x_j + \partial u_j(x)/\partial x_i)/2$, $\varepsilon_{ij}(\widetilde{u})(y) = (\partial\widetilde{u}_i(y)/\partial y_j + \partial\widetilde{u}_j(y)/\partial y_i)/2$, etc. Due to $Du(x) = D\widetilde{u}(y)/\rho$ and to (4.3), we have

$$|u|^2_{1,\mathcal{O}} = \int_{\mathcal{O}} \sigma(u) : \varepsilon(u)\,dx = \frac{1}{\rho^2}\int_{\widetilde{\mathcal{O}}} \sigma(\widetilde{u}) : \varepsilon(\widetilde{u})\,\rho^n dy.$$

Hence

$$|u|_{1,\mathcal{O}} = \rho^{(n-2)/2}\,|\widetilde{u}|_{1,\widetilde{\mathcal{O}}}\,. \tag{4.6}$$

Similarly, we have

$$\begin{aligned}|v|_{1/2,\Gamma_R} &= \inf\{|u|_{1,C(R/2,R)}\,; \quad u = v \quad \text{on} \quad \Gamma_R\}\\ &= \inf\{\rho^{(n-2)/2}\,|\widetilde{u}|_{1,C(R/(2\rho),R/\rho)}\,; \quad \widetilde{u} = \widetilde{v} \quad \text{on} \quad \Gamma_{R/\rho}\}.\end{aligned}$$

Hence

$$|v|_{1/2,\Gamma_R} = \rho^{(n-2)/2}\,|\widetilde{v}|_{1/2,\Gamma_{R/\rho}}\,. \tag{4.7}$$

LEMMA 4.1. *For $\varphi \in H^{-1/2}(\partial\omega)^n$ such that*

$$\int_{\partial\omega} \varphi\,d\gamma(x) = 0,$$

let v be the solution to the problem

$$\left\{\begin{array}{rcl} -\operatorname{div}\sigma(v) & = & 0 \quad \text{in } \mathbb{R}^n\backslash\overline{\omega}, \\ v & = & 0 \quad \text{at } \infty, \\ \sigma(v)\mathbf{n} & = & \varphi \quad \text{on } \partial\omega. \end{array}\right. \tag{4.8}$$

The function v is split into

$$v(y) = V(y) + W(y),$$
$$V(y) = -\int_{\partial\omega} (DE(y)x)p(x)\, d\gamma(x),$$

where E is defined in (3.11)*. There exists $c > 0$ such that*

$$|V|_{1,C(R/(2\rho),R/\rho)} \leq c\rho^{n/2}\|\varphi\|_{-1/2,\partial\omega}, \tag{4.9}$$
$$|W|_{1,C(R/(2\rho),R/\rho)} \leq c\rho^{n/2+1}\|\varphi\|_{-1/2,\partial\omega}.$$

Proof. With the help of a single layer potential representation, the function v reads

$$v(y) = \int_{\partial\omega} E(y-x)p(x)\, d\gamma(x), \quad y \in \mathbb{R}^n\backslash\overline{\omega},$$

where (see also (3.13))

$$\frac{p(y)}{2} + \int_{\partial\omega} \sigma_y(E(y-x)p(x))\mathbf{n}(y)\, d\gamma(x) = -\varphi(y) \quad \forall y \in \partial\omega, \tag{4.10}$$

$$\int_{\partial\omega} p(x)\, d\gamma(x) = 0. \tag{4.11}$$

Using a Taylor expansion of E computed at the point y, (4.11), and the well-posedness of (4.10) [8], we have

$$||DV(y)|| \leq \frac{c}{|y|^n}\|\varphi\|_{-1/2,\partial\omega}, \quad ||DW(y)|| \leq \frac{c}{|y|^{n+1}}\|\varphi\|_{-1/2,\partial\omega},$$

from which (4.9) follows straightforwardly. □

LEMMA 4.2. *For $\varphi \in H^{1/2}(\Gamma_R)^n$, let v_ρ be the solution to the problem*

$$\left\{\begin{array}{rcl} -\operatorname{div}\sigma(v_\rho) & = & 0 \quad \text{in } D_\rho, \\ v_\rho & = & \varphi \quad \text{on } \Gamma_R, \\ \sigma(v_\rho)\mathbf{n} & = & 0 \quad \text{on } \partial\omega_\rho. \end{array}\right. \tag{4.12}$$

There exist a constant $c > 0$ (independent of φ and ρ) and $\rho_1 > 0$ such that for all $0 < \rho < \rho_1$,

$$|v_\rho|_{1,D_\rho} \leq c\,|\varphi|_{1/2,\Gamma_R}.$$

Proof. Let w solve the problem

$$\left\{\begin{array}{rcl} -\operatorname{div}\sigma(w) & = & 0 \quad \text{in } C(R/2,R), \\ w & = & \varphi \quad \text{on } \Gamma_R, \\ \sigma(w)\mathbf{n} & = & 0 \quad \text{on } \Gamma_{R/2}, \end{array}\right.$$

and let $w = w_0 + \gamma$, where the constant γ is such that $\int_{C(R/2,R)} w_0\,dx = 0$. For $\rho = 0$, (4.12) is well-posed, and there exists a constant c such that

$$|v_0|_{1,D_0} = |v_0 - \gamma|_{1,D_0} \le c\,\|\varphi - \gamma\|_{1/2,\Gamma_R} \le c\,\|w_0\|_{1/2,\partial C(R/2,R)} \le c\,\|w_0\|_{1,C(R/2,R)}\,,$$

where the last inequality follows from the trace theorem. On the subspace of $H^1(C(R/2,R))$ consisting in the functions u such that $\int_{C(R/2,R)} u\,dx = 0$, the semi-norm $|u|_1$ is equivalent to the norm $\|u\|_1$. Hence, using (4.4), we deduce that

$$|v_0|_{1,D_0} \le c\,|w_0|_{1,C(R/2,R)} = c\,|w|_{1,C(R/2,R)} = c\,|\varphi|_{1/2,\Gamma_R}\,.$$

Then let $\rho_1 > 0$ be such that $\omega_\rho \subset D_0$ for all $\rho < \rho_1$. The function v_ρ minimizes $|v|_{1,D_\rho}$ over the affine space $\left\{v \in H^1(D_\rho)^n;\ v = \varphi \text{ on } \Gamma_R\right\}$. Hence, if $\widehat{v_0}$ denotes the restriction of v_0 to D_ρ, we have

$$|v_\rho|_{1,D_\rho} \le |\widehat{v_0}|_{1,D_\rho} \le |v_0|_{1,D_0} \le c\,|\varphi|_{1/2,\Gamma_R}\,. \qquad \square$$

LEMMA 4.3. *For $\rho > 0$ and $\psi \in H^1(D_0)^n$ such that*

$$\operatorname{div}\sigma(\psi) = 0 \quad \text{in } D_0,$$

let u_ρ be the solution to the problem

$$\left\{\begin{array}{rcll} -\operatorname{div}\sigma(u_\rho) & = & 0 & \text{in } D_\rho, \\ u_\rho & = & 0 & \text{on } \Gamma_R, \\ \sigma(u_\rho)\mathbf{n} & = & \sigma(\psi)\mathbf{n} & \text{on } \partial\omega_\rho. \end{array}\right. \tag{4.13}$$

There exist a constant $c > 0$ (independent of ψ and ρ) and $\rho_1 > 0$ such that for all $0 < \rho < \rho_1$,

$$|u_\rho|_{1,C(R/2,R)} \le c\rho^n \|\sigma(\psi)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega}.$$

Proof. First we note that

$$\int_{\partial\omega} \sigma(\psi)(\rho x)\mathbf{n}\,d\gamma(x) = \frac{1}{\rho^{n-1}} \int_{\partial\omega_\rho} \sigma(\psi)(x)\mathbf{n}\,d\gamma(x) = \frac{1}{\rho^{n-1}} \int_{\omega_\rho} \operatorname{div}\sigma(\psi)\,d\gamma(x) = 0.$$

Hence we can define the solution $\widetilde{v_\rho}$ to the problem

$$\left\{\begin{array}{rcll} -\operatorname{div}\sigma(\widetilde{v_\rho}) & = & 0 & \text{in } \mathbb{R}^n\backslash\overline{\omega}, \\ \widetilde{v_\rho} & = & 0 & \text{at } \infty, \\ \sigma(\widetilde{v_\rho})\mathbf{n} & = & \rho\sigma(\psi)(\rho y)\mathbf{n} & \text{on } \partial\omega. \end{array}\right.$$

The function u_ρ can be written

$$u_\rho = v_\rho - w_\rho,$$

where $v_\rho(x) = \widetilde{v_\rho}(x/\rho)$. The function w_ρ itself is the solution to

$$\left\{\begin{array}{rcll} -\operatorname{div}\sigma(w_\rho) & = & 0 & \text{in } D_\rho, \\ w_\rho & = & v_\rho & \text{on } \Gamma_R, \\ \sigma(w_\rho)\mathbf{n} & = & 0 & \text{on } \partial\omega_\rho. \end{array}\right.$$

It follows from Lemma 4.2 that there exist $c > 0$ and $\rho_1 > 0$ such that for all $0 < \rho < \rho_1$,

$$|w_\rho|_{1,D_\rho} \leq c\,|v_\rho|_{1/2,\Gamma_R}\,. \tag{4.14}$$

Using (4.7), (4.4), and Lemma 4.1, we have

$$\begin{aligned} |v_\rho|_{1/2,\Gamma_R} &= \rho^{(n-2)/2}\,|\widetilde{v_\rho}|_{1/2,\Gamma_{R/\rho}} \leq \rho^{(n-2)/2}\,|\widetilde{v_\rho}|_{1,C(R/(2\rho),R/\rho)} \\ &\leq \rho^{(n-2)/2} c\rho^{n/2}\|\rho\sigma(\psi)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega} = c\rho^n\|\sigma(\psi)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega}. \end{aligned} \tag{4.15}$$

Again using Lemma 4.1, we also have

$$\begin{aligned} |v_\rho|_{1,C(R/2,R)} &= \rho^{(n-2)/2}\,|\widetilde{v_\rho}|_{1,C(R/(2\rho),R/\rho)} \\ &\leq c\rho^n\|\sigma(\psi)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega}. \end{aligned} \tag{4.16}$$

Hence, gathering together (4.14), (4.15), and (4.16), we obtain

$$\begin{aligned} |u_\rho|_{1,C(R/2,R)} &= |v_\rho - w_\rho|_{1,C(R/2,R)} \leq |v_\rho|_{1,C(R/2,R)} + |w_\rho|_{1,D_\rho} \\ &\leq c\rho^n\|\sigma(\psi)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega}. \qquad \square \end{aligned}$$

Lemmas 4.2 and 4.3 are summarized in the following lemma.

Lemma 4.4. *Let v_ρ be the solution to the problem*

$$\left\{\begin{array}{rcll} -\operatorname{div}\sigma(v_\rho) &=& 0 & \text{in } D_\rho, \\ v_\rho &=& \varphi & \text{on } \Gamma_R, \\ \sigma(v_\rho)\mathbf{n} &=& \sigma(\psi)\mathbf{n} & \text{on } \partial\omega_\rho, \end{array}\right.$$

where $\varphi \in H^{1/2}(\Gamma_R)^n$ and $\psi \in H^1(D_0)^n$ with

$$\operatorname{div}\sigma(\psi) = 0 \quad \text{in } D_0.$$

There exist a constant $c > 0$ (independent of φ, ψ and ρ) and $\rho_1 > 0$ such that for all $0 < \rho < \rho_1$,

$$|v_\rho|_{1,C(R/2,R)} \leq c\,|\varphi|_{1/2,\Gamma_R} + c\rho^n\|\sigma(\psi)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega}.$$

4.3. Variation of the boundary operator. The linear operator δ_T (independent of ρ) is defined as follows:

$$\begin{array}{rcl} \delta_T : H^{1/2}(\Gamma_R)^n & \longrightarrow & H^{-1/2}(\Gamma_R)^n, \\ \varphi & \longmapsto & \delta_T\varphi := \sigma(Q_w - P_\omega)\mathbf{n}, \end{array} \tag{4.17}$$

where P_ω and Q_w are defined by (3.15) and (3.18), with $\sigma(u_0^\varphi)(0)$ substituted for $\sigma(u_\Omega)(0)$ in (3.10), the function u_0^φ being the solution to (3.4). (We are interested in $\delta_T u_0$, and in that case there is no substitution to do.)

Proposition 4.5. *The asymptotic expansion of T_ρ is*

$$\|T_\rho - T_0 - \rho^n\delta_T\|_{\mathcal{L}(H^{1/2}(\Gamma_R)^n;\,H^{-1/2}(\Gamma_R)^n)} = O(\rho^{n+1}). \tag{4.18}$$

Proof. Let $\varphi \in H^{1/2}(\Gamma_R)^n$, and denote $u_\rho = u_\rho^\varphi$, $u_0 = u_0^\varphi$ as the solutions to (3.4) and (3.5). The function v_ω is defined by (3.10) (with $\sigma(u_0)(0)$ substituted for

$\sigma(u_\Omega)(0)$). For $y = x/\rho$, recall ((3.14)–(3.16)) that $v_\omega(y) = P_\omega(y) + W_\omega(y)$ with $P_\omega(x/\rho) = \rho^{n-1}P_\omega(x)$ and $W_\omega(y) = O(1/||y||^n)$. Let

$$\psi_\rho(x) = (T_\rho - T_0 - \rho^n \delta_T)\varphi(x).$$

We have

$$\begin{aligned} \psi_\rho(x) &= \sigma(u_\rho(x) - u_0(x) + \rho^n P_\omega(x) - \rho^n Q_\omega(x))\mathbf{n} \\ &= \sigma(w_\rho(x) - \rho W_\omega(x/\rho))\mathbf{n}, \end{aligned}$$

where w_ρ is defined by

$$w_\rho(x) = u_\rho(x) - u_0(x) + \rho v_\omega(x/\rho) - \rho^n Q_\omega(x).$$

The function w_ρ is the solution to

$$\left\{ \begin{array}{rcll} -\operatorname{div}\sigma(w_\rho) &=& 0 & \text{in } D_\rho, \\ w_\rho &=& \rho v_\omega(x/\rho) - \rho^n Q_\omega(x) & \text{on } \Gamma_R, \\ \sigma(w_\rho)\mathbf{n} &=& \sigma(-u_0(x) + \rho v_\omega(x/\rho) - \rho^n Q_\omega(x))\mathbf{n} & \text{on } \partial\omega_\rho. \end{array} \right.$$

In order to apply Lemma 4.4, we have to estimate the two right-hand sides.

On Γ_R, due to (3.18), (3.16), and (3.14), we have

$$\rho v_\omega(x/\rho) - \rho^n Q_\omega(x) = \rho W_\omega(x/\rho).$$

Using (4.7), (4.4), Lemma 4.1, and the elliptic regularity of problem (3.4), we have

$$\begin{aligned} |\rho v_\omega(x/\rho) - \rho^n Q_\omega(x)|_{1/2,\Gamma_R} &= |\rho W_\omega(x/\rho)|_{1/2,\Gamma_R} \\ &= \rho^{n/2}\,|W_\omega(y)|_{1/2,\Gamma_{R/\rho}} \\ &\le \rho^{n/2}\,|W_\omega|_{1,C(R/(2\rho),R/\rho)} \\ &\le c\rho^{n+1}\|\sigma(u_0)(0)\mathbf{n}\|_{-1/2,\partial\omega} \\ &\le c\rho^{n+1}\|\varphi\|_{1/2,\Gamma_R}. \end{aligned} \tag{4.19}$$

On $\partial\omega_\rho$, due to the definition (3.10) of v_ω, we have

$$\sigma_x(\rho v_\omega(x/\rho))\mathbf{n} = \sigma_y(v_\omega(y))\mathbf{n} = \sigma_x(u_0)(0)\mathbf{n},$$

where the subscript in σ denotes the differentiated variable. Hence, using $\sigma(\varepsilon(u_0)(0)x) = \sigma(u_0)(0)$ for all x, we have on $\partial\omega_\rho$

$$\begin{aligned} \sigma(w_\rho)(x)\mathbf{n} &= \sigma(-u_0(x) + \rho v_\omega(x/\rho) - \rho^n Q_\omega(x))\mathbf{n} \\ &= \sigma(-u_0(x) + \varepsilon(u_0)(0)x - \rho^n Q_\omega(x))\mathbf{n} \\ &= \rho\sigma(\theta_\rho)\mathbf{n}, \end{aligned}$$

where the function $\theta_\rho := (-u_0(x) + \varepsilon(u_0)(0)x - \rho^n Q_\omega(x))/\rho$ satisfies $\operatorname{div}\sigma(\theta_\rho) = 0$ in D_0. Applying Lemma 4.4 and using (4.19), we obtain

$$\begin{aligned} |w_\rho|_{1,C(R/2,R)} &\le c\,|\rho v_\omega(x/\rho) - \rho^n Q_\omega(x)|_{1/2,\Gamma_R} + c\rho^n\|\rho\sigma(\theta_\rho)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega} \\ &\le c\rho^{n+1}(\|\varphi\|_{1/2,\Gamma_R} + \|\sigma(\theta_\rho)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega}). \end{aligned}$$

The analyticity of $\sigma(\theta_\rho)(\rho y) = (-\sigma(u_0)(\rho y) + \sigma(u_0)(0))/\rho - \rho^{n-1}\varepsilon(Q_\omega)(\rho y)$ (u_0 and Q_ω are analytical in D_0) and the elliptic regularity of problems (3.4) and (3.18) [15] yield for small ρ

$$\begin{aligned}\|\sigma(\theta_\rho)(\rho y)\mathbf{n}\|_{-1/2,\partial\omega} &\le c\|\sigma(\theta_\rho)(\rho y)\mathbf{n}\|_{L^\infty(\partial\omega)}\\ &\le c\|u_0\|_{W^{2,\infty}(B(0,R/2))} + c\|Q_\omega\|_{W^{1,\infty}(B(0,R/2))}\\ &\le c\|\varphi\|_{1/2,\Gamma_R}.\end{aligned}$$

Hence

$$|w_\rho|_{1,C(R/2,R)} \le c\rho^{n+1}\|\varphi\|_{1/2,\Gamma_R}. \tag{4.20}$$

We have $\operatorname{div}\sigma(w_\rho) = 0$, and $\operatorname{div}\sigma(W_\omega(x/\rho)) = 0$ follows from (3.17); thus using (4.5) and (4.6) yields

$$\begin{aligned}\|\psi_\rho\|_{-1/2,\Gamma_R} &= \|\sigma(w_\rho(x) - \rho W_\omega(x/\rho))\mathbf{n}\|_{-1/2,\Gamma_R}\\ &\le |w_\rho|_{1,C(R/2,R)} + |\rho W_\omega(x/\rho)|_{1,C(R/2,R)}\\ &\le |w_\rho|_{1,C(R/2,R)} + \rho^{n/2}\,|W_\omega|_{1,C(R/(2\rho),R/\rho)}\,,\end{aligned}$$

and we conclude by using (4.20) and (4.19) once more. □

4.4. Variation of the bilinear form. The asymptotic expansion of the bilinear form a_ρ is a direct consequence of Proposition 4.5.

PROPOSITION 4.6. *Let*

$$\delta_a(u,v) = \int_{\Gamma_R} \delta_T u.v \; d\gamma(x), \quad u,\, v \in \mathcal{V}_R.$$

The asymptotic expansion of the bilinear form a_ρ with respect to ρ is

$$\|a_\rho - a_0 - \rho^n\delta_a\|_{\mathcal{L}_2(\mathcal{V})} = O(\rho^{n+1}).$$

Hence the fundamental assumption (2.2) is satisfied, and it follows from Theorem 2.2 that

$$\begin{aligned}j(\rho) &= j(0) + \rho^n g(x_0) + o(\rho^n),\\ g(x_0) &= \delta_a(u_0,v_0) + \delta_J(u_0),\end{aligned} \tag{4.21}$$

which achieves the proof of Theorem 3.2.

5. Numerical results. According to (4.21), the topological sensitivity gives an information on the opportunity of creating a small hole around x_0. Suppose that the function j has to be minimized. Then creating a hole where $g(x) < 0$ may decrease the function j. Following the presentation of Céa, Gioan, and Michel in [5], (4.21) leads to the following optimality condition:

$$g(x) \ge 0 \quad \forall x \in \ \Omega.$$

This condition may be used in the following way to derive a topology optimization algorithm.

5.1. Algorithm (see [5]). Consider an initial domain Ω_0 which represents the design domain. The optimal domain is sought in $\{\Omega' \subset \Omega_0;\ \Omega' \text{ is open}\}$. Let $(m_k)_{k\geq 0}$ be a decreasing sequence of volume constraints with $m_0 = \text{meas}(\Omega_0)$. For example, a geometrical sequence may be chosen. At the kth iteration, the topological sensitivity is denoted by $g_k(x)$, and c_{k+1} is chosen in such a way that

$$\begin{cases} \Omega_{k+1} = \{x \in \Omega_k,\ g_k(x) \geq c_{k+1}\}, \\ \text{meas}(\Omega_{k+1}) = m_{k+1}. \end{cases}$$

The process stops when a target such as a volume constraint (or a constraint on the maximum of the Von Mises stress) is reached. Hence the algorithm is the following.

```
ALGORITHM: TOPOLOGY OPTIMIZATION WITH VOLUME CONSTRAINT.
• Initialization: chose the initial domain $\Omega_0$, and set $k = 0$.
• Repeat
    1. solve the linear elasticity problem in $\Omega_k$,
    2. compute the topological sensitivity $g_k$,
    3. set $\Omega_{k+1} = \{x \in \Omega_k,\ g_k(x) \geq c_{k+1}\}$, where $c_{k+1}$ is chosen such that
       $\text{meas}(\Omega_{k+1}) = m_{k+1}$,
    4. $k \leftarrow k + 1$.
• until target is reached.
```

The topological sensitivity is computed on each element. Then the elements are sorted with respect to this sensitivity. The lowest elements are removed. The number of elements removed at each step is given by the volume ratio (volume of elements removed) / (volume of the previous structure). In the following examples, this volume ratio is taken between 5% and 10%. At each iteration, most of the computational time is required for solving the elasticity problem. As only a few iterations are needed, this method is not expensive in terms of computational cost.

5.2. Applications. The following examples are taken from [12].

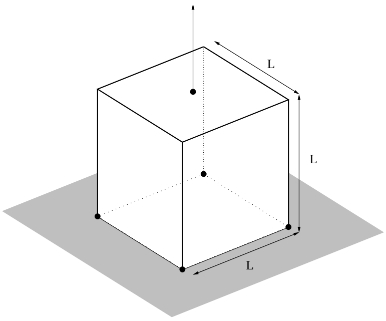


FIG. 5.1. *The design domain and boundary conditions for the first example.*

In the first example, the initial design domain Ω_0 is a cube (see Figure 5.1). The four vertices of the bottom face can slide in the horizontal plan, and the vertical displacement is zero at those latter points. A load is applied on the center of the top face. Fifty iterations were computed, and 9% of the material was removed at each step. Figure 5.2 shows several intermediate designs obtained during the optimization process. The mesh was refined after 25 iterations.

When the four bottom corners are clamped (homogeneous Dirichlet condition), the horizontal lattices between the four supports disappear, and the optimal design consists in four rods joined in a pyramidal structure.

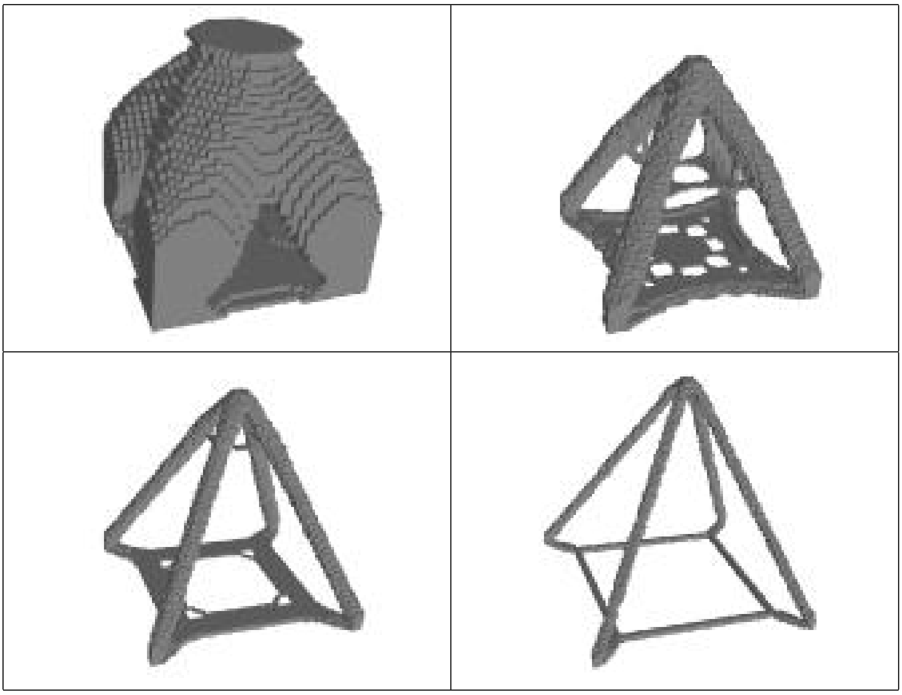


FIG. 5.2. *The pyramids obtained after* 8, 28, 37, *and* 50 *iterations. Volumes represent, respectively,* 50%, 7%, 3%, *and* 1% *of the initial volume.*

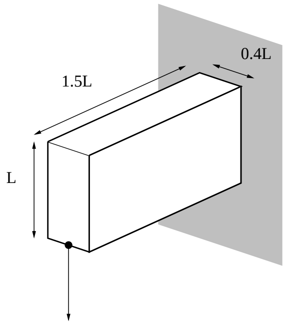


FIG. 5.3. *The design domain and boundary conditions for the slender cantilever beam.*

The second example is a slender cantilever beam (see Figure 5.3). The initial domain Ω_0 is a thin hexahedron with one face clamped, i.e., a homogeneous Dirichlet boundary condition is imposed on the whole face, and a pointwise load is applied at the lower edge of the opposite face. In that case, the minimization of the maximum of the Von Mises stress is also considered. Here 7% of the material is removed at each step. The optimal domains (see Figure 5.4) are obtained after 20 iterations. The results shown on Figure 5.4 recalls the 2D case (see, e.g., [2]); in fact, one dimension is very small with respect to the two others. One can observe that the structures obtained by minimization of the compliance and minimization of the maximum Von Mises stress are slightly different.

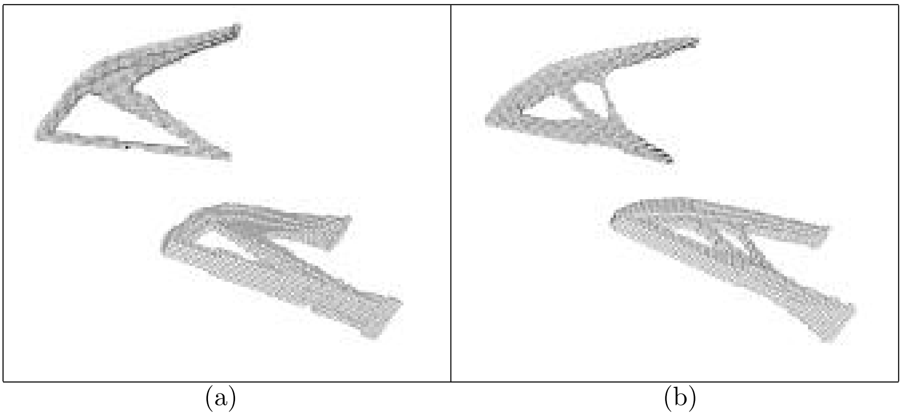


FIG. 5.4. *The slender cantilever beam for a volume equal to* 20% *of the initial volume, after* 20 *iterations. The compliance case is shown in* (a)*, and the Von Mises case is shown in* (b)*.*

6. Conclusion. Using the mathematical framework presented in this paper, it is possible to write an asymptotic expansion of a general functional with respect to the creation of a small hole. This approach is general and can be adapted to various equations and boundary conditions. Moreover, as the topological sensitivity involves only the direct solution and possibly the adjoint state, its computation is cheap, and efficient topology optimization algorithms can be implemented.

Acknowledgment. The authors are grateful to the referees for their thorough reading and their suggestions concerning the presentation.

REFERENCES

[1] G. Allaire and R. Kohn, *Optimal bounds on the effective behavior of a mixture of two well-ordered elastic materials*, Quart. Appl. Math., 51 (1994), pp. 643–674.

[2] M. Becker, *Optimisation topologique de structure en variables discrètes*, Université de Lièges, Lieǵes, Belgium, 1996.

[3] M. Bendsøe, *Optimal Topology Design of Continuum Structure: An Introduction*, Technical report, Department of Mathematics, Technical University of Denmark, DK2800 Lyngby, Denmark, 1996.

[4] J. Cea, *Conception optimale ou identification de forme. Calcul rapide de la dérivée directionnelle de la fonction coût*, RAIRO Modél. Math. Anal. Numèr., 20 (1986), pp. 371–402.

[5] J. Cea, A. Gioan, and J. Michel, *Quelques résultats sur l'identification de domains*, CL, 10 (1973), pp. 207–232.

[6] J. Cea, S. Garreau, Ph. Guillaume, and M. Masmoudi, *The shape and topological optimizations connection*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 713–726.

[7] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[8] R. Dautray and J. L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Commisoriat à l'Energie Atomique, Masson, Paris, 1987.

[9] J. Giroire, *Formulations variationnelles par équations intégrales de problèmes aux limites extérieurs*, Thesis, Ecole Polytechnique, Palaiseau, France, 1976.

[10] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological sensitivity for linear isotropic elasticity*, in Proceedings of the European Conference on Computationnal Mechanics, The German Association for Computational Mechanics, Munich, Germany, 1999, report MIP 99.45.

[11] P. Guillaume and M. Masmoudi, *Computation of high order derivatives in optimal shape design*, Numer. Math., 67 (1994), pp. 231–250.

[12] J. Jacobsen, N. Oolhoff, and E. Rønholt, *Generalized Shape Optimization of Three-Dimensional Structures Using Materials with Optimum Microstructures*, Technical report,

Institute of Mechanical Engineering, Aalborg University, DK-9920 Aalborg, Denmark, 1996.

[13] V. D. Kupradze, T. G. Gregelia, M. O. Basheleuishvili, and T. V. Burchauladze, *Three Dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity*, North-Holland Ser. Appl. Math. Mech. 25, North-Holland, Amsterdam, 1979.

[14] T. Lewinski and J. Sokolowski, *Topological Derivative for Nucleation of Non-circular Voids*, Technical report RR-3798, INRIA, Lorraine, France, 1999.

[15] J. L. Lions and E. Magenes, *Problèmes aux limites non homogenes et applications*, Dunod, Paris, 1968.

[16] M. Masmoudi, *Outils pour la conception optimale de formes*, Thèse d'état, Université de Nice, Nice, France 1987.

[17] M. Masmoudi, *The topological asymptotic*, in Computational Methods for Control Applications, H. Kawarada and J. Periaux, eds., GAKUTO Internat. Ser. Math. Sci. Appl. Gakkōtosho, Tokyo, to appear.

[18] F. Murat and J. Simon, *Sur le contrôle par un domaine géométrique*, Thése d'état, Université Pierre et Marie Curie, Paris, 1976.

[19] F. Murat and L. Tartar, *Calcul des variations et homogénéisation*, in Les Méthodes de l'Homogénéisation: Théorie et Applications en Physique, Eyrolles, Paris, 1985, pp. 319–369.

[20] J. Necas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[21] A. Schumacher, *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lopchpositionierungkrieterien*, Thesis, Universität-Gesamthochschule Siegen, Siegen, Germany, 1995.

[22] M. Shoenauer, L. Kallel, and F. Jouve, *Mechanical inclusions identification by evolutionary computation*, Rev. Européenne Élém. Finis, 5 (1996), pp. 619–648.

[23] J. Sokolowski and A. Żochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.

[24] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 267–278.

[25] D. J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of Conductivity Imperfections of Small Diameter by Boundary Measurements, Continuous Dependence and Computational Reconstruction*, Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN, 1997.

SIAM J. Control Optim.
Vol. 39, No. 6, pp. 1779–1816

OPTIMAL CONTROL PROBLEMS FOR STOCHASTIC REACTION-DIFFUSION SYSTEMS WITH NON-LIPSCHITZ COEFFICIENTS[*]

SANDRA CERRAI[†]

Abstract. By using the dynamic programming approach, we study a control problem for a class of stochastic reaction-diffusion systems with coefficients having polynomial growth. In the cost functional a non-Lipschitz term appears, and this allows us to treat the quadratic case, which is of interest in the applications. The corresponding Hamilton–Jacobi–Bellman equation is first resolved by a fixed point argument in a small time interval and then is extended to arbitrary time intervals by suitable a priori estimates. The main ingredient in the proof is the smoothing effect of the transition semigroup associated with the uncontrolled system.

Key words. stochastic reaction-diffusion systems, Hamilton–Jacobi–Bellman equations in infinite dimension, stochastic optimal control problems

AMS subject classifications. 60H15, 69J35, 93C20, 93E20

PII. S0363012999356465

1. Introduction.

In this paper we consider the following class of stochastic reaction-diffusion systems with distributed parameter controls in bounded domains $\mathcal{O}$ of $\mathbb{R}^d$, with $d \leq 3$:

$$\tag{1.1} \left\{ \begin{array}{l} \dfrac{\partial y_k}{\partial s}(s,\xi) = \mathcal{A}_k\, y_k(s,\xi) + f_k(\xi, y_1(s,\xi),\dots,y_r(s,\xi)) + z_k(s,\xi) + Q_k \dfrac{\partial^2 w_k}{\partial s \partial \xi}(s,\xi), \\[2ex] y_k(t,\xi) = x_k(\xi), \quad 0 \leq t < s \leq T, \quad \xi \in \overline{\mathcal{O}}, \\[2ex] \mathcal{B}_k\, y_k(s,\xi) = 0, \qquad \xi \in \partial\mathcal{O}, \qquad k = 1,\dots,r. \end{array} \right.$$

Here $\mathcal{A} = (\mathcal{A}_1,\dots,\mathcal{A}_r)$ is a uniformly elliptic second order differential operator with regular real coefficients, and $\mathcal{B} = (\mathcal{B}_1,\dots,\mathcal{B}_r)$ is a first order differential operator acting on the boundary of $\mathcal{O}$. The reaction term $f = (f_1,\dots,f_r) : \overline{\mathcal{O}} \times \mathbb{R}^r \to \mathbb{R}^r$ is continuous, and $f(\xi,\cdot) : \mathbb{R}^r \to \mathbb{R}^r$ is twice differentiable, has polynomial growth together with its derivatives, and fulfills appropriate dissipativity conditions. $Q = (Q_1,\dots,Q_r)$ is a nonnegative bounded linear operator from $H = L^2(\mathcal{O};\mathbb{R}^r)$ into itself, and $\partial^2 w_k/\partial t\,\partial\xi$ are independent space-time white noises defined on a stochastic basis $(\Omega,\mathcal{F},\mathcal{F}_t,\mathbf{P})$. The control $z = (z_1,\dots,z_r)$ is taken in the set of adapted processes of $L^2(\Omega;L^2(0,T;H))$. We remark that the dimension d is taken less than or equal to 3 because the noise should be, in a sense, nondegenerate, and the solution of the system (1.1) has to take value in $E = C(\mathcal{O};\mathbb{R}^r)$.

We are here concerned with the *cost functional*

$$\tag{1.2} J(t,x;z) = \mathbf{E}\,\varphi(y(T)) + \int_t^T \mathbf{E}\,(g(y(s)) + k(z(s)))\,ds,$$

[*]Received by the editors May 19, 1999; accepted for publication (in revised form) November 16, 2000; published electronically April 12, 2001.
http://www.siam.org/journals/sicon/39-6/35646.html

[†]Dipartimento di Matematica per le Decisioni, Università di Firenze, Via Lombroso 6/17, 50134 Firenze, Italy (cerrai@cibs.sns.it).

where $y(s) = y(s,t;x,z)$ is the solution of the problem (1.1), φ and g are bounded and Lipschitz continuous functions from H into $\mathbb{R}$, and $k : H \to]-\infty, +\infty]$ is a measurable function which fulfills suitable conditions. Our aim is to prove that the *value function* corresponding to the cost functional (1.2), which is defined by

$$V(t,x) = \inf \left\{ J(t,x;z)\,;\, z \in L^2(\Omega; L^2(0,T;H))\,\text{adapted} \right\},$$

satisfies the Hamilton–Jacobi–Bellman equation

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial t}(t,x) + \mathcal{L}u(t,x) - K(Du(t,x)) + g(x) = 0, \\[2ex] u(T,x) = \varphi(x), \end{array} \right. \tag{1.3}$$

where $\mathcal{L}$ is the differential operator

$$\mathcal{L}\psi(x) = \frac{1}{2}\mathrm{Tr}\,\left[Q^2 D^2 u(t,x)\right] + \langle \mathcal{A}x + f(\cdot;x), Du(t,x)\rangle_H$$

and K is the Legendre transform of k. Notice that the hamiltonian K is not assumed to be Lipschitz continuous so that we can cover the important case of quadratic hamiltonians. Moreover, it is important to stress that in the present paper we are only able to treat the case when data φ and g are Lipschitz continuous.

After proving in the first part that there exists a unique mild solution $u(t,x)$ for (1.3), we show that for any adapted control $z \in L^2(\Omega; L^2(0,T;H))$ and for any $x \in H$ and $t \in [0,T]$ the following identity holds:

$$\begin{aligned} &J(t,x;z) = u(t,x) \\ &+ \int_t^T \mathbf{E}\,\left[K(Du(s,y(s))) + \langle z(s), Du(s,y(s))\rangle_H + k(z(s))\right]\,ds. \end{aligned}$$

Thus, in particular, we have $V(t,x) \geq u(t,x)$. Now if we could prove the existence of a solution $y^\star(t)$ for the *closed loop* equation

$$dy(t) = (\mathcal{A}y(t) + f(\cdot;y(t)) - DK(Du(t,y(t))))\,dt + Q\,dw(t), \qquad y(0) = x, \tag{1.4}$$

then $z^\star(t) = -DK(Du(t,y^\star(t)))$ would be an optimal control for the minimizing problem related to the functional (1.2). But unfortunately here we are only able to prove C^1 regularity for the solution of the Hamilton–Jacobi–Bellman equation (1.3), so that we cannot prove the existence of a solution for (1.4) which is adapted to the filtration $\mathcal{F}_t$. Actually, as the solution of (1.3) is only C^1, the *closed loop* equation (1.4) admits only martingale solutions, and hence there is no reason why the optimal control which we could get from it is adapted to the filtration we fixed at the beginning. Thus at present we restrict ourselves to the proof of the verification theorem. In the future it will be interesting to check if, by introducing the notion of *relaxed controls* (see [17] and [29] for the definition), it will be possible to prove the existence of an optimal control. However, in dimension $d = 1$ we are able to show that under some additional assumptions, the closed loop equation has a unique solution so that there exists a unique optimal control.

In order to prove the opposite inequality $V(t,x) \leq u(t,x)$, we introduce an approximating cost functional $J_\alpha(t,x;z)$, and we prove that it satisfies a verification theorem and admits a unique optimal control for each $\alpha > 0$. Due to suitable a priori estimates, we show that there exists a subset $\mathcal{M}_R$ of the space of adapted processes in $L^2(\Omega; L^2(0,T;H))$ such that for any $\alpha > 0$

$$V_\alpha(t,x) = \inf\left\{ J_\alpha(t,x;z)\ ;\ z \in \mathcal{M}^2_R(T) \right\}.$$

Moreover, we show that for any $x \in C(\overline{\mathcal{O}};\mathbb{R}^r)$ the functional $J_\alpha(t,x;z)$ converges to $J(t,x;z)$ as α goes to zero, uniformly for $z \in \mathcal{M}_R$, so that

$$\lim_{\alpha\to 0} V_\alpha(t,x) \geq V(t,x).$$

Thus, by showing that

$$\lim_{\alpha\to 0} V_\alpha(t,x) = u(t,x),$$

we have that $u(t,x) \geq V(t,x)$, and the verification theorem holds for $x \in C(\overline{\mathcal{O}};\mathbb{R}^r)$. The general case $x \in H$ follows by further approximation arguments.

Hamilton–Jacobi–Bellman equations in infinite dimensional spaces have been studied by several authors by using both semigroup techniques and the approach of viscosity solutions (see [3], [4], [12], [13], [20], [21], [23], [24], [25], and all references quoted therein). In particular, in [20] and [21] abstract semilinear stochastic problems are studied, and the nonlinear term f is assumed to be Lipschitz continuous. Instead, in the present paper we are able to skip the condition of Lipschitz continuity for f, and we can consider the case of reaction terms which have polynomial growth (and hence are not well defined in H).

In order to solve the problem (1.3), we introduce the transition semigroup P_t associated with the system (1.1) by setting for any bounded Borel function φ from H into $\mathbb{R}$ and for any $x \in H$

$$P_t\varphi(x) = \mathbf{E}\,\varphi(y(t;x)), \qquad t \geq 0,$$

where $y(t;x)$ is the solution of the uncontrolled system (1.1) starting from x at time zero. Due to Itô's formula and the variation of constants formula, we write (1.3) in the mild form

$$u(t,x) = P_t\varphi(x) - \int_0^t P_{t-s}K(Du(s,\cdot))(x)\,ds + \int_0^t P_{t-s}g(x)\,ds,$$

and by using a fixed point argument we show that for any $\varphi, g \in C^1_b(H)$ there exists a unique differentiable solution $u(t,x)$ which is defined only in a small time interval $[0,T_0]$, as K is only locally Lipschitz continuous. We want to emphasize that the crucial point in our argument is given by the smoothing effect of the semigroup P_t. Actually, P_t maps the space of bounded Borel functions defined on H into the space of differentiable functions, and the estimate

$$\sup_{x\in H} |D(P_t\varphi)(x)| \leq c(t\wedge 1)^{-\frac{1+\epsilon}{2}} \sup_{x\in H} |\varphi(x)|$$

holds for some constant $\epsilon < 1$ depending on the dimension $d \leq 3$ (see [7] for the proof). In order to have a global solution we need to obtain some a priori estimates. To this purpose we first approximate the reaction term f by a Lipschitz continuous sequence $\{f_\alpha\}_{\alpha>0}$, and then we consider the approximating Hamilton–Jacobi–Bellman equation, with the nonlinear term f replaced by f_α. By a Galerkin argument we prove some a priori estimates for the corresponding solutions $u_\alpha(t,x)$, and, by taking the limit as α goes to zero, we get the good estimates for $u(t,x)$.

2. Notations and preliminary results. Let $\mathcal{O}$ be a bounded regular open set of $\mathbb{R}^d$, with $d \leq 3$, having a regular boundary. Here and in what follows we denote by H the Hilbert space $L^2(\mathcal{O};\mathbb{R}^r)$, endowed with the scalar product $\langle\cdot,\cdot\rangle_H$ and the norm $|\cdot|_H$. For any $p \geq 1$, $p \neq 2$, we denote by $|\cdot|_p$ the norm in $L^p(\mathcal{O};\mathbb{R}^r)$. Moreover, we denote by E the Banach space $C(\overline{\mathcal{O}};\mathbb{R}^r)$ endowed with the *sup-norm* and the duality pairing $\langle\cdot,\cdot\rangle_E$ in $E \times E^\star$.

If X and Y are two separable Banach spaces, $B_b(X;Y)$ is the Banach space of all bounded Borel functions $\varphi : X \to Y$ endowed with the *sup-norm*

$$\|\varphi\|_0^X = \sup_{x\in X} |\varphi(x)|_Y.$$

$C_b(X;Y)$ is the subspace of uniformly continuous functions. For any integer $k \geq 1$, we denote by $C_b^k(X;Y)$ the subspace of k-times Fréchet differentiable functions, having bounded and uniformly continuous derivatives, up to the kth order. If we set for any $j = 1,\dots,k$

$$[\varphi]_j^X = \sup_{x\in X} |D^j\varphi(x)|_{\mathcal{L}^j(X;Y)},$$

we have that $C_b^k(X;Y)$ is a Banach space endowed with the norm

$$\|\varphi\|_k^X = \|\varphi\|_0^X + \sum_{j=1}^k [\varphi]_j^X.$$

We denote by $\mathrm{Lip}_b(X;Y)$ the subspace of functions $\varphi \in C_b(X;Y)$ such that

$$[\varphi]_{\mathrm{Lip}}^X = \sup_{\substack{x,y\in X\\ x\neq y}} \frac{|\varphi(x)-\varphi(y)|_Y}{|x-y|_X} < \infty.$$

$\mathrm{Lip}_b(X;Y)$ is a Banach space endowed with the norm

$$\|\varphi\|_{\mathrm{Lip}}^X = \|\varphi\|_0^X + [\varphi]_{\mathrm{Lip}}^X.$$

When $Y = \mathbb{R}$, we denote $B_b(X;Y)$, $C_b(X;Y)$, $C_b^k(X;Y)$, and $\mathrm{Lip}_b(X;Y)$, respectively, by $B_b(X)$, $C_b(X)$, $C_b^k(X)$, and $\mathrm{Lip}_b(X)$.

2.1. The Nemytskii operator. We assume that for any $k = 1,\dots,r$ there exist two continuous functions $g_k : \overline{\mathcal{O}} \times \mathbb{R} \to \mathbb{R}$ and $h_k : \overline{\mathcal{O}} \times \mathbb{R}^r \to \mathbb{R}$ such that for any $\xi \in \overline{\mathcal{O}}$ and $\sigma = (\sigma_1,\dots,\sigma_r) \in \mathbb{R}^r$ it holds that

$$f_k(\xi,\sigma_1,\dots,\sigma_r) = g_k(\xi,\sigma_k) + h_k(\xi,\sigma_1,\dots,\sigma_r).$$

The functions g_k and h_k are assumed to enjoy the following conditions.

Hypothesis 1.

1. *For any $\xi \in \overline{\mathcal{O}}$ the function $h_k(\xi,\cdot)$ is of class C^2 and has bounded derivatives, uniformly with respect to $\xi \in \overline{\mathcal{O}}$. Moreover, the mapping $D^j_\sigma h_k : \overline{\mathcal{O}} \times \mathbb{R}^r \to \mathcal{L}^j(\mathbb{R}^r)$ is continuous for $j = 1, 2$.*
2. *For any $\xi \in \overline{\mathcal{O}}$, the function $g_k(\xi,\cdot)$ is of class C^2, and there exists $m \geq 0$ such that*
$$\sup_{\xi\in\overline{\mathcal{O}}} \sup_{t\in\mathbb{R}} \frac{|D^j_t g_k(\xi,t)|}{1+|t|^{2m+1-j}} < \infty.$$
Moreover, the mapping $D^j_t g_k : \overline{\mathcal{O}} \times \mathbb{R}^r \to \mathbb{R}$ is continuous for $j = 1, 2$.
3. *There exist $a > 0$ and $c \in \mathbb{R}$ such that*
$$\sup_{\xi\in\overline{\mathcal{O}}} D_t g_k(\xi,t) \leq -a\, t^{2m} + c, \qquad t \in \mathbb{R}. \tag{2.1}$$

The Nemytskii operator F associated with the function $(\xi,\sigma) \mapsto f(\xi,\sigma)$ is defined as
$$F(x)(\xi) = f(\xi, x(\xi)), \quad \xi \in \mathcal{O}.$$
If we denote
$$p_\star = 2m+2, \qquad q_\star = \frac{2m+2}{2m+1},$$
it is possible to show that if $m \geq 1$, then F is twice Fréchet differentiable from $L^{p_\star}(\mathcal{O};\mathbb{R}^r)$ into $L^{q_\star}(\mathcal{O};\mathbb{R}^r)$, and it holds that
$$|D^j F(x)|_{\mathcal{L}^j(L^{p_\star},\, L^{q_\star})} \leq c\left(1+|x|^{2m+1-j}_{p_\star}\right), \qquad x \in L^{p_\star}(\mathcal{O};\mathbb{R}^r).$$
From (2.1) we obtain that for any $x, h \in L^{p_\star}(\mathcal{O};\mathbb{R}^r)$
$$\langle DF(x)h, h\rangle_H \leq -a|x^m h|^2_H + c|h|^2_H,$$
and, in particular,
$$\langle DF(x)h, h\rangle_H \leq c\,|h|^2_H.$$
Moreover, from (2.1) it follows that for any $\sigma, \rho \in \mathbb{R}^r$
$$\sup_{\xi\in\overline{\mathcal{O}}} \langle f(\xi,\sigma+\rho) - f(\xi,\rho), \sigma\rangle_{\mathbb{R}^r} \leq -a|\sigma|^{2m+2} + c\left(1+|\rho|^{2m+1}\right)$$
for some constants $a > 0$ and $c \in \mathbb{R}$, possibly different from those introduced in (2.1). This implies that
$$\langle F(x+h) - F(x), h\rangle_H \leq -a\,|h|^{p^\star}_{p_\star} + c\left(1+|x|^{p_\star}_{p_\star}\right).$$

By using similar arguments, it is immediate to prove that $F : E \to E$ is twice differentiable, and
$$|D^j F(x)|_{\mathcal{L}^j(E)} \leq c\left(1+|x|^{2m+1-j}_E\right), \qquad x \in E. \tag{2.2}$$

For any $h, y \in E$ we define

$$\langle \delta_h, y \rangle_E = \begin{cases} \dfrac{1}{|h|_E} \displaystyle\sum_{k=1}^r y_k(\xi_k) h_k(\xi_k) & \text{if } h \neq 0, \\ \delta_0 & \text{if } h = 0, \end{cases} \tag{2.3}$$

where $|h_k(\xi_k)| = |h_k|_{C(\overline{\mathcal{O}})}$ for any $k = 1, \ldots, r$, and δ_0 is any element of the unitary ball of $E^\star$. It is possible to show that $\delta_h \in \partial\, |h|_E$ (see [11] and [6] for more details), and for any $x, h \in E$

$$\langle F(x+h) - F(x), \delta_h \rangle_E \leq c\, |h|_E.$$

Remark 2.1. For any $k = 1, \ldots, r$, let us define

$$g_k(\xi, t) = -c_k(\xi) t^{2m+1} + \sum_{j=0}^{2m} c_{kj}(\xi) t^j,$$

where c_k, c_{kj} are bounded continuous functions from $\overline{\mathcal{O}}$ into $\mathbb{R}$. If we assume that

$$\inf_{\xi \in \overline{\mathcal{O}}} c_k(\xi) > 0,$$

then it is possible to check that g_k fulfills parts 2 and 3 of Hypothesis 1.

Due to Hypothesis 1, there exists $c \in \mathbb{R}$ such that the mapping $\gamma(\xi, \cdot) = f(\xi, \cdot) - cI$ is dissipative for any $\xi \in \overline{\mathcal{O}}$. Then for any $\alpha > 0$ we can define the function

$$\gamma_\alpha : \overline{\mathcal{O}} \times \mathbb{R}^r \to \mathbb{R}^r, \qquad (\xi, \sigma) \mapsto \gamma(\xi, J_\alpha(\xi, \sigma)),$$

where

$$J_\alpha(\xi, \sigma) = (I - \alpha \gamma(\xi, \cdot))^{-1}(\sigma), \qquad \sigma \in \mathbb{R}^r.$$

As proved in [9, appendix A], the function $J_\alpha(\xi, \cdot)$ is of class C^2 for any fixed $\xi \in \overline{\mathcal{O}}$. Now if we set

$$f_\alpha(\xi, \sigma) = \gamma_\alpha(\xi, \sigma) + c J_\alpha(\xi, \sigma),$$

we have that $f_\alpha(\xi, \cdot)$ is Lipschitz continuous, uniformly with respect to $\xi \in \overline{\mathcal{O}}$, is twice differentiable, and

$$\sup_{\xi \in \overline{\mathcal{O}}} \langle f_\alpha(\xi, \sigma + \rho) - f_\alpha(\xi, \sigma), \rho \rangle_{\mathbb{R}^r} \leq c\, |\rho|^2 \tag{2.4}$$

for some constant c independent of α. Moreover, by using well-known properties of the function $J_\alpha(\xi, \sigma)$ (see [11] for the definitions and main results and see [9, chapter 9, appendix A],

$$\sup_{\xi \in \overline{\mathcal{O}}} |f_\alpha(\xi, \sigma) - f(\xi, \sigma)| \leq \alpha\, c(1 + |\sigma|^{4m+1}). \tag{2.5}$$

For any fixed $\xi \in \overline{\mathcal{O}}$ the function $f_\alpha(\xi, \cdot)$ is of class C^2, and for any $R > 0$

$$\lim_{\alpha \to 0} \sup_{\xi \in \overline{\mathcal{O}}} \sup_{|\sigma| \leq R} |D^j_\sigma f_\alpha(\xi, \sigma) - D^j_\sigma f(\xi, \sigma)| = 0. \tag{2.6}$$

Moreover, it is possible to show that

$$\sup_{\xi \in \overline{\mathcal{O}}} \frac{|D^j_\sigma f_\alpha(\xi,\sigma)|}{1+|\sigma|^{2m+1-j}} \leq c < \infty \tag{2.7}$$

for a constant c independent of α.

For each $\alpha > 0$, let F_α be the Nemytskii operator associated with the function f_α. Clearly F_α is Lipschitz continuous both as an operator in E and as an operator in H and is twice Fréchet differentiable in E, and, thanks to (2.4), there exists a constant c independent of α such that if $x, y \in H$,

$$\langle F_\alpha(x) - F_\alpha(y), x-y\rangle_H \leq c\,|x-y|^2_H, \tag{2.8}$$

and if $x, y \in E$,

$$\langle F_\alpha(x) - F_\alpha(y), \delta_{x-y}\rangle_E \leq c\,|x-y|_E, \tag{2.9}$$

where δ_{x-y} is the element in $\partial\,|x-y|_E$ introduced in (2.3). Furthermore, due to (2.6), for each $j = 0, 1, 2$ it holds that

$$\lim_{\alpha\to 0} \sup_{|x|_E \leq R} |D^j F_\alpha(x) - D^j F(x)|_{\mathcal{L}^j(E)} = 0 \tag{2.10}$$

for any $R > 0$, and due to (2.7)

$$|D^j F_\alpha(x)|_{\mathcal{L}^j(E)} \leq c\left(1 + |x|_E^{2m+1-j}\right), \qquad x \in E. \tag{2.11}$$

2.2. The operators A and Q and the stochastic convolution. We shall denote by $\mathcal{A}$ the second order differential operator defined for each $x \in H$ by $\mathcal{A}x = (\mathcal{A}_1 x_1, \ldots, \mathcal{A}_r x_r)$. For any $k = 1, \ldots, r$ we have

$$\mathcal{A}_k(\xi, D) = \sum_{i,j=1}^d a^{ij}_k(\xi) \frac{\partial^2}{\partial\xi_i \partial\xi_j} + \sum_{i=1}^d b^i_k(\xi)\frac{\partial}{\partial\xi_i}, \qquad \xi \in \overline{\mathcal{O}}.$$

The coefficients a^{ij}_k and b^i_k are assumed to be of class $C^1(\overline{\mathcal{O}})$, and for any $\xi \in \overline{\mathcal{O}}$ the matrix $[a^{ij}_k(\xi)]$ is symmetric and satisfies the uniform ellipticity condition

$$\inf_{\xi \in \overline{\mathcal{O}}} \sum_{i,j=1}^d a^{ij}_k(\xi) h_i h_j \geq \nu |h|^2, \quad h \in \mathbb{R}^d,$$

for some $\nu > 0$. The boundary operator $\mathcal{B}$ is defined by $\mathcal{B}x = (\mathcal{B}_1 x_1, \ldots, \mathcal{B}_r x_r)$, and for each $k = 1, \ldots, r$ we have

$$\mathcal{B}_k(\xi, D) = I \quad \text{or} \quad \mathcal{B}_k(\xi, D) = \sum_{i,j=1}^d a^{ij}_k(\xi)\nu_j(\xi)\frac{\partial}{\partial\xi_i}, \qquad \xi \in \ \partial\,\mathcal{O}. \tag{2.12}$$

We denote by A the realization in H of the elliptic operator $\mathcal{A}$, with the boundary conditions given by $\mathcal{B}$, that is,

$$D(A) = \left\{\, x \in\ H\ :\ \mathcal{A}x \in\ H,\ \mathcal{B}x_{|_{\partial D}} = 0\ \right\}, \quad Ax = \mathcal{A}x,\ x \in\ D(A).$$

The operator A generates an analytic semigroup e^{tA}. The semigroup e^{tA} is also analytic in each $L^p(\mathcal{O};\mathbb{R}^r)$, for $p \in (1,+\infty]$ (see [26, chapter 3] for all details).

In what follows it will not be restrictive to assume that for any $p \in [2,\infty]$

$$|e^{tA}x|_p \leq M|x|_p$$

for some constant $M > 0$ independent of p (see also [9, chapter 4]). In particular, we will have

$$\langle Ax, x\rangle_H \leq 0, \qquad x \in H. \tag{2.13}$$

Finally, if we denote by A the realization in E of the operator $\mathcal{A}$ with the boundary conditions given by $\mathcal{B}$, we have that A generates an analytic semigroup $e^{t\,A}$ of negative type; that is, for any $x \in E$ and $\delta_x \in \partial\,|x|_E$ defined as in (2.3)

$$\langle Ax, \delta_x\rangle_E \leq 0. \tag{2.14}$$

Now for any $k = 1,\dots,r$ we define

$$\mathcal{G}_k(\xi, D) = \sum_{i=1}^{d}\left(b_k^i(\xi) - \sum_{j=1}^{d}\frac{\partial a_k^{ij}}{\partial \xi_j}(\xi)\right)\frac{\partial}{\partial \xi_i}, \qquad \xi \in \mathcal{O},$$

and by difference we define $\mathcal{C}_k = \mathcal{A}_k - \mathcal{G}_k$. The second order elliptic operators $\mathcal{C} = (\mathcal{C}_1,\dots,\mathcal{C}_r)$, generate a negative analytic semigroup e^{tC} in each $L^p(\mathcal{O};\mathbb{R}^r)$ for any $p \in (1,\infty]$ and in E. The semigroup e^{tC} enjoys the same properties as $e^{t\,A}$ and, due to the boundary conditions (2.12), is self-adjoint in H. Moreover, for any $\delta \in \mathbb{R}$ we have that $D((-A)^\delta) = D((-C)^\delta)$ and

$$c_1\,|(-A)^\delta x|_H \leq |(-C)^\delta x|_H \leq c_2\,|(-A)^\delta x|_H \tag{2.15}$$

for suitable positive constants c_1 and c_2 depending only on δ.

Concerning the realization of the operator $\mathcal{G}$, as the coefficients a_k^{ij} and b_k^i are assumed to be smooth, it is easy to check that $D(G^\star) \subset D((-C)^{1/2})$ and

$$|G^\star x|_H \leq c\,|(-C)^{1/2}x|_H, \qquad x \in D(G^\star). \tag{2.16}$$

Finally, we denote by Q the bounded linear operator of components $(Q_1,\dots,Q_r)$. In what follows we shall assume that the operators Q and C fulfill the following conditions.

Hypothesis 2.

1. *There exists a complete orthonormal basis $\{e_k\}$ in H which diagonalizes C such that* $\sup_{k\in\mathbb{N}}|e_k|_E < \infty$. *The corresponding set of eigenvalues is denoted by $\{-\alpha_k\}$.*
2. *The bounded linear operator $Q : H \to H$ is nonnegative and diagonal with respect to the complete orthonormal basis $\{e_k\}$ which diagonalizes C. Moreover, if $\{\lambda_k\}$ is the corresponding set of eigenvalues, we have*

$$\sum_{k=1}^{\infty}\frac{\lambda_k^2}{\alpha_k^{1-\gamma}} < +\infty$$

for some $\gamma > 0$.

3. *There exists* $\epsilon < 1$ *such that*

$$D((-C)^{\frac{\epsilon}{2}}) \subset D(Q^{-1}). \tag{2.17}$$

Remark 2.2. It is known (see, for example, the book by Agmon [1]) that when the elliptic operator $\mathcal{A}$ with the boundary conditions $\mathcal{B}$ is smooth enough, then

$$\alpha_k \asymp k^{2/d}.$$

In this case, it is possible to prove that if $d \leq 3$, then there exists an operator Q which fulfills the conditions of parts 2 and 3 of Hypothesis 2.

Actually, if we assume that

$$\lambda_k \asymp \alpha_k^{-\rho}$$

for some $\rho > (d-2)/4$, then

$$\frac{\lambda_k^2}{\alpha_k^{1-\gamma}} \asymp \alpha_k^{-(1-\gamma+2\rho)} \asymp k^{-\frac{2(1-\gamma+2\rho)}{d}}.$$

As $1+2\rho > d/2$, we can fix $\gamma > 0$ such that $1-\gamma+2\rho > d/2$, and this implies that

$$\sum_{k=1}^{\infty} \frac{\lambda_k^2}{\alpha_k^{1-\gamma}} \asymp \sum_{k=1}^{\infty} k^{-\frac{2(1-\gamma+2\rho)}{d}} < \infty.$$

On the other hand, if $\rho \leq \epsilon/2$, then (2.17) holds. This means that if $d \leq 3$, it is possible to find ρ such that Q enjoys conditions 2 and 3 in Hypothesis 2. Notice that in dimension $d = 1$ one can take $\epsilon = 0$.

Let $\{w_k(t)\}$ be a sequence of mutually independent real-valued Brownian motions defined on a stochastic basis $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbf{P})$ and adapted to the nonanticipative filtration $\mathcal{F}_t$, $t \geq 0$. We define the cylindrical Wiener process $w(t)$ as

$$w(t) = \sum_{k=1}^{\infty} e_k w_k(t),$$

where $\{e_k\}$ is the complete orthonormal system of H introduced in part 1 of Hypothesis 2. The series above defining $w(t)$ does not converge in H, but it is convergent in any Hilbert space U such that the embedding $H \subset U$ is Hilbert–Schmidt (see [15, chapter 4]).

Now we consider the Ornstein–Uhlenbeck problem associated with the system (1.1)

$$dv(t) = Av(t)\,dt + Q\,dw(t), \qquad v(s) = 0,$$

for $0 \leq s \leq t \leq T$. Due to parts 1 and 2 of Hypothesis 2, such a problem admits a unique solution $w^A(t,s)$, which is the mean-square Gaussian process with values in H given by

$$w^A(t,s) = \int_s^t e^{(t-r)\,A} Q\,dw(r) \tag{2.18}$$

(see, e.g., [15] for a proof). Moreover, as shown in [8], $w^A(\cdot,s) \in C([s,T]\times\overline{\mathcal{O}})$, $\mathbf{P}$-almost surely (a.s.), and for any $p \geq 1$ it holds that

$$\mathbf{E}\,|w^A(\cdot,s)|^p_{C([s,T];E)} < \infty. \tag{2.19}$$

For any $n \in \mathbb{N}$ we define

$$A_n = P_n A P_n, \qquad C_n = C P_n, \qquad G_n = P_n G P_n, \qquad Q_n = Q P_n,$$

where P_n is the projection of H onto the finite dimensional space H_n generated by the eigenfunctions $\{e_1,\dots,e_n\}$. If we denote by $w^A_n(t,s)$ the solution of the problem

$$dv(t) = A_n v(t)\,dt + Q_n\,dw(t), \qquad v(s)=0,$$

by using a factorization argument (see [15]) it is not difficult to prove that for any $p \geq 1$

$$\lim_{n\to+\infty} \mathbf{E}\,|w^A(\cdot,s) - w^A_n(\cdot,s)|^p_{C([s,T];H)} = 0.$$

3. The state equation. By using the notations introduced in the previous section, the controlled system (1.1) can be rewritten in the abstract form

$$dy(t) = (Ay(t) + F(y(t)) + z(t))\;dt + Q\,dw(t), \qquad y(s) = x, \tag{3.1}$$

for $0 \leq s \leq t \leq T$.

Definition 3.1.

1. *Let us fix an adapted process $z \in L^2(\Omega; L^2(0,T;E))$ and $x \in E$. An E-valued predictable process $y(t) = y(t,s;x,z)$ is a* mild *solution for the problem* (3.1) *if*

$$y(t) = e^{(t-s)\,A}x + \int_s^t e^{(t-r)\,A}\,(F(y(r)) + z(r))\;dr + w^A(t,s),$$

where the process $w^A(t,s)$ is given by (2.18).

2. *Let us fix an adapted process $z \in L^2(\Omega; L^2(0,T;H))$ and $x \in H$. A H-valued process $y(t,s;x,z)$ is a* generalized *solution for the problem* (3.1) *if, for any sequences $\{x_n\} \subset E$ converging to x in H and $\{z_n\} \subset L^2(\Omega;L^2(0,T;E))$ converging to z in $L^2(\Omega;L^2(0,T;H))$, the corresponding sequence of mild solutions $\{y(\cdot,s;x_n,z_n)\}$ converges to $y(\cdot,s;x,z)$ in $C([s,T];H)$, $\mathbf{P}$-a.s.*

In [15] (see also [6] and [7]) the following existence and uniqueness result is proved for the uncontrolled system. When $z \neq 0$, the proof is analogous, and we do not repeat it.

Theorem 3.2. *Assume Hypotheses* 1 *and* 2, *and fix $0 \leq s \leq T$.*

1. *For any $z \in L^2(\Omega; L^p(s,T;H))$ with $p > 4/(4-d)$ and for any $x \in E$, the problem* (3.1) *admits a unique mild solution $y(\cdot,s;x,z) \in L^2(\Omega; C((s,T];E) \cap L^\infty(s,T;E))$ such that*

$$|y(t,s;x,z)|_E \leq c_T\Big(|x|_E + |z|^{2m+1}_{L^p(s,t;H)} + \sup_{r\in\,[s,t]} |w^A(r,s)|^{2m+1}_E\Big), \qquad \mathbf{P} - a.s. \tag{3.2}$$

2. *For any $z \in L^2(\Omega; L^2(s,T;H))$ and $x \in H$, the problem* (3.1) *admits a unique generalized solution $y(\cdot, s; x, z) \in L^2(\Omega; C([s,T]; H))$ such that*

$$|y(t,s;x,z)|_H \leq c_T\Big(|x|_H + |z|^{2m+1}_{L^2(s,T;H)} + \sup_{r\in [s,t]} |w^A(r,s)|^{2m+1}_E\Big), \qquad \mathbf{P} - a.s. \tag{3.3}$$

3. *The unique generalized solution $y(\cdot, s; x, z)$ belongs to $L^{p_\star}(t,T; L^{p_\star}(\mathcal{O}))$, $\mathbf{P}$-a.s., and*

$$y(t,s;x,z) = e^{(t-s)\,A}x + \int_s^t e^{(t-r)\,A}\left(F(y(r,s;x,z)) + z(r)\right) dr + w^A(t,s).$$

4. *For any $x_1, x_2 \in H$ and $z_1, z_2 \in L^2(\Omega; L^2(s,T;H))$, we have*

$$|y(t,s;x_1,z_1) - y(t,s;x_2,z_2)|_H \leq c_T\left(|x_1 - x_2|_H + |z_1 - z_2|_{L^2(s,T;H)}\right). \tag{3.4}$$

For any $\alpha > 0$ we consider the approximating problem

$$dy(t) = (Ay(t) + F_\alpha(y(t)) + z(t))\,dt + Q\,dw(t), \qquad y(s) = x, \tag{3.5}$$

$s \leq t \leq T$. Clearly an existence theorem analogous to Theorem 3.2 holds for (3.5). Actually, for each $x \in E$ and $z \in L^2(\Omega; L^p(0,T;H))$, with $p > 4/(4-d)$, there exists a unique mild solution $y_\alpha(\cdot, s; x, z)$ in $L^2(\Omega; C((s,T]; E) \cap L^\infty(s,T;E))$, and for each $x \in H$ and $z \in L^2(\Omega; L^2(0,T;H))$ there exists a unique generalized solution $y_\alpha(\cdot, s; x, z)$ which belongs to $L^2(\Omega; C([s,T]; H))$. Moreover, estimates analogous to (3.2) and (3.3) hold for every $\alpha > 0$.

LEMMA 3.3. *Under Hypotheses* 1 *and* 2, *if $x \in E$ and $z \in L^p(\Omega; L^\infty(0,T;H))$ for p sufficiently large, then for any fixed $q \geq 1$*

$$\lim_{\alpha\to 0} \mathbf{E}\,|y_\alpha(t,s;x,z) - y(t,s;x,z)|^q_E = 0, \tag{3.6}$$

uniformly with respect to $t \in [0,T]$, x in bounded subsets of E and z in the set

$$\mathcal{M}_R = \Big\{\, z \in L^2(\Omega; L^2(0,T;H))\ ;\ \sup_{t\in [0,T]} |z(t)|_H \leq R,\ \mathbf{P} - a.s. \Big\}$$

for any $R > 0$.

Proof. If we set $v_\alpha(t) = y_\alpha(t) - y(t)$, we have that v_α is the unique solution of the problem

$$\frac{dv}{dt}(t) = Av(t) + F_\alpha(y_\alpha(t)) - F(y(t)), \qquad v(s) = 0.$$

Thus, by using classical properties of the subdifferential of the norm in E introduced in (2.3) (see [11] for all properties), if $\delta_{v_\alpha(t)} \in \partial\,|v_\alpha(t)|_E$, we have

$$\frac{d}{dt}^{-}|v_\alpha(t)|_E \leq \left\langle Av_\alpha(t), \delta_{v_\alpha(t)}\right\rangle_E + \left\langle F_\alpha(y_\alpha(t)) - F(y(t)), \delta_{v_\alpha(t)}\right\rangle_E.$$

From (2.9) and (2.14) this easily implies that

$$\frac{d}{dt}^{-}|v_\alpha(t)|_E \leq c\,|v_\alpha(t)|_E + |F_\alpha(y(t)) - F(y(t))|_E$$

so that, due to the Gronwall lemma and (2.5), we have

$$|v_\alpha(t)|_E \leq \alpha\, c \int_s^t e^{c(t-r)} \left(1 + |y(r)|_E^{4m+1}\right) dr.$$

This implies (3.6), as from (2.19) and (3.2) for any $q \geq 1$ we have

$$\sup_{z \in \mathcal{M}_R} \mathbf{E}\, |y(\cdot, s; x, z)|^q_{C([s,T];E)} < \infty. \qquad \square \tag{3.7}$$

Next, for any $n \in \mathbb{N}$ and $\alpha > 0$ we define

$$F_{\alpha,n}(x) = P_n F_\alpha(P_n x), \qquad x \in H.$$

It is immediate to check that for any $x, y \in H$ it holds that

$$\langle F_{\alpha,n}(x) - F_{\alpha,n}(y), x - y \rangle_H \leq c\, |x - y|_H^2 \tag{3.8}$$

for a constant c independent of n and α. Moreover,

$$|F_{\alpha,n}(x) - F_{\alpha,n}(y)|_H \leq c_\alpha\, |x - y|_H \tag{3.9}$$

for some constant c_α independent of n. In correspondence with each $n \in \mathbb{N}$, $\alpha > 0$, and $0 \leq s \leq T$, we consider the approximating problem

$$dy(t) = (A_n y(t) + F_{\alpha,n}(y(t)) + z_n(t))\, dt + Q_n\, dw(t), \qquad y(s) = P_n x, \tag{3.10}$$

where $z_n(t) = P_n z(t)$ and z is an adapted process in $L^2(\Omega; L^2(s,T;H))$. Such a problem is a finite dimensional problem with Lipschitz coefficients. Thus for any $x \in H$ there exists a unique strong solution $y_{\alpha,n}(\cdot, s; x, z) \in L^2(\Omega; C([s,T];H))$.

LEMMA 3.4. *Let z be an adapted process in $L^2(\Omega; L^2(s,T;H))$. If $y_{\alpha,n}(\cdot, s; x, z)$ is the unique solution of the approximating problem* (3.10), *it holds that*

$$\lim_{n \to +\infty} y_{\alpha,n}(\cdot, s; x, z) = y_\alpha(\cdot, s; x, z) \qquad \text{in } L^2(\Omega; C([s,T];H)), \tag{3.11}$$

uniformly for x in bounded subsets of H.

Proof. For each $n, k \in \mathbb{N}$, we consider the problem

$$dy(t) = (A_n y(t) + F_{\alpha,n}(y(t)) + z_{n \wedge k}(t))\, dt + Q_{k \wedge n}\, dw(t), \qquad y(0) = P_n x. \tag{3.12}$$

By using a factorization argument, we have that for any $p \geq 1$

$$\lim_{k \to +\infty} \sup_{n \in \mathbb{N}} \mathbf{E} \sup_{t \in [s,T]} |w^A_{n,k}(t,s)|_H^p = 0,$$

and, since $z \in L^2(\Omega; L^2(s,T;H))$, we have

$$\lim_{k \to +\infty} \sup_{n \in \mathbb{N}} \mathbf{E}\, |z_n - z_{n \wedge k}|^2_{L^2(s,T;H)} = 0.$$

Thus, by some calculations, if we denote by $y^k_{\alpha,n}(t)$ the solution of (3.12), we have

$$\lim_{k \to +\infty} \sup_{n \in \mathbb{N}} |y^k_{\alpha,n}(\cdot, s; x, z) - y_{\alpha,n}(\cdot, s; x, z)|_{L^2(\Omega; C([s,T];H))} = 0. \tag{3.13}$$

Now for any $k \in \mathbb{N}$ we consider the problem

$$dy(t) = (Ay(t) + F_\alpha(y(t)) + z_k(t))\, dt + Q_k\, dw(t), \qquad y(s) = x.$$

It is immediate to check that $w_k^A \in L^p(\Omega, C((s,T]; D((-A)^\delta)))$ for any $\delta \in \mathbb{R}$ and $p \geq 1$. Hence, by straightforward computations, thanks to (2.15) we have that such a problem admits a unique mild solution $y_\alpha^k(\cdot, s; x, z)$ such that

$$\begin{aligned} & |y_\alpha^k(t)|_H^2 + \int_0^t |(-C)^{1/2} y_\alpha^k(s)|_H^2\, ds \\ & \leq c_T \Big(|x|_H^2 + \sup_{t \in [s,T]} |w_k^A(t,s)|^2_{D((-A)^{1/2})} + |z|^2_{L^2(s,T;H)} \Big). \end{aligned} \tag{3.14}$$

Moreover, it is possible to show that for any fixed $k \in \mathbb{N}$

$$\lim_{n \to +\infty} y_{\alpha,n}^k(\cdot, s; x, z) = y_\alpha^k(\cdot, s; x, z) \qquad \text{in } L^2(\Omega; C([s,T]; H)). \tag{3.15}$$

Finally, we have that

$$\lim_{k \to +\infty} y_\alpha^k(\cdot, s; x, z) = y_\alpha(\cdot, s; x, z) \qquad \text{in } L^2(\Omega; C([s,T]; H)). \tag{3.16}$$

Indeed, if we define $v_\alpha^k(t) = y_\alpha(t) - y_\alpha^k(t) - w^A(t,s) + w_k^A(t,s)$, we have that $v_\alpha^k(t)$ is the unique solution for the problem

$$\frac{dv}{dt}(t) = Av(t) + F_\alpha(y_\alpha(t)) - F_\alpha(y_\alpha^k(t)) + z(t) - z_k(t), \qquad v(s) = 0.$$

Thus, by multiplying each side by $v_\alpha^k(t)$, we have

$$\begin{aligned} & \frac{1}{2}\frac{d}{dt}|v_\alpha^k(t)|_H^2 + |(-A)^{1/2} v_\alpha^k(t)|_H^2 \\ & = \left\langle F_\alpha(y_\alpha(t)) - F_\alpha(y_\alpha^k(t)), v_\alpha^k(t) \right\rangle_H + \left\langle z(t) - z_k(t), v_\alpha^k(t) \right\rangle_H \end{aligned}$$

so that, as F_α is Lipschitz continuous, we easily get

$$\frac{1}{2}\frac{d}{dt}|v_\alpha^k(t)|_H^2 \leq c_\alpha\, |v_\alpha^k(t)|_H^2 + c_\alpha\, |w^A(t,s) - w_k^A(t,s)|_H^2 + c\, |z(t) - z_k(t)|_H^2.$$

By applying the Gronwall lemma, by taking the supremum over $t \in [s,T]$, and, finally, by taking the expectation, we get

$$\mathbf{E} \sup_{t \in [s,T]} |v_\alpha^k(t)|_H^2 \leq c_{\alpha,T} \int_s^T \mathbf{E} \left(|w^A(t,s) - w_k^A(t,s)|_H^2 + |z(t) - z_k(t)|_H^2 \right) dt,$$

and this immediately implies (3.16).

Now we can conclude. Actually, due to (3.13) and (3.16), for any $\epsilon > 0$ there exists $k_\epsilon \in \mathbb{N}$ such that for any $n \in \mathbb{N}$ it holds that

$$\mathbf{E} \sup_{t \in [s,T]} \left(|y_{\alpha,n}^{k_\epsilon}(t) - y_{\alpha,n}(t)|_H^2 + |y_\alpha^{k_\epsilon}(t) - y_\alpha(t)|_H^2 \right) < \epsilon/2.$$

Besides, due to (3.15) there exists $n_\epsilon \in \mathbb{N}$ such that

$$\mathbf{E} \sup_{t \in [s,T]} |y_{\alpha,n}^{k_\epsilon}(t) - y_\alpha^{k_\epsilon}(t)|_H^2 < \epsilon/2$$

for any $n \geq n_\epsilon$ so that (3.11) follows. $\square$

4. The first variation equation. Here and in what follows, we shall denote, respectively, by $y(t;x)$, $y_\alpha(t;x)$ and $y_{\alpha,n}(t;x)$ the mild solutions of the problems (3.1), (3.5), and (3.10) when $z=0$ and $s=0$.

In the present section we study the first variation equation associated with the problem (3.1):

$$\frac{dv}{dt}(t) = Av(t) + DF(y(t;x))v(t), \qquad v(0)=h. \tag{4.1}$$

In [6, Theorem 4.2] we have proved that for any $x,h \in E$ there exists a unique mild solution for the problem (4.1), and for any $t \geq 0$ and $x,h \in E$ such a solution is given by $Dy(t;x)h$, the Fréchet derivative of the mapping

$$E \to L^2(\Omega;E), \qquad x \mapsto y(t;x),$$

along the direction h. Moreover, in [7, Proposition 4.1] we have proved that if $x,h \in H$, then the problem (4.1) admits a unique generalized solution $v(x,h)$.

As proved in [5], under Hypotheses 1 and 2 the solution $y_\alpha(t;x)$ is twice mean-square differentiable with respect to $x \in H$. Moreover, the first derivative $Dy_\alpha(t;x)h$ is the unique solution of the first variation equation corresponding to the problem (3.5), which is

$$\frac{dv}{dt}(t) = Av(t) + DF_\alpha(y_\alpha(t))v(t), \qquad v(0)=h.$$

We have the following approximation result.

Lemma 4.1. *Under Hypotheses* 1 *and* 2, *for any $x \in E$ and $t \geq 0$, it holds that*

$$\lim_{\alpha\to 0} \mathbf{E} \sup_{|h|_H \leq 1} |Dy_\alpha(\cdot;x)h - Dy(\cdot;x)h)|^2_{L^\infty(0,T;H)\cap L^2(0,T;D((-A)^{1/2}))} = 0, \tag{4.2}$$

uniformly for x in bounded sets of E.

Proof. As proved in [6], for any $h \in H$

$$\sup_{x\in H}\left(|Dy(t;x)h|^2_H + \int_0^t |(-A)^{1/2}Dy(s;x)h|^2_H\right) \leq c_T|h|^2_H, \qquad \mathbf{P}-\text{a.s.} \tag{4.3}$$

If we define $v_\alpha(t) = Dy_\alpha(t;x)h - Dy(t;x)h$, we have that $v_\alpha(t)$ is the unique solution to the problem

$$\frac{dv}{dt}(t) = Av(t) + DF_\alpha(y_\alpha(t;x))Dy_\alpha(t;x)h - DF(y(t;x))Dy(t;x)h, \quad v(0)=0.$$

Thus we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|v_\alpha(t)|^2_H + |(-A)^{1/2}v_\alpha(t)|^2_H \\
&= \langle DF_\alpha(y_\alpha(t))v_\alpha(t), v_\alpha(t)\rangle_H + \langle\left(DF_\alpha(y_\alpha(t)) - DF(y(t))\right)Dy(t;x)h, v_\alpha(t)\rangle_H \\
&\leq c\,|v_\alpha(t)|^2_H + c\,e^{cT}\,|DF_\alpha(y_\alpha(t)) - DF(y(t))|^2_E|h|^2_H,
\end{aligned}$$

the last inequality following from (2.8), (4.3), and the Young inequality. Since F_α verifies the estimate (2.11), for any $x,y \in E$ we have

$$\begin{aligned}
&|DF_\alpha(x) - DF(y)|_E \leq |DF_\alpha(x) - DF_\alpha(y)|_E + |DF_\alpha(y) - DF(y)|_E \\
&\leq c\left(1 + |x|^{2m-1}_E + |y|^{2m-1}_E\right)|x-y|_E + |DF_\alpha(y) - DF(y)|_E.
\end{aligned}$$

Therefore, thanks to the Gronwall lemma and the above inequality, we have

$$|v_\alpha(t)|_H^2 \leq c_T \int_0^t \left(1+|y_\alpha(s)|_E^{4m-2}+|y(s)|_E^{4m-2}\right)|y_\alpha(s)-y(s)|_E^2\,ds|h|_H^2$$

$$+c_T\int_0^t |DF_\alpha(y(s))-DF(y(s))|_E\,ds|h|_H^2.$$

Due to (2.9) it is immediate to check that

$$\sup_{\alpha>0}|y_\alpha(t;x)|_E \leq c_T\Big(|x|_E+\sup_{t\in\,[0,T]}|w^A(t,0)|_E\Big), \qquad \mathbf{P}-\text{a.s.}, \tag{4.4}$$

and hence, by using (2.10), (3.2), and (3.6), we have

$$\lim_{\alpha\to 0}\mathbf{E}\sup_{|h|_H\leq 1}|v_\alpha(t)|_H^2=0.$$

This immediately yields

$$\lim_{\alpha\to 0}\mathbf{E}\sup_{|h|_H\leq 1}\int_0^t|(-A)^{1/2}v_\alpha(s)|_H^2\,ds=0,$$

and (4.2) holds true. □

Due to part 2 of Hypothesis 2, and the closed graph theorem, we have that the operator $\Gamma_\epsilon = Q^{-1}(-A)^{-\epsilon/2}$ is bounded in H. Thus, for any $x\in D((-A)^{1/2})$, by interpolation we get

$$|Q^{-1}x|_H \leq c\,|(-A)^{1/2}x|_H^\epsilon\,|x|_H^{1-\epsilon}. \tag{4.5}$$

Therefore, from (4.2) we get

$$\lim_{\alpha\to 0}\mathbf{E}\sup_{|h|_H\leq 1}\int_0^t|Q^{-1}(Dy_\alpha(s;x)h-Dy(s;x)h)|_H^2\,ds=0, \tag{4.6}$$

uniformly for x in bounded sets of E.

For each $n\in\mathbb{N}$ and $\alpha>0$, the solution of (3.10) is twice mean-square differentiable with respect to $x\in H$. In the next lemma we show that we can approximate in a suitable sense $Dy_\alpha(t;x)h$ by means of $Dy_{\alpha,n}(t;x)h$.

Lemma 4.2. *Assume that Hypotheses* 1 *and* 2 *hold. Then*

$$\lim_{n\to+\infty}\mathbf{E}\sup_{|h|_H\leq 1}|Dy_{\alpha,n}(\cdot;x)h-Dy_\alpha(\cdot;x)P_nh|^2_{L^\infty(0,T;H)\cap L^2(0,T;D((-A)^{1/2}))}=0, \tag{4.7}$$

uniformly for x in bounded subsets of H.

Proof. If we set $v_{\alpha,n}(t)=Dy_{\alpha,n}(t;x)h-Dy_\alpha(t;x)P_nh$, we have that $v_{\alpha,n}(t)$ is the unique solution of the problem

$$\frac{dv}{dt}(t)=Cv(t)+G_nDy_{\alpha,n}(t)h-GDy_\alpha(t)P_nh$$

$$+DF_{\alpha,n}(y_{\alpha,n}(t;x))Dy_{\alpha,n}(t)h-DF_\alpha(y_\alpha(t;x))Dy_\alpha(t)P_nh, \qquad v(0)=0.$$

By using (2.15), (2.16), and (3.9) by some computations, we get

$$\tag{4.8}\begin{aligned}&\frac{d}{dt}|v_{\alpha,n}(t)|_H^2+|(-C)^{1/2}v_{\alpha,n}(t)|_H^2\le c_\alpha\,|v_{\alpha,n}(t)|_H^2\\&+c_{\alpha,T}\left(\|P_n-I\|^2_{\mathcal{L}(D((-C)^{1/2});H)}+|y_{\alpha,n}-y_\alpha|^2_{C([0,T];H)}\right)|Dy_\alpha(t;x)P_nh|^2_{D((-A)^{1/2})}.\end{aligned}$$

In [6] it is proved that for each $h\in H$

$$\sup_{x\in H}\int_0^t|Dy_\alpha(s;x)h|^2_{D((-A)^{1/2})}\,ds\le c_T\,|h|_H^2,\qquad \mathbf{P}-\text{a.s.},$$

and then, by using the Gronwall lemma, this yields

$$|v_{\alpha,n}(t)|_H^2\le c_{\alpha,T}\left(\|P_n-I\|^2_{\mathcal{L}(D((-C)^{1/2});H)}+|y_{\alpha,n}-y_\alpha|^2_{C([0,T];H)}\right)|h|_H^2.$$

Thus, as

$$\lim_{n\to+\infty}\|P_n-I\|_{\mathcal{L}(D((-C)^{1/2});H)}=0,$$

from Lemma 3.4 we get

$$\lim_{n\to+\infty}\mathbf{E}\sup_{|h|_H\le1}\sup_{t\in[0,T]}|Dy_{\alpha,n}(t;x)h-Dy_\alpha(t;x)P_nh|_H^2=0.$$

Thanks to (4.8), from the limit above we get

$$\lim_{n\to+\infty}\mathbf{E}\sup_{|h|_H\le1}\int_0^t|(-A)^{1/2}(Dy_{\alpha,n}(s;x)h-Dy_\alpha(s;x)P_nh)|_H^2\,ds=0$$

so that (4.7) follows. □

By using the interpolation inequality (4.5), we get

$$\tag{4.9}\lim_{n\to+\infty}\mathbf{E}\sup_{|h|_H\le1}\int_0^t\left|Q^{-1}\left(Dy_{\alpha,n}(s;x)h-Dy_\alpha(s;x)P_nh\right)\right|_H^2\,ds=0.$$

5. The transition semigroup. The transition semigroup P_t associated with the system (1.1) is defined for any $\varphi\in B_b(H)$ and $x\in H$ by

$$P_t\varphi(x)=\mathbf{E}\,\varphi(y(t;x)),\qquad t\ge0,$$

where $y(t;x)$ is the solution of the problem (1.1), with $z=0$, starting from x at time zero.

As proved in [7], P_t is a contraction semigroup on $C_b(H)$. In general, P_t is not strongly continuous in $C_b(H)$. Nevertheless, $y(\cdot;x)\in L^2(\Omega;C([0,T];H))$ for any fixed $x\in H$ so that the mapping

$$[0,+\infty)\to\mathbb{R},\qquad t\mapsto P_t\varphi(x),$$

is continuous for any $\varphi\in C_b(H)$.

In [7, Theorem 5.1] we have also proved that P_t has a smoothing effect. Namely, we have shown that $P_t : B_b(H) \to C_b^1(H)$ for any $t > 0$, and if ϵ is the constant introduced in part 3 of Hypothesis 2.

$$\|P_t\varphi\|_j^H \leq c_0\,(t \wedge 1)^{-\frac{(j-i)(1+\epsilon)}{2}}\,\|\varphi\|_i^H, \qquad i \leq j \leq 0, 1, \tag{5.1}$$

for some constant $c_0 > 0$. Moreover, if $\varphi \in C_b(H)$, for any $x, h \in H$ it holds that

$$\langle D(P_t\varphi)(x), h\rangle_H = \frac{1}{t}\,\mathbf{E}\,\varphi(y(t;x)) \int_0^t \left\langle Q^{-1}v(s;x,h), dw(s)\right\rangle_H, \tag{5.2}$$

where $v(s;x,h)$ is the unique generalized solution of the problem (4.1). The formula above is a generalization to the degenerate case of the Bismut–Elworthy formula (see [2] and [16] for the finite dimension and [27] for the infinite dimension).

Now for any $\alpha > 0$ we define P_t^α as the transition semigroup corresponding to the approximating problem (3.5) with $z = 0$. As proved in [5], the semigroup P_t^α maps $B_b(H)$ into $C_b^2(H)$ for any $t > 0$, and if $\varphi \in C_b(H)$, it holds that

$$\langle D(P_t^\alpha\varphi)(x), h\rangle_H = \frac{1}{t}\,\mathbf{E}\,\varphi(y_\alpha(t;x)) \int_0^t \left\langle Q^{-1}Dy_\alpha(s;x)h, dw(s)\right\rangle_H$$

for all $x, h \in H$. Moreover, for $i \leq j = 0, 1, 2$

$$\|P_t^\alpha\varphi\|_j^H \leq c_\alpha\,(t \wedge 1)^{-\frac{(j-i)(1+\epsilon)}{2}}\,\|\varphi\|_i^H. \tag{5.3}$$

Due to (2.8), by proceeding as in [6] it is possible to show that

$$\sup_{x \in H}\left(|Dy_\alpha(t;x)h|_H^2 + \int_0^t |(-A)^{1/2}Dy_\alpha(s;x)h|_H^2\right) \leq c_T|h|_H^2, \qquad \mathbf{P}-\text{a.s.}, \tag{5.4}$$

for a constant c_T independent of α. Thus if $j = 1$, for each $i = 0, 1$ we have

$$\|P_t^\alpha\varphi\|_1^H \leq c\,(t \wedge 1)^{-\frac{(1-i)(1+\epsilon)}{2}}\,\|\varphi\|_i^H, \tag{5.5}$$

and the constant c is independent of α.

From Lemma 3.3, we easily have that for any $\varphi \in C_b(H)$ it holds that

$$\lim_{\alpha \to 0} P_t^\alpha\varphi(x) = P_t\varphi(x), \tag{5.6}$$

uniformly with respect to $t \in [0, T]$ and x in bounded subsets of E. Moreover, from Lemma 4.1, we have that

$$\lim_{\alpha \to 0} |D(P_t^\alpha\varphi)(x) - D(P_t\varphi)(x)|_H = 0, \qquad t > 0, \tag{5.7}$$

uniformly for x in bounded sets of E. Actually, for each $\alpha > 0$ it holds that

$$\langle D(P_t^\alpha\varphi)(x), h\rangle_H = \frac{1}{t}\mathbf{E}\,\varphi(y_\alpha(t;x)) \int_0^t \left\langle Q^{-1}Dy_\alpha(s;x)h, dw(s)\right\rangle_H,$$

and then by easy calculations we obtain

$$\begin{aligned}
&|\langle D(P_t^\alpha\varphi)(x) - D(P_t\varphi)(x), h\rangle_H| \\
&\leq \frac{\|\varphi\|_1^H}{t}\left(\mathbf{E}\,|y_\alpha(t,x) - y(t;x)|_H^2\right)^{1/2}\left(\mathbf{E}\int_0^t |Q^{-1}Dy_\alpha(s;x)h|_H^2\,ds\right)^{1/2} \\
&+\frac{\|\varphi\|_1^H}{t}\left(\mathbf{E}\int_0^t |Q^{-1}(Dy_\alpha(s;x)h - Dy(s;x)h)|_H^2\,ds\right)^{1/2}.
\end{aligned}$$

Thus (5.7) follows from (3.6) and (4.6).

In correspondence of each $n \in \mathbb{N}$, we can introduce the transition semigroup $P_t^{\alpha,n}$ associated with the system (3.10). The semigroup $P_t^{\alpha,n}$ fulfills all the regularizing properties described above for P_t^α. In particular, due to (3.9) it is not difficult to check that for $i = 0, 1$

$$\|P_t^{\alpha,n}\|_1^H \leq c\,(t \wedge 1)^{-\frac{(1-i)(1+\epsilon)}{2}}\,\|\varphi\|_i^H, \qquad t > 0, \tag{5.8}$$

for a constant c which does not depend on n and α. In the next theorem we prove that it is possible to approximate $P_t^\alpha \varphi$ and its first derivative by means of $P_t^{\alpha,n}$ and its first derivative.

Proposition 5.1. *Under Hypotheses* 1 and 2, *for any $\varphi \in C_b(H)$ we have*

$$\lim_{n \to +\infty} P_t^{\alpha,n}\varphi(x) = P_t^\alpha \varphi(x), \tag{5.9}$$

uniformly for x in bounded sets of H and $t \in [0, T]$. Moreover,

$$\lim_{n \to +\infty} |D(P_t^{\alpha,n}\varphi)(x) - D(P_t^\alpha \varphi)(x)|_H = 0, \tag{5.10}$$

uniformly for x in bounded sets of H and $t \in [\delta, T]$, with $\delta > 0$.

Proof. The limit (5.9) follows directly from Lemma 3.4. As far as the limit (5.10) is concerned, we have

$$\begin{aligned}
&\langle D(P_t^\alpha \varphi)(x) - D(P_t^{\alpha,n}\varphi)(x), P_n h\rangle_H \\
&= \frac{1}{t}\mathbf{E}\left(\varphi(y_\alpha(t;x)) - \varphi(y_{\alpha,n}(t;x))\right)\int_0^t \left\langle Q^{-1} D y_\alpha(s;x) P_n h, dw(s)\right\rangle_H \\
&+ \frac{1}{t}\mathbf{E}\,\varphi(y_{\alpha,n}(t;x))\int_0^t \left\langle Q^{-1}(D y_\alpha(s;x) P_n h - D y_{\alpha,n}(s;x)h), dw(s)\right\rangle_H .
\end{aligned}$$

Thus we get

$$\begin{aligned}
&|\langle D(P_t^\alpha \varphi)(x) - D(P_t^{\alpha,n}\varphi)(x), P_n h\rangle_H|^2 \\
&\leq \frac{2}{t^2}\mathbf{E}\,|\varphi(y_\alpha(t;x)) - \varphi(y_{\alpha,n}(t;x))|^2\,\mathbf{E}\int_0^t |Q^{-1} D y_\alpha(s;x) P_n h|_H^2\,ds \\
&+ \frac{2}{t^2}\|\varphi\|_0^2\,\mathbf{E}\int_0^t |Q^{-1}(D y_\alpha(s;x) P_n h - Dx y_{\alpha,n}(s;x)h)|_H^2\,ds.
\end{aligned}$$

By taking the supremum over $|h|_H \leq 1$, due to (3.11), (4.6), and (5.4), it follows that

$$\lim_{n \to +\infty} |P_n D(P_t^\alpha \varphi)(x) - D(P_t^{\alpha,n}\varphi)(x)|_H = 0,$$

and, as

$$\lim_{n \to +\infty} |P_n D(P_t^\alpha \varphi)(x) - D(P_t^\alpha \varphi)(x)|_H = 0,$$

we obtain (5.10). □

6. The Hamilton–Jacobi–Bellman equation. We are here concerned with the infinite dimensional Cauchy problem

$$\frac{\partial u}{\partial t}(t,x) = \mathcal{L}u(t,x) - K(Du(t,x)) + g(x), \qquad u(0,x) = \varphi(x), \tag{6.1}$$

where $\mathcal{L}$ is the differential operator defined by

$$\mathcal{L}\psi(x) = \frac{1}{2}\mathrm{Tr}\,[Q^2 D^2\psi(x)] + \langle Ax + F(x), D\psi(x)\rangle_H\,, \qquad x \in D(A) \cap D(F).$$

In addition to Hypotheses 1 and 2, the following condition shall be assumed.

HYPOTHESIS 3. *The hamiltonian $K : H \to \mathbb{R}$ is Fréchet differentiable and locally Lipschitz continuous together with its derivative. Moreover, $K(0) = 0$.*

Notice that the requirement $K(0) = 0$ is not restrictive, as we can substitute g by $g - K(0)$.

The problem (6.1) can be rewritten in the *mild* form

$$u(t,x) = P_t\varphi(x) - \int_0^t P_{t-s}\left(K(Du(s,\cdot))\right)(x)\,ds + \int_0^t P_{t-s}g(x)\,ds. \tag{6.2}$$

As we noticed in the previous section, the semigroup P_t is not strongly continuous in $C_b(H)$ in general. Nevertheless, the mapping

$$[0,+\infty) \to \mathbb{R}, \qquad t \mapsto P_t\varphi(x)$$

is continuous for any fixed $\varphi \in C_b(H)$ and $x \in H$. Thus the integrals in the formula (6.2) have a meaning only for fixed $x \in H$.

We define $\mathcal{V}^1_T$ as the space of all continuous and bounded functions $u : [0,T]\times H \to \mathbb{R}$, such that $u(t,\cdot) \in C^1_b(H)$ for all $t \in (0,T]$, and the mapping

$$(0,T] \times H \to H, \qquad (t,x) \mapsto Du(t,x)$$

is bounded and measurable. It is easy to check that $\mathcal{V}^1_T$, endowed with the norm

$$\|u\|_{\mathcal{V}^1_T} = \sup_{t\in[0,T]} \|u(t,\cdot)\|_0^H + \sup_{t\in(0,T]} \|Du(t,\cdot)\|_0^H,$$

is a Banach space.

Moreover, we define $\mathcal{Z}^1_T$ as the space of bounded continuous functions $y : [0,T] \times H \to \mathbb{R}$, such that $y(t,\cdot) \in C^1_b(H)$ for all $t \in (0,T]$, and the mapping

$$(0,T] \times H \to H, \qquad (t,x) \mapsto t^{\frac{1+\epsilon}{2}} Dy(t,x)$$

is bounded and measurable. It is easy to check that $\mathcal{Z}^1_T$, endowed with the norm

$$\|u\|_{\mathcal{Z}^1_T} = \sup_{t\in[0,T]} \|y(t,\cdot)\|_0^H + \sup_{t\in(0,T]} (t\wedge 1)^{\frac{1+\epsilon}{2}} \|Dy(t,\cdot)\|_0^H,$$

is a Banach space.

Finally, we say that a function $y \in \mathcal{V}^1_T$ belongs to the space $\mathcal{Z}^2_T$ if $y(0,\cdot) \in C^1_b(H)$, the function $y(t,\cdot)$ is in $C^2_b(H)$ for any $t > 0$, and the mapping

$$(0,T] \times H \to \mathcal{L}(H), \qquad (t,x) \mapsto (t\wedge 1)^{\frac{1+\epsilon}{2}} D^2 y(t;x)$$

is bounded and measurable. $\mathcal{Z}^2_T$, endowed with the norm

$$\|u\|_{\mathcal{Z}^2_T} = \sup_{t\in\,[0,T]} \|y(t,\cdot)\|_1^H + \sup_{t\in\,(0,T]} (t\wedge 1)^{\frac{1+\epsilon}{2}} \|D^2 y(t,\cdot)\|_0^H,$$

is a Banach space.

A proof of the following lemma, in the case when $F = 0$, can be found in [20, Lemmas 4.8 and 4.12]. Such a proof completely adapts to our case where $F \neq 0$; thus we do not repeat it.

LEMMA 6.1. *Let us fix $T > 0$, and, for $\psi : [0,T]\times H \to \mathbb{R}$ bounded and measurable, let us define*

$$\lambda(\psi)(t,x) = \int_0^t P_{t-s}\psi(s,\cdot)(x)\,ds.$$

Then $\lambda(\psi)$ is continuous and bounded, $\lambda(\psi)(t,\cdot)\in C^1_b(H)$ for any $t\geq 0$, and

$$\sup_{t\in\,[0,T]} \|\lambda(\psi)(t,\cdot)\|_1^H < \infty.$$

It is immediate to check that Lemma 6.1 adapts to the approximating semigroups P^α_t and $P^{\alpha,n}_t$.

For each $\alpha > 0$, we consider the approximating problem

$$\frac{\partial u}{\partial t}(t,x) = \mathcal{L}_\alpha u(t,x) - K(Du(t,x)) + g(x), \qquad u(0,x) = \varphi(x),$$

where

$$\mathcal{L}_\alpha\psi(x) = \frac{1}{2}\mathrm{Tr}\,[Q^2 D^2\psi(x)] + \langle Ax + F_\alpha(x), D\psi(x)\rangle_H\,.$$

In mild form it can be rewritten as

$$u(t,x) = P^\alpha_t\varphi(x) - \int_0^t P^\alpha_{t-s}\left(K(Du(s,\cdot))\right)(x)\,ds + \int_0^t P^\alpha_{t-s} g(x)\,ds. \tag{6.3}$$

The first part of the following theorem was proved in [7], under the assumption of Lipschitz continuity for the hamiltonian K. Here the proof is more delicate, as K is only locally Lipschitz.

THEOREM 6.2. *Assume that Hypotheses* 1, 2, *and* 3 *hold, and fix $T > 0$. Then for any $\varphi, g\in \mathrm{Lip}_b(H)$,* (6.2) *admits a unique solution $u(t,x)$ in $\mathcal{V}^1_T$.*

If $u_\alpha(t,x)$ denotes the unique mild solution for the approximating problem (6.3), *we have*

$$\lim_{\alpha\to 0} |u_\alpha(t,x) - u(t,x)| + |Du_\alpha(t,x) - Du(t,x)|_H = 0, \tag{6.4}$$

uniformly for t in compact subsets of $(0,T]$ and for x in bounded subsets of E. Moreover, if $\varphi, g\in C^1_b(H)$, then the limit (6.4) *is uniform for $t\in[0,T]$ and for x in bounded subsets of E.*

We first prove some preliminary results.

LEMMA 6.3. *Fix $\varphi, g \in C^1_b(H)$ and $R\geq 2c_0\,\|\varphi\|_1^H$, where c_0 is the constant introduced in* (5.1). *Then the problem* (6.2) *admits a unique local solution $u(t,x)$ in $[0,\tau_R]$ for some constant*

$$\tau_R = \tau_R\left(\|\varphi\|_1^H, \|g\|_0^H, \|K\|_1^{B_R^H}\right).$$

Proof. For any $\tau > 0$, we define $\Lambda_R(\tau)$ as the set of all bounded continuous functions $u : [0,\tau] \times H \to \mathbb{R}$ such that $u(t,\cdot) \in C_b^1(H)$ for all $t \in [0,\tau]$, the mapping

$$[0,\tau] \times H \to H, \qquad (t,x) \mapsto Du(t,x)$$

is bounded and measurable, and

$$\sup_{t \in [0,\tau]} \|u(t,\cdot)\|_1^H \leq R.$$

We claim that for some τ_R sufficiently small, the operator Γ defined by

$$\Gamma(v)(t,x) = P_t\varphi(x) - \int_0^t P_{t-s}\left(K(Dv(s,\cdot))\right)(x)\,ds + \int_0^t P_{t-s}g(x)\,ds$$

maps $\Lambda_R(\tau_R)$ into itself as a contraction. Due to Lemma 6.1, $\Gamma(v)(t,x)$ is well defined for any x and t. Due to (5.1) we have

$$\|P_t\varphi\|_1^H \leq c_0\,\|\varphi\|_1^H \leq \frac{R}{2}.$$

Moreover, if we set

$$\Gamma_1(v)(t,x) = -\int_0^t P_{t-s}\left(K(Dv(s,\cdot))\right)(x)\,ds + \int_0^t P_{t-s}g(x)\,ds,$$

we have

$$\|\Gamma_1(v)(t,\cdot)\|_0^H \leq \int_0^t \|K(Dv(s,\cdot))\|_0^H\,ds + t\|g\|_0^H \leq \tau\Big(\sup_{x \in B_R^H} |K(x)| + \|g\|_0^H\Big).$$

Concerning the derivative, due to the estimate (5.1) it holds that

$$\|D(\Gamma_1(v))(t,\cdot)\|_0^H \leq c_0 \int_0^t (t-s)^{-\frac{1+\epsilon}{2}}(\|K(Dv(s,\cdot))\|_0^H + \|g\|_0^H)\,ds$$

$$\leq c_0\,\tau^{\frac{1-\epsilon}{2}}\Big(\sup_{x \in B_R^H} |K(x)| + \|g\|_0^H\Big).$$

This implies that

$$\sup_{t \in [0,\tau]} \|\Gamma(v)(t,\cdot)\|_1^H \leq \frac{R}{2} + c\left(\tau + \tau^{\frac{1-\epsilon}{2}}\right)\Big(\sup_{x \in B_R^H} |K(x)| + \|g\|_0^H\Big)$$

so that it is possible to find $\bar{\tau}_R$ sufficiently small such that

$$\sup_{t \in [0,\bar{\tau}_R]} \|\Gamma(v)(t,\cdot)\|_1^H \leq R.$$

In a completely analogous way it is possible to show that Γ is a contraction on $\Lambda_R(\tau_R)$ for some $\tau_R \leq \bar{\tau}_R$. This allows us to conclude that there exists a unique fixed point u for Γ in $\Lambda_R(\tau_R)$, which is the unique solution of (6.2) in $[0,\tau_R]$. □

Remark 6.4. In an identical way it is possible to prove that for each $\alpha > 0$ the mapping

$$\Gamma_\alpha(v)(t,x) = P_t^\alpha\varphi(x) - \int_0^t P_{t-s}^\alpha\left(K(Dv(s,\cdot))\right)(x)\,ds + \int_0^t P_{t-s}^\alpha g(x)\,ds$$

is a contraction in $\Lambda_R(\tau_R)$, where τ_R is the same as in Lemma 6.3. This implies that there exists a unique solution $u_\alpha(t,x)$ for the problem (6.3).

Moreover, it is useful to remark that thanks to (5.5) the contraction constant of Γ_α in $\Lambda_R(\tau_R)$ can be taken as the same for all $\alpha > 0$.

LEMMA 6.5. *If $u(t,x)$ and $u_\alpha(t,x)$ are, respectively, the solutions of the problems* (6.2) *and* (6.3) *with $\varphi, g \in C_b^1(H)$, we have*

$$\lim_{\alpha\to 0} |u_\alpha(t,x) - u(t,x)| + |Du_\alpha(t,x) - Du(t,x)|_H = 0, \tag{6.5}$$

uniformly for t in $[0,\tau_R]$ and for x in bounded subsets of E.

Proof. In order to prove the existence of the solutions $u(t,x)$ and $u_\alpha(t,x)$ for the problems (6.2) and (6.3), we have applied a contraction theorem. Hence, due to Lemma 6.3 and Remark 6.4, for each $\epsilon > 0$ there exists $k_\epsilon \in \mathbb{N}$ such that

$$\sup_{t\in\,[0,\tau_R]} \left(\|u_\alpha(t,\cdot) - \Gamma_\alpha^{k_\epsilon}(0)(t,\cdot)\|_1^H + \|u(t,\cdot) - \Gamma^{k_\epsilon}(0)(t,\cdot)\|_1^H \right) \leq \epsilon/2 \tag{6.6}$$

for each $\alpha > 0$. Now, from Proposition 5.1, by using an induction argument we can prove that for each $k \in \mathbb{N}$

$$\lim_{\alpha\to 0} \left|\Gamma_\alpha^k(0)(t,x) - \Gamma^k(0)(t,x)\right| + \left|D\Gamma_\alpha^k(0)(t,x) - D\Gamma^k(0)(t,x)\right|_H = 0, \tag{6.7}$$

uniformly for (t,x) in bounded subsets of $[0,\tau_R]\times E$. Actually, for $k = 1$, (6.7) follows directly from (5.6) and (5.7). Now assume that (6.7) holds for some $k \geq 1$. We have

$$\Gamma_\alpha^{k+1}(0)(t,x) - \Gamma^{k+1}(0)(t,x) = \Gamma_\alpha(\Gamma_\alpha^k(0))(t,x) - \Gamma(\Gamma^k(0))(t,x)$$

$$= P_t^\alpha\varphi(x) - P_t\varphi(x) + \int_0^t \left(P_{t-s}^\alpha g(x) - P_{t-s}g(x)\right) ds$$

$$- \int_0^t \left(P_{t-s}^\alpha \left[K(D(\Gamma_\alpha^k(0))(s,\cdot))\right](x) - P_{t-s}\left[K(D(\Gamma^k(0))(s,\cdot))\right](x)\right) ds.$$

Since $\Gamma_\alpha^k(0)$ and $\Gamma^k(0)$ belong to $\Lambda_R(\tau_R)$ and (6.7) holds for k, by using (5.6) and the boundedness of K on bounded subsets of H, from the dominated convergence theorem it follows that

$$\lim_{\alpha\to 0} \int_0^t \left(P_{t-s}^\alpha \left[K(D(\Gamma_\alpha^k(0))(s,\cdot))\right](x) - P_{t-s}\left[K(D(\Gamma^k(0))(s,\cdot))\right](x)\right) ds = 0,$$

uniformly on bounded sets of $[0,\tau_R]\times E$. By using (5.6) once more, we have

$$\lim_{\alpha\to 0} P_t^\alpha\varphi(x) - P_t\varphi(x) + \int_0^t \left(P_{t-s}^\alpha g(x) - P_{t-s}g(x)\right) ds = 0,$$

uniformly on bounded sets of $[0,\tau_R]\times E$, so that we get

$$\lim_{\alpha\to 0} \Gamma_\alpha^{k+1}(0)(t,x) - \Gamma^{k+1}(0)(t,x) = 0.$$

The second part of the limit (6.7) for $k+1$ follows by analogous arguments. By induction we can conclude that (6.7) holds for any $k \in \mathbb{N}$.

Now, from (6.6) we have that

$$|u_\alpha(t,x)-u(t,x)|+|Du_\alpha(t,x)-Du(t,x)|_H$$

$$\leq \sup_{t\in\,[0,\tau_R]}\|u_\alpha(t,\cdot)-\Gamma^{k_\epsilon}_\alpha(0)(t,\cdot)\|_1^H+|\Gamma^{k_\epsilon}_\alpha(0)(t,x)-\Gamma^{k_\epsilon}(0)(t,x)|$$

$$+|D\Gamma^{k_\epsilon}_\alpha(0)(t,x)-D\Gamma^{k_\epsilon}(0)(t,x)|_H+\sup_{t\in\,[0,\tau_R]}\|u(t,\cdot)-\Gamma^{k_\epsilon}(0)(t,\cdot)\|_1^H$$

$$\leq \epsilon/2+|\Gamma^{k_\epsilon}_\alpha(0)(t,x)-\Gamma^{k_\epsilon}(0)(t,x)|+\left|D\Gamma^{k_\epsilon}_\alpha(0)(t,x)-D\Gamma^{k_\epsilon}(0)(t,x)\right|_H,$$

and due to (6.7) we can conclude that (6.5) holds. □

Proof of Theorem 6.2. Let us fix $T>0$ and $\varphi,g\in C^1_b(H)$, and let us define

$$R=2c\,(1+T)\,e^{c\,T}\left((1+2c_0)\|\varphi\|_1^H+\|g\|_1^H\right).$$

Due to Lemma 6.3, there exists a mild solution $u(t,x)$ defined for $t\in\,[0,\tau_R]$. Moreover, from Lemma 6.5 we have that (6.4) holds, uniformly with respect to $t\in\,[0,\tau_R]$ and x in bounded sets of E.

From Proposition A.3, we have that

$$\sup_{t\in\,[0,\tau_\star]}\|u_\alpha(t,\cdot)\|_1^H\leq c\,(1+T)e^{c\,T}\left(\|\varphi\|_1^H+\|g\|_1^H\right).$$

According to Lemma 6.5, this implies that for any $t\in\,[0,\tau_\star]$ and $x\in\,E$

$$|u(t,x)|+|Du(t,x)|_H\leq c\,(1+T)e^{c\,T}\left(\|\varphi\|_1^H+\|g\|_1^H\right),$$

and, since $u(t,\cdot)\in\,C^1_b(H)$ for $t\in\,[0,\tau_\star]$, we have

$$\sup_{t\in\,[0,\tau_\star]}\|u(t,\cdot)\|_1^H\leq c\,(1+T)e^{c\,T}\left(\|\varphi\|_1^H+\|g\|_1^H\right). \tag{6.8}$$

In particular, due to the definition of R we have that

$$\|u(\tau_\star,\cdot)\|_1^H\leq\frac{R}{2}.$$

This allows us to repeat all of the same arguments we have been using until now in the intervals $[\tau_\star,2\tau_\star]$, $[2\tau_\star,3\tau_\star]$, and so on, up to time T, and hence to get a global solution.

Now, assume that $\varphi,g\in\,\mathrm{Lip}_b(H)$. It is possible to find two bounded sequences $\{\varphi_k\}$ and $\{g_k\}$ in $C^1_b(H)$ converging, respectively, to φ and g in $C_b(H)$. In correspondence with each k, there exists a unique solution $u_k(t,x)$ to the problem

$$u_k(t,x)=P_t\varphi_k(x)-\int_0^t P_{t-s}\left(K(Du_k(s,\cdot))\right)(x)\,ds+\int_0^t P_{t-s}g_k(x)\,ds.$$

Our aim is to show that $\{u_k\}$ is a Cauchy sequence in $\mathcal{Z}^1_T$ and that the limit u fulfills (6.2).

For each $k,n\in\,\mathbb{N}$ we have

$$u_k(t,x)-u_h(t,x)=P_t\left(\varphi_k-\varphi_h\right)(x)$$

$$-\int_0^t P_{t-s}\left(K(Du_k(s,\cdot))-K(Du_h(s,\cdot))\right)(x)\,ds+\int_0^t P_{t-s}\left(g_k-g_h\right)(x)\,ds.$$

Due to (6.8) we easily have

$$\sup_{k\in\mathbb{N}}\sup_{t\in[0,T]}\|Du_k(t,\cdot)\|_1^H\leq c\,(1+T)e^{c\,T}\sup_{k\in\mathbb{N}}\left(\|\varphi_k\|_1+\|g_k\|_1^H\right)=c_T. \tag{6.9}$$

If M is the Lipschitz constant of K in $B^H_{c_T}$, we have

$$\begin{aligned}&\|u_k(t,\cdot)-u_h(t,\cdot)\|_0^H\leq\|\varphi_k-\varphi_h\|_0^H\\&+M\int_0^t\|Du_k(s,\cdot)-Du_h(s,\cdot)\|_0^H\,ds+t\,\|g_k-g_h\|_0.\end{aligned} \tag{6.10}$$

Moreover, we have

$$\begin{aligned}&Du_k(t,x)-Du_h(t,x)=DP_t(\varphi_k-\varphi_h)(x)\\&-\int_0^t DP_{t-s}\left(K(Du_k(s,\cdot))-K(Du_h(s,\cdot))\right)(x)\,ds+\int_0^t DP_{t-s}(g_k-g_h)(x)\,ds,\end{aligned}$$

so that, thanks to (5.1), we get

$$\begin{aligned}&\|Du_k(t,\cdot)-Du_h(t,\cdot)\|_0^H\leq c\,t^{-\frac{1+\epsilon}{2}}\|\varphi_k-\varphi_h\|_0^H\\&+c\,M\int_0^t(t-s)^{-\frac{1+\epsilon}{2}}\|Du_k(s,\cdot)-Du_h(s,\cdot)\|_0^H\,ds+c\int_0^t(t-s)^{-\frac{1+\epsilon}{2}}\,ds\|g_k-g_h\|_0^H.\end{aligned}$$

This implies that

$$\begin{aligned}&t^{\frac{1+\epsilon}{2}}\|Du_k(t,\cdot)-Du_h(t,\cdot)\|_0^H\leq c\,\|\varphi_k-\varphi_h\|_0^H\\&+c\,Mt^{\frac{1+\epsilon}{2}}\int_0^t(t-s)^{-\frac{1+\epsilon}{2}}\|Du_k(s,\cdot)-Du_h(s,\cdot)\|_0^H\,ds+c\,t\,\|g_k-g_h\|_0^H.\end{aligned} \tag{6.11}$$

By combining (6.10) and (6.11), we conclude that

$$\begin{aligned}&\|u_k(t,\cdot)-u_h(t,\cdot)\|_0^H+t^{\frac{1+\epsilon}{2}}\|Du_k(t,\cdot)-Du_h(t,\cdot)\|_0^H\\&\leq c\left(\|\varphi_k-\varphi_h\|_0^H+T\|g_k-g_h\|_0^H\right)+M(1+c\,T^{\frac{1+\epsilon}{2}})\int_0^t s^{-\frac{1+\epsilon}{2}}\left((t-s)^{-\frac{1+\epsilon}{2}}+1\right)\\&\left(\|u_k(s,\cdot)-u_h(s,\cdot)\|_0^H+s^{\frac{1+\epsilon}{2}}\|Du_k(s,\cdot)-Du_h(s,\cdot)\|_0^H\right)\,ds.\end{aligned}$$

Thus, from a generalization of the Gronwall lemma, we can say that

$$\|u_k-u_h\|_{\mathcal{Z}^1_T}\leq c_T\left(\|\varphi_k-\varphi_h\|_0^H+T\|g_k-g_h\|_0^H\right) \tag{6.12}$$

for some constant c_T independent of k and h. This implies that $\{u_k\}$ is a Cauchy sequence in $\mathcal{Z}^1_T$, and hence it converges to a limit $u\in\mathcal{Z}^1_T$. Moreover, from (6.9) we have that

$$\sup_{t\in[0,T]}\|Du(t,\cdot)\|_0^H<+\infty,$$

so that $u \in \mathcal{V}_T^1$.

Now, we show that u is the mild solution of the problem (6.2). Actually, for any $s > 0$ and $x \in H$

$$\lim_{k \to +\infty} K(Du_k(s,x)) = K(Du(s,x)).$$

Due to (6.9) we can apply the dominated convergence theorem, and we get

$$\lim_{k \to +\infty} \int_0^t P_{t-s} K(Du_k(s,\cdot))(x)\,ds = \int_0^t P_{t-s} K(Du(s,\cdot))(x)\,ds.$$

Therefore, since

$$\lim_{k \to +\infty} P_t \varphi_k(x) = P_t \varphi(x)$$

and

$$\lim_{k \to +\infty} \int_0^t P_{t-s} g_k(x)\,ds = \int_0^t P_{t-s} g(x)\,ds,$$

we conclude that u is a solution of (6.2).

Finally, uniqueness follows from the Gronwall lemma and local Lipschitzianity of K. Indeed, if u_1 and u_2 are two solutions in $\mathcal{V}_T^1$, we have

$$u_1(t,x) - u_2(t,x) = -\int_0^t P_{t-s} \left(K(Du_1(s,\cdot)) - K(Du_2(s,\cdot)) \right)(x)\,ds,$$

and then, if M is the Lipschitz constant of K in $B^H_{c_T}$, where

$$c_T = \|u_1\|_{\mathcal{V}_T^1} + \|u_2\|_{\mathcal{V}_T^1},$$

we have

$$\|u_1 - u_2\|_{\mathcal{V}_T^1} \leq M \int_0^t \left(1 + (t-s)^{-\frac{1+\epsilon}{2}} \right) ds \|u_1 - u_2\|_{\mathcal{V}_T^1}.$$

This implies that $u_1 = u_2$. □

7. Application to control. We apply here the results proved in the previous section to a stochastic control problem. Let $k : H \to]-\infty, +\infty]$ be a convex lower semicontinuous function, and let K be its Legendre transform; that is,

$$K(x) = \sup_{z \in H} \left\{ -\langle x, z \rangle_H - k(z) \right\}, \qquad x \in H.$$

We assume that k is such that K fulfills Hypothesis 3. We consider here the cost functional

$$J(t,x;z) = \mathbf{E} \int_t^T \left(g(y(s)) + k(z(s)) \right) ds + \mathbf{E}\, \varphi(y(T)),$$

where $y(s) = y(s,t;x,z)$ is the unique solution of the controlled system (1.1) at time s, starting from x at time t. We want to minimize the functional J over all adapted controls $z \in L^2(\Omega; L^2([0,T];H))$.

The *value function* corresponding to the cost functional J is defined by

$$V(t,x) = \inf \left\{ J(t,x;z) \,:\, z \in L^2(\Omega; L^2([0,T];H)) \text{ adapted} \right\}$$

and is related to the Hamilton–Jacobi–Bellman equation (6.1). Namely, we are showing that for every $t \in [0,T]$ and $x \in H$

$$V(t,x) = u(T-t,x),$$

where $u(t,x)$ is the unique mild solution of the problem (6.1).

For any $\alpha > 0$ we introduce the approximating cost functional

$$J_\alpha(t,x;z) = \mathbf{E} \int_t^T (g(y_\alpha(s)) + k(z(s)))\, ds + \mathbf{E}\, \varphi(y_\alpha(T)), \tag{7.1}$$

where $y_\alpha(s) = y_\alpha(s,t;x,z)$ is the unique solution to the problem (3.5). In what follows we will denote by $V_\alpha(t,x)$ the corresponding value function.

LEMMA 7.1. *Assume Hypotheses* 1, 2, *and* 3, *and assume that* $\varphi, g \in \mathrm{Lip}_b(H)$. *If* u *is the mild solution of the problem* (6.2), *for any control* $z \in L^2(\Omega; L^2([0,T];H))$, $x \in H$, *and* $t \in [0,T]$, *the following identity holds:*

$$\begin{aligned} &J(t,x;z) = u(T-t,x) \\ &+ \int_t^T \mathbf{E}\, (K(Du(T-s,y(s))) + \langle z(s), Du(T-s,y(s))\rangle_H + k(z(s)))\, ds, \end{aligned} \tag{7.2}$$

where $y(s) = y(s,t;x,z)$ *is the solution of the problem* (3.1).

Moreover, the same identity holds with $J(t,x;z)$, $u(t,x)$, $Du(t,x)$, *and* $y(t)$ *replaced, respectively, by* $J_\alpha(t,x;z)$, $u_\alpha(t,x)$, $Du_\alpha(t,x)$, *and* $y_\alpha(t)$.

Proof. We first assume that $\varphi, g \in C_b^1(H)$. Let $u_{\alpha,n}(t,x)$ be the solution of (A.2), and let $y_{\alpha,n}(s) = y_{\alpha,n}(s,t;x,z)$ be the solution to the problem (3.10). Since $u_{\alpha,n}$ is smooth (in fact, $u_{\alpha,n} \in Z^2(T)$) and $y_{\alpha,n}$ is a strong solution, we can apply Itô's formula to the function $s \mapsto u_{\alpha,n}(T-s, y_{\alpha,n}(s))$ for $t \leq s \leq T$, and we get

$$\begin{aligned} du_{\alpha,n}(T-s,y_{\alpha,n}(s)) &= \langle dy_{\alpha,n}(s), Du_{\alpha,n}(T-s,y_{\alpha,n}(s))\rangle_H \\ &+ \left(\frac{1}{2} \mathrm{Tr}\, \left[Q_n^2 D^2 u_{\alpha,n}(T-s,y_{\alpha,n}(s)) \right] - \frac{\partial u_{\alpha,n}}{\partial t}(T-s,y_{\alpha,n}(s)) \right) ds. \end{aligned}$$

By integrating with respect to $s \in [t,T]$ and by taking the expectation, we get

$$\begin{aligned} \mathbf{E}\, \varphi(y_{\alpha,n}(T)) - u_{\alpha,n}(T-t,x) &= \mathbf{E} \int_t^T (K(Du_{\alpha,n}(T-s,y_{\alpha,n}(s))) \\ &+ \langle z(s), Du_{\alpha,n}(T-s,y_{\alpha,n}(s))\rangle_H - g(y_{\alpha,n}(s)))\, ds. \end{aligned} \tag{7.3}$$

Now, due to Lemma 3.4 and (A.3), we can take the limit as n goes to $+\infty$ in each side of (7.3), and, rearranging all terms, we get

$$\begin{aligned} &\mathbf{E}\, \varphi(y_\alpha(T)) + \mathbf{E} \int_t^T g(y_\alpha(s))\, ds \\ &= u_\alpha(T-t,x) + \mathbf{E} \int_t^T (K(Du_\alpha(T-s,y_\alpha(s))) + \langle z(s), Du_\alpha(T-s,y_\alpha(s))\rangle_H)\, ds. \end{aligned} \tag{7.4}$$

This implies (7.2).

Now, let $\varphi, g \in \mathrm{Lip}_b(H)$. As in the proof of Theorem 6.2, let $\{\varphi_k\}$ and $\{g_k\}$ be two bounded sequences in $C^1_b(H)$ converging, respectively, to φ and g in $C_b(H)$. If we denote by $u^k_\alpha(t,x)$ the solutions of the problem (6.2) corresponding to φ_k and g_k, we have

$$\mathbf{E}\,\varphi_k(y_\alpha(T)) + \mathbf{E}\int_t^T g_k(y_\alpha(s))\,ds = u^k_\alpha(T-t,x)$$

$$+\mathbf{E}\int_t^T \left(K(Du^k_\alpha(T-s,y_\alpha(s))) + \left\langle z(s), Du^k_\alpha(T-s,y_\alpha(s))\right\rangle_H\right)\,ds.$$

It is immediate to check that the sequence $\{u^k_\alpha\}$ fulfills an estimate analogous to (6.12), and then the sequence $\{u^k_\alpha\}$ converges to u_α in $\mathcal{Z}^1_T$, as k goes to infinity. Moreover, due to (2.19), (3.3), and (6.9), we can apply the dominated convergence theorem and, by taking the limit for k going to $+\infty$, we get (7.4) for any $\varphi, g \in \mathrm{Lip}_b(H)$.

Now, if $x \in E$, then $y_\alpha(s) \in E$ and (4.4) holds. Thus, due to (3.11) and (6.4), we can take the limit as α goes to zero in each side of (7.4), and we get (7.2) for $x \in E$. Finally, if $x \in H$, we fix a sequence $\{x_n\} \subset E$ converging to x in H. Thanks to (3.4) we have that $y(s,t;x_n,z)$ converges to $y(s,t;x,z)$ in H, and then, as $u(t,\cdot) \in C^1_b(H)$, we easily get (7.2) for any $x \in H$. □

Now we can conclude by giving the main result of this section.

THEOREM 7.2. *Under Hypotheses* 1, 2, *and* 3, *for any* $\varphi, g \in \mathrm{Lip}_b(H)$ *the value function* $V(t,x)$ *coincides with* $u(T-t,x)$, *where* $u(t,x)$ *is the solution of the problem* (6.2). *Moreover,*

$$V(t,x) = \lim_{\alpha\to 0}\, \min\,\left\{ J_\alpha(t,x;z),\ z \in L^2(\Omega; L^2(0,T;H)),\ \textit{adapted}\right\},$$

where $J_\alpha(t,x;z)$ *is the cost functional defined in* (7.1).

Proof. From (7.2) we immediately have that $J(t,x;z) \geq u(T-t,x)$ for any $z \in L^2(\Omega; L^2(0,T;H))$, so that $V(t,x) \geq u(T-t,x)$. Now we prove the opposite inequality.

Since J_α fulfills a formula analogous to (7.2), we have $V_\alpha(t,x) \geq u_\alpha(T-t,x)$. In fact, it holds that $V_\alpha(t,x) = u_\alpha(T-t,x)$. Actually, by a general property of the Legendre transform, for each $t \in [0,T]$ the mapping

$$H \to \mathbb{R}, \qquad z \mapsto \langle z, Du_\alpha(T-t,y(t))\rangle_H + k(z),$$

attains its maximum for

$$z_\alpha(t) = -DK(Du_\alpha(T-t,y(t))), \qquad t \in [0,T].$$

Thus, if we prove that the *closed loop* equation

(7.5)
$$dy(t) = (Ay(t) + F_\alpha(y(t)) - DK(Du_\alpha(T-t,y(t))))\,dt + Q\,dw(t), \qquad y(0) = x,$$

admits a unique solution $y^\star_\alpha(t)$, and if we define

$$z^\star_\alpha(t) = -DK(Du_\alpha(T-t,y^\star_\alpha(t))),$$

due to (7.2) for J_α we have that $J_\alpha(t,x;z^\star_\alpha) = u(T-t,x)$, so that $y^\star_\alpha(t)$ and $z^\star_\alpha(t)$ are, respectively, the unique optimal state and the unique optimal control for the minimizing problem corresponding to the functional J_α.

Assume that $\varphi, g \in C_b^1(H)$. Due to (5.3) it is possible to show that the solution u_α of the problem (6.3) belongs to $Z^2(T)$ and

$$\|u_\alpha\|_{Z^2(T)} \leq c_{\alpha,T} \left(\|\varphi\|_1^H + T\,\|g\|_1^H\right). \tag{7.6}$$

Thus, if we define for $(t,x) \in [0,T] \times H$

$$U_\alpha(t,x) = -DK(Du_\alpha(t,x)),$$

we have that the function U_α fulfills the conditions of Lemma A.1, so that there exists a unique solution $y_\alpha^\star(t)$ for the closed loop equation (7.5).

Now, assume that $\varphi, g \in \mathrm{Lip}_b(H)$. As in the proofs of Theorem 6.2 and of Lemma 7.1, we approximate them in $C_b(H)$ by two bounded sequences $\{\varphi_k\}$ and $\{g_k\}$ in $C_b^1(H)$. For each k there exists a unique solution $u_{\alpha,k}$ for the problem (6.3), with data φ_k and g_k. Thus, as proved above, in correspondence with each $u_{\alpha,k}$ there exists a unique solution $y_{\alpha,k}^\star(t)$ for the problem (7.5). Let us define $v_{h,k}^\alpha(t) = y_{\alpha,k}^\star(t) - y_{\alpha,h}^\star(t)$. We have that $v_{h,k}^\alpha(t)$ is the unique solution of the problem

$$\begin{aligned}\frac{dv}{dt}(t) = \quad & Av(t) + F_\alpha(y_{\alpha,k}^\star(t)) - F_\alpha(y_{\alpha,h}^\star(t)) - DK(Du_{\alpha,k}(T-t, y_{\alpha,k}^\star(t))) \\ & +DK(Du_{\alpha,h}(T-t, y_{\alpha,h}^\star(t))), \qquad v(0) = 0.\end{aligned}$$

Thus, by multiplying each side by $v_{h,k}^\alpha(t)$, due to the Lipschitz continuity of F_α and (2.13) we get

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}|v_{h,k}^\alpha(t)|_H^2 \leq c_\alpha\,|v_{h,k}^\alpha(t)|_H^2 \\ &+|DK(Du_{\alpha,k}(T-t, y_{\alpha,k}^\star(t))) - DK(Du_{\alpha,h}(T-t, y_{\alpha,h}^\star(t)))|_H |v_{h,k}^\alpha(t)|_H.\end{aligned} \tag{7.7}$$

Since Proposition A.3 holds, the sequences $\{\varphi_k\}$ and $\{g_k\}$ are bounded in $C_b^1(H)$, and DK is locally Lipschitz continuous, there exists $c > 0$ such that

$$\begin{aligned}&|DK(Du_{\alpha,k}(T-t, y_{\alpha,k}^\star(t))) - DK(Du_{\alpha,h}(T-t, y_{\alpha,h}^\star(t)))|_H \\ &\leq c\,|Du_{\alpha,k}(T-t, y_{\alpha,k}^\star(t)) - Du_{\alpha,h}(T-t, y_{\alpha,h}^\star(t))|_H.\end{aligned}$$

Now, for any $t > 0$ and $x, y \in H$, due to (7.6) we have

$$\begin{aligned}&|Du_{\alpha,k}(t,x) - Du_{\alpha,k}(t,y)|_H \\ &\leq c_{\alpha,T} t^{-\frac{1+\epsilon}{2}} \left(\|\varphi_k\|_1^H + T\,\|g_k\|_1^H\right) |x-y|_H \leq c_{\alpha,T} t^{-\frac{1+\epsilon}{2}} |x-y|_H\end{aligned}$$

for some constant independent of k. Moreover, we can repeat all arguments used in the proof of Theorem 6.2, and we have

$$|Du_{\alpha,k}(t,y) - Du_{\alpha,h}(t,y)|_H \leq c_{\alpha,T} t^{-\frac{1+\epsilon}{2}} \left(\|\varphi_k - \varphi_h\|_0^H + T\|g_k - g_h\|_0^H\right).$$

Therefore, we get

$$\begin{aligned}&\left|K(Du_{\alpha,k}(T-t, y_{\alpha,k}^\star(t))) - DK(Du_{\alpha,h}(T-t, y_{\alpha,h}^\star(t)))\right|_H \\ &\leq c_{\alpha,T}(T-t)^{-\frac{1+\epsilon}{2}} \left(|v_{h,k}^\alpha(t)|_H + \|\varphi_k - \varphi_h\|_0^H + T\|g_k - g_h\|_0^H\right),\end{aligned}$$

so that, from (7.7) we conclude

$$\frac{d}{dt}|v^\alpha_{h,k}(t)|^2_H \leq c_{\alpha,T}\left(1+(T-t)^{-\frac{1+\epsilon}{2}}\right)|v^\alpha_{h,k}(t)|^2_H$$

$$+c_{\alpha,T}(T-t)^{-\frac{1+\epsilon}{2}}\left(\|\varphi_k-\varphi_h\|^H_0+T\|g_k-g_h\|^H_0\right)^2.$$

Due to the Gronwall lemma this yields

$$|y^\star_{\alpha,k}(t)-y^\star_{\alpha,h}(t)|^2_H \leq c_{\alpha,T}\left(\|\varphi_k-\varphi_h\|^H_0+T\|g_k-g_h\|^H_0\right)^2,$$

and the sequence $\{y^\star_{\alpha,k}\}$ converges to some $y^\star_\alpha$ in $C([0,T];H)$, $\mathbf{P}$-a.s. and in mean-square, and clearly $y^\star_\alpha$ is the unique solution of the closed loop (7.5).

Since $z^\star_\alpha(t)=-DK(Du_\alpha(T-t,y^\star_\alpha(t)))$, then due to (7.6) there exists a constant R such that

$$\sup_{\alpha>0}\sup_{t\in[0,T]}|z^\star_\alpha(t)|_H=R,\qquad \mathbf{P}-\text{a.s.}$$

This means that if we define the set $\mathcal{M}_R$ as in the Lemma 3.3, then for any $\alpha>0$

$$V_\alpha(t,x)=\inf\{\,J_\alpha(t,x;z)\;;\;z\in\mathcal{M}_R\,\}. \tag{7.8}$$

By using (6.4) we have that if $\varphi,g\in C^1_b(H)$, then for any $t\in[0,T]$ and $x\in E$

$$\lim_{\alpha\to 0}V_\alpha(t,x)=u(T-t,x).$$

Moreover,

$$|J_\alpha(t,x;z)-J(t,x;z)|\leq\|\varphi\|^H_1\mathbf{E}\,|y_\alpha(T)-y(T)|_E+\|g\|^H_1\int_0^t\mathbf{E}\,|y_\alpha(s)-y(s)|_E\,ds,$$

so that, due to Lemma (3.3) we have

$$\lim_{\alpha\to 0}\sup_{z\in\mathcal{M}_R}|J_\alpha(t,x;z)-J(t,x;z)|=0.$$

Due to (7.8) we have that

$$u(T-t,x)=\lim_{\alpha\to 0}V_\alpha(t,x)=\inf\{\,J(t,x;z)\;;\;z\in\mathcal{M}_R\,\}\geq V(t,x),$$

and since $u(T-t,x)\leq V(t,x)$, we conclude that $u(T-t,x)=V(t,x)$ for $\varphi,g\in C^1_b(H)$ and $x\in E$.

Now, if $x\in H$ and $\{x_n\}$ is a sequence in E converging to x in H, by using (3.4) we can prove that

$$\lim_{n\to+\infty}\sup_{z\in\mathcal{M}_R}|J(t,x_n;z)-J(t,x;z)|=0.$$

Therefore, since $u(t,x_n)$ converges to $u(t,x)$, we get the theorem for any $x\in H$. Finally, if $\varphi,g\in\text{Lip}_b(H)$, let $\{\varphi_k\}$ and $\{g_k\}$ be two bounded sequences in $C^1_b(H)$, converging, respectively, to φ and g in $C_b(H)$. We have

$$\lim_{k\to+\infty}\mathbf{E}\,\varphi_k(y(T))+\mathbf{E}\int_0^t(g_k(y(s))+k(z(s)))\,ds=J(t,x;z),$$

uniformly with respect to z, and then, thanks to (6.12), the theorem holds for any $\varphi, g \in \mathrm{Lip}_b(H)$. □

In some particular cases the closed loop equation admits a unique solution, and then there exist a unique optimal control and a unique state for the control problem.

THEOREM 7.3. *Assume the hypotheses of Theorem 7.2, and take the space dimension $d = 1$.*

1. *If the constant m in Hypothesis 1 is less than 2, then for any $\varphi, g \in Lip_b(H)$, and $x \in H$ there exists a unique optimal control for the minimizing problem associated with the functional J. Furthermore, the optimal control $z^\star$ is related to the corresponding optimal state $y^\star$ by the feedback formula*
$$z^\star(t) = -DK(D_x V(T-t, y^\star(t))), \qquad t \in [0,T].$$
2. *If DK can be extended as a Lipschitz continuous mapping from $E^\star$ into itself, then the same conclusion of 1 holds for any $x \in E$.*

Proof. We first prove 1. As we have seen in the proof of the previous theorem, the only thing we have to show is that for any $\varphi, g \in C_b^1(H)$ the derivative with respect to x of the solution u of the problem (6.1) is Lipschitz continuous, and for any $x, y \in H$

$$|Du(t,x) - Du(t,y)|_H \leq c_T\,(t \wedge 1)^{-\frac{1+\epsilon}{2}}\,\left(\|\varphi\|_1^H + T\,\|g\|_1^H\right)|x-y|_H. \tag{7.9}$$

Actually, if we define for $(t,x) \in [0,T] \times H$
$$U(t,x) = -DK(Du(T-t,x)),$$
the function U verifies the conditions of Lemma A.1, and then there exists a unique solution $y^\star(t)$ for the closed loop equation. Thanks to Lemma 7.1 this implies the existence of a unique optimal control and state. Finally, as in the proof of the previous theorem, the general case of $\varphi, g \in \mathrm{Lip}_b(H)$ follows by approximation.

We have seen that u is the unique solution of (6.2) in $C_b^1(H)$ and
$$\|u\|_1^H \leq c_T\,\left(\|\varphi\|_1^H + T\,\|g\|_1^H\right).$$
Clearly, if we show that the function $D(P_t\varphi)$ is Lipschitz continuous for any $\varphi \in C_b^1(H)$ and $t > 0$ and

$$|D(P_t\varphi)(x) - D(P_t\varphi)(y)|_H \leq c\,(t \wedge 1)^{-\frac{1+\epsilon}{2}}\,\|\varphi\|_1^H|x-y|_H, \tag{7.10}$$

then, by using the same arguments of section 6, we have that $u(t,\cdot) \in C_b^1(H)$ and
$$|Du(t,x) - Du(t,y)|_H \leq c\,(t \wedge 1)^{-\frac{1+\epsilon}{2}}\,|x-y|_H,$$
where the constant c depends only on g, φ, and T. Since $D(P_t\varphi)$ is given by the formula (5.2), (7.10) immediately follows once one proves that for any $x, y \in E$ it holds that

$$|v(t;x,h) - v(t;y,h)|_H^2 + \int_0^t |Q^{-1}v(s;x,h) - v(s;y,h)|_H^2 \leq c_T|h|_H^2|x-y|_H^2, \tag{7.11}$$

P-a.s. Let us define $z(t) = v(t;x,h) - v(t;y,h)$. We have that z is the unique solution of the problem

$$\left\{\begin{array}{ll} \dfrac{dz}{dt}(t) = & (Az(t) + DF(y(t;x))z(t))\,dt \\[2ex] & + \left(DF(y(t;x)) - DF(y(t;y))\right)v(t;y,h)\,dt, \qquad z(0) = 0. \end{array}\right.$$

Thus we have

$$\frac{1}{2}\frac{d}{dt}|z(t)|_H^2 + |z(t)|^2_{D((-A)^{1/2})} \leq c\,|z(t)|_H^2$$

$$+\,|\langle (DF(y(t;x)) - DF(y(t;y)))\,v(t;y,h), z(t)\rangle_H|\,.$$

Due to the Sobolev embedding theorem, for any $\delta > 0$ we have

$$|\langle (DF(y(t;x)) - DF(y(t;y)))\,v(t;y,h), z(t)\rangle_H|$$

$$\leq |z(t)|_{D\left((-A)^{(1+\delta)/4}\right)}\,|(DF(y(t;x)) - DF(y(t;y)))\,v(t;y,h)|_1\,.$$

In [6] it is proved that

$$\sup_{x\in\,E} |y(t;x)|_E \leq k(t)(t\wedge 1)^{-\frac{1}{2m}}, \qquad \mathbf{P}-\text{a.s.},$$

for some process $k(t)$ having all moments finite. Thus, since

$$\sup_{x\in\,H} |v(t;x,h)|_H \leq c(t)|h|_H,$$

for some continuous increasing function $c(t)$, by interpolation we get

$$|\langle (DF(y(t;x)) - DF(y(t;y)))\,v(t;y,h), z(t)\rangle_H|$$

$$\leq |z(t)|^{(1+\delta)/2}_{D\left((-A)^{1/2}\right)}|z(t)|_H^{(1-\delta)/2}|x-y|_H|h|_H\,c(t)\,(t\wedge 1)^{-(2m-1)/2m}.$$

As we can write

$$(t\wedge 1)^{-(2m-1)/2m} = (t\wedge 1)^{-(1-\delta)/2}(t\wedge 1)^{-\frac{1}{2}(1+\delta-1/m)},$$

thanks to the Young inequality we get

$$\frac{1}{2}\frac{d}{dt}|z(t)|_H^2 + |z(t)|^2_{D((-A)^{1/2})} \leq c\,|z(t)|_H^2 + \frac{1}{2}|z(t)|^2_{D\left((-A)^{1/2}\right)}$$

$$+c\,|x-y|_H^2|h|_H^2(t\wedge 1)^{-(1-\delta)} + c(t)\,(t\wedge 1)^{-2(1+\delta-1/m)/(1-\delta)}\,|z(t)|_H^2,$$

where $c(t)$ is a process having all moments finite. Now, if $m < 2$, it is possible to find some $\delta \in (0,1)$ such that

$$2(1+\delta-1/m)/(1-\delta) < 1,$$

and then, by using the Gronwall lemma, (7.11) follows.

Concerning the proof of 2, we recall that in [6] it has been proved that for any $\varphi \in C_b^1(E) \supset C_b^1(H)$ and $t > 0$ it holds that

$$|D(P_t\varphi)(x) - D(P_t\varphi)(y)|_{E^\star} \leq c\,(t\wedge 1)^{-\frac{1+\epsilon}{2}}\,\|\varphi\|_1^E|x-y|_E, \qquad x,y \in\ E.$$

Then as before we have that $u(t,\cdot) \in C_b^1(E)$ and

$$|Du(t,x) - Du(t,y)|_{E^\star} \leq c\,(t\wedge 1)^{-\frac{1+\epsilon}{2}}\,|x-y|_E, \tag{7.12}$$

where the constant c depends only on g, φ, and T. This makes it possible to prove that the closed loop equation admits a unique mild solution. Actually, due to the Sobolev embedding theorem, as the dimension d equals 1, for any $\epsilon > 0$ we have that $D((-C)^{1/4+\epsilon})$ is continuously embedded into E, and then

$$\begin{aligned}
&\left|\int_0^t e^{(t-s)C} DK(u(T-s,y(s;x)))\,ds\right|_E \\
&\leq c\int_0^t \left|e^{(t-s)C} DK(u(T-s,y(s;x)))\right|_{D((-C)^{1/4+\epsilon})}\,ds \\
&\leq c\int_0^t (t-s)^{-1/2-2\epsilon}\,|DK(u(T-s,y(s;x)))|_{\left(D((-C)^{1/4+\epsilon})\right)^\star} \\
&\leq c\int_0^t (t-s)^{-1/2-2\epsilon}\,|DK(u(T-s,y(s;x)))|_{E^\star}\,.
\end{aligned}$$

Therefore, as DK is Lipschitz continuous on $E^\star$ and (7.12) holds, it is easy to show that the closed loop equation admits a unique mild solution. □

Notice that if $K(x) = |x|_H^2$, then $DK(x) = x$, so that DK can be extended as a Lipschitz continuous mapping from $E^\star$ into itself.

Appendix A. An a priori estimate.

We prove here an a priori C^1 estimate for the solution u_α of the approximating problem (6.3). As in [21] we represent u_α and Du_α by means of the transition semigroups associated with suitable stochastic problems. This allows us to have a maximum principle both for u_α and Du_α.

LEMMA A.1. *Let $U : [0,T]\times H \to H$ be a bounded and measurable mapping, such that $U(t,\cdot)$ is Lipschitz continuous for any $t > 0$ and*

$$\sup_{t\in[\epsilon,T]} \|U(t,\cdot)\|_{\mathrm{Lip}}^H < \infty$$

for any $\epsilon > 0$. Then, for any $\alpha > 0$ the stochastic problem

$$dy(t) = (Ay(t) + F_\alpha(y(t)) + U(T-t,y(t)))\,dt + Q\,dw(t), \qquad y(r) = x,$$

admits a unique solution $y_\alpha(t,r;x) \in L^2(\Omega; C([r,T];H)\cap L^\infty(r,T;H))$.

Proof. For any $\epsilon > 0$ the function $U(T-t,\cdot)$ is Lipschitz continuous, uniformly for $t \in [0,T-\epsilon]$, and then there exists a unique solution $y_\alpha(t)$ in the interval $[0,T-\epsilon]$ for any $\epsilon > 0$. If we define $v_\alpha(t) = y_\alpha(t) - w^A(t,r)$, we have that $v_\alpha(t)$ is the unique solution of the problem

$$\frac{dv}{dt}(t) = Av(t) + F_\alpha(y_\alpha(t)) + U(T-t,y_\alpha(t)), \qquad v(r) = x. \tag{A.1}$$

Thus, by multiplying each side of (A.1) by $v_\alpha(t)$, we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|v_\alpha(t)|_H^2 + |(-A)^{1/2}v_\alpha(t)|_H^2 = \left\langle F_\alpha(v_\alpha(t)+w^A(t,r)) - F_\alpha(w^A(t,r)), v_\alpha(t)\right\rangle_H \\
&+ \left\langle F_\alpha(w^A(t,r)), v_\alpha(t)\right\rangle_H + \langle U(T-t,y_\alpha(t)), v_\alpha(t)\rangle_H\,.
\end{aligned}$$

Due to the Lipschitz continuity of F_α and the boundedness of U, this implies that

$$\frac{1}{2}\frac{d}{dt}|v_\alpha(t)|_H^2 \leq c_\alpha\,|v_\alpha(t)|_H^2 + c_\alpha\left(|w^A(t,r)|_H^2 + 1\right).$$

Therefore, by integrating with respect to t, by taking the supremum for $t \in [r, T-\epsilon]$, and, finally, by taking the expectation, due to the Gronwall lemma we have

$$\mathbf{E}\sup_{t\in\,[r,T-\epsilon]}|v_\alpha(t)|_H^2 \leq c_{\alpha,T}\Big(|x|_H^2 + \mathbf{E}\sup_{t\in\,[r,T]}|w^A(t,r)|_H^2\Big).$$

Thanks to the regularity of $w^A(t,r)$, this allows us to conclude that

$$\mathbf{E}\sup_{t\in\,[r,T-\epsilon]}|y_\alpha(t)|_H^2 \leq 2\sup_{t\in\,[r,T-\epsilon]}\left(|v_\alpha(t)|_H^2 + |w^A(t,r)|_H^2\right) \leq c_{\alpha,T}\left(|x|_H^2+1\right).$$

As the constant $c_{\alpha,T}$ does not depend on ϵ, by a uniqueness argument we have that the solution $y_\alpha(t)$ is defined for any $t \in [r,T)$ and $y_\alpha(t,r;x) \in L^2(\Omega; C([r,T);H) \cap L^\infty(r,T;H))$. □

For each $\alpha > 0$ and $n \in \mathbb{N}$, we introduce the approximating Hamilton–Jacobi–Bellman equation

$$\text{(A.2)}\qquad \frac{\partial u}{\partial t}(t,x) = \mathcal{L}_{\alpha,n}u(t,x) - K_n(Du(t,x)) + g_n(x), \qquad u(0,x) = \varphi_n(x),$$

where

$$\mathcal{L}_{\alpha,n}\psi(x) = \frac{1}{2}\mathrm{Tr}\,\left[Q_n^2 D^2\psi(x)\right] + \langle A_n x + F_{\alpha,n}(x), D\psi(x)\rangle_H\,.$$

By arguing as for the problem (6.3) (see Remark 6.4), if $\varphi, g \in \mathrm{Lip}_b(H)$, the problem (A.2) admits a unique local solution $u_{\alpha,n}$ in $\Lambda_R(\tau_R)$. The solution $u_{\alpha,n}$ is the unique fixed point for the functional

$$\Gamma_{\alpha,n}(v)(t,x) = P_t^{\alpha,n}\varphi(x) - \int_0^t P_{t-s}^{\alpha,n}K(Dv(s,\cdot))(x)\,ds + \int_0^t P_{t-s}^{\alpha,n}g(x)\,ds,$$

and due to (5.8) the contraction constant of $\Gamma_{\alpha,n}$ is independent of n and α. Thus we can proceed as in the proof of Lemma 6.5, and thanks to Proposition 5.1 we conclude that for any $\alpha > 0$

$$\text{(A.3)}\qquad \lim_{n\to+\infty}|u_{\alpha,n}(t,x) - u_\alpha(t,x)| + |Du_{\alpha,n}(t,x) - Du_\alpha(t,x)|_H = 0,$$

uniformly for $t \in [0,\tau_R]$ and for x in bounded sets of H.

According to (5.3) it is possible to show that u_α and $u_{\alpha,n}$ have a stronger regularity.

Lemma A.2. *If $\varphi, g \in C_b^1(H)$, then the solutions u_α and $u_{\alpha,n}$ of the problems* (6.2) *and* (6.3) *belong to $\mathcal{Z}^2_{\tau_\star}$ for some $\tau_\star = \tau_\star(\alpha) \leq T$, which can be taken independent of n.* For a proof we refer to [9, chapter 9].

Proposition A.3. *Let us fix $\varphi, g \in C_b^1(H)$, and assume that u_α is the unique solution of the problem* (6.3) *in $\mathcal{Z}^2_{\tau_\star}$, with $\tau_\star = \tau_\star(\alpha) \leq T$. Then, under Hypotheses* 1, 2, *and* 3, *we have*

$$\sup_{\alpha>0}\Big(\sup_{t\in\,[0,\tau_\star]}\|u_\alpha(t,\cdot)\|_1^H\Big) \leq c\,(1+T)e^{c\,T}\left(\|\varphi\|_1^H + \|g\|_1^H\right).$$

Proof. If we define

$$U_{\alpha,n}(t,x)=\int_0^1 DK_n(\lambda Du_{\alpha,n}(t,x))\,d\lambda, \tag{A.4}$$

the problem (A.2) can be rewritten as

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x)=\dfrac{1}{2}\mathrm{Tr}\,[Q_n^2D^2u(t,x)]+\langle A_nx+F_{\alpha,n}(x)+U_{\alpha,n}(t,x),Du(t,x)\rangle+g_n(x),\\[2ex] u(0,x)=\varphi_n(x). \end{cases}$$

Since $\varphi,g\in C_b^1(H)$, we have that the solution $u_{\alpha,n}\in\mathcal{Z}^2_{\tau_\star}$ and then $U_{\alpha,n}:[0,\tau_\star]\times H\to H_n$ is continuous and bounded. Moreover, since DK is locally Lipschitz continuous, if we define $M_{\alpha,n}$ as the Lipschitz constant of DK in the ball $\{\,x\in H\ ;\ |x|_H\leq\|u_{\alpha,n}\|_{\mathcal{Z}^2_{\tau_\star}}\,\}$ for any $x,y\in H$ and $t>0$, we have that

$$|U_{\alpha,n}(t,x)-U_{\alpha,n}(t,y)|_H\leq\int_0^1|DK_n(\lambda Du_{\alpha,n}(t,x))-DK_n(\lambda Du_{\alpha,n}(t,y))|_H\,d\lambda$$

$$\leq M_{\alpha,n}\,|Du_{\alpha,n}(t,x)-Du_{\alpha,n}(t,y)|_H\leq c\,M_{\alpha,n}\sup_{z\in H}|D^2u_{\alpha,n}(t,z)||x-y|_H$$

$$\leq c\,M_{\alpha,n}\,t^{-\frac{1+\epsilon}{2}}\|u_{\alpha,n}\|_{\mathcal{Z}^2_{\tau_\star}}|x-y|_H.$$

This means that the function $U_{\alpha,n}$ fulfills the hypotheses of Lemma A.1 so that for each $0\leq r<T$ the stochastic problem

$$dy(t)=(A_ny(y)+F_{\alpha,n}(y(t))+U_{\alpha,n}(t,y(t)))\,dt+Q_ndw(t),\qquad y(r)=P_nx, \tag{A.5}$$

admits a unique strong solution $y_{\alpha,n}(t,r;x)\in L^2(\Omega;C([r,\tau_\star);H)\cap L^\infty(r,\tau_\star;H))$.

If we denote by $R^{\alpha,n}_{s,t}$ the corresponding transition semigroup, that is,

$$R^{\alpha,n}_{s,t}\varphi(x)=\mathbf{E}\,\varphi(y_{\alpha,n}(t,s;x)),\qquad 0\leq s\leq t\leq\tau_\star,$$

for $\varphi\in B_b(H)$ and $x\in H$, we have

$$u_{\alpha,n}(t,x)=R^{\alpha,n}_{\tau_\star-t,\tau_\star}\varphi(x)+\int_0^t R^{\alpha,n}_{\tau_\star-t,\tau_\star-s}g(x)\,ds. \tag{A.6}$$

Indeed, since $y_{\alpha,n}(t)$ is a strong solution and $u_{\alpha,n}(t,x)$ is regular, we can apply Itô's formula to the function $s\mapsto u_{\alpha,n}(\tau_\star-s,y_{\alpha,n}(s,\tau_\star-t;x))$, and by integrating with respect to $s\in[\tau_\star-t,\tau_\star]$ and by taking the expectation, we get

$$u_{\alpha,n}(t,x)=\mathbf{E}\,\varphi(y_{\alpha,n}(\tau_\star,\tau_\star-t;x))+\int_0^t\mathbf{E}\,g(y_{\alpha,n}(\tau_\star-s,\tau_\star-t;x))\,ds,$$

which coincides with (A.6). As an immediate consequence this yields

$$\sup_{t\in[0,\tau_\star]}\|u_{\alpha,n}(t,\cdot)\|_0^H\leq\|\varphi\|_0^H+T\|g\|_0^H. \tag{A.7}$$

The proof of the analogous estimate for the derivative of $u_{\alpha,n}(t,x)$ is more complicated but is based on similar arguments.

The problem (A.2) can be rewritten as

$$(\mathrm{A.8})\quad \begin{cases} \dfrac{\partial u}{\partial t}(t,x) = & \dfrac{1}{2}\displaystyle\sum_{k=1}^{n}\lambda_k^2 D_k^2 u(t,x) + \sum_{k,h=1}^{n} a_{k,h}x_h D_k u(t,x) - K_n(Du(t,x)) \\ & +\displaystyle\sum_{k=1}^{n}\left\langle F_{\alpha,n}(x), e_k\right\rangle_H D_k u(t,x) + g_n(x), \\ u(0,x) = & \varphi_n(x), \end{cases}$$

where for each $k,h \in \mathbb{N}$ we denote $D_k u = \langle Du, e_k\rangle$ and $a_{k,h} = \langle Ae_k, e_h\rangle$. By differentiating each side of (A.8) with respect to x_j and by setting $v_j = D_j u$, we get

$$\frac{\partial v_j}{\partial t} = \frac{1}{2}\sum_{k=1}^{n}\lambda_k^2 D_k^2 v_j + \sum_{k=1}^{n} a_{k,h}x_h D_k v_j + \sum_{k=1}^{n} a_{k,j}v_k + \sum_{k=1}^{n}\langle F_{\alpha,n}(x), e_k\rangle D_k v_j$$

$$+\sum_{k=1}^{n}\langle DF_{\alpha,n}(x)e_j, e_k\rangle v_k - \sum_{k=1}^{n}\langle DK_n(Du_{\alpha,n}), e_k\rangle D_k v_j + \langle Dg_n(x), e_j\rangle .$$

By multiplying each side by v_j and by summing up on j, we obtain

$$\frac{1}{2}\frac{\partial}{\partial t}\sum_{j=1}^{n} v_j^2 = \frac{1}{2}\sum_{k,j=1}^{n}\lambda_k^2 v_j D_k^2 v_j + \sum_{k,h,j=1}^{n} a_{k,h}x_h v_j D_k v_j + \sum_{k,j=1}^{n} a_{k,j}v_k v_j$$

$$+\sum_{k,j=1}^{n}\langle F_{\alpha,n}(x), e_k\rangle v_j D_k v_j + \sum_{k,j=1}^{n}\langle DF_{\alpha,n}(x)v_j e_j, v_k e_k\rangle$$

$$-\sum_{k,j=1}^{n}\langle DK_n(Du_{\alpha,n}(t,x)), e_k\rangle v_j D_k v_j + \sum_{j=1}^{n}\langle Dg_n(x), v_j e_j\rangle .$$

Now, if we define $2z_{\alpha,n}(t,x) = |Du_{\alpha,n}(t,x)|_H^2$, it holds that

$$\sum_{k,j=1}^{n}\lambda_k^2 v_j D_k^2 v_j = \sum_{k=1}^{n}\lambda_k^2 D_k^2 z_{\alpha,n} - \sum_{k,j=1}^{n}\lambda_k^2 (D_k v_j)^2,$$

$$\sum_{k,h,j=1}^{n} a_{k,h}x_h v_j D_k v_j + \sum_{k,j=1}^{n} a_{k,j}v_k v_j = \langle A_n x, Dz_{\alpha,n}\rangle + \langle A_n Du_{\alpha,n}, Du_{\alpha,n}\rangle ,$$

$$\sum_{k,j=1}^{n}\langle F_{\alpha,n}(x), e_k\rangle v_j D_k v_j = \langle F_{\alpha,n}(x), Dz_{\alpha,n}\rangle ,$$

$$\sum_{k,j=1}^{n}\langle DK_n(Du_{\alpha,n}), e_k\rangle v_j D_k v_j = \langle DK_n(Du_{\alpha,n}), Dz_{\alpha,n}\rangle .$$

Moreover, we have

$$\sum_{k,j=1}^{n} \langle DF_{\alpha,n}(x)v_j e_j, v_k e_k\rangle = \langle DF_{\alpha,n}(x)Du_{\alpha,n}, Du_{\alpha,n}\rangle,$$

$$\sum_{j=1}^{n} \langle Dg_n, v_j e_j\rangle = \langle Dg_n, Du_{\alpha,n}\rangle\,.$$

Thus, by substituting and by taking into account of (2.13) and (3.8), we can conclude that

$$\frac{\partial z_{\alpha,n}}{\partial t}(t,x) \leq \mathcal{M}_{\alpha,n} z_{\alpha,n}(t,x) + c\, z_{\alpha,n}(t,x) + |Dg_n|_H^2,$$

where the differential operator $\mathcal{M}_{\alpha,n}$ is defined by

$$\mathcal{M}_{\alpha,n}\psi(x) = \frac{1}{2}\mathrm{Tr}\,[Q_n^2 D^2\psi(x)] + \langle A_n x + F_{\alpha,n}(x) - DK_n(Du_{\alpha,n}(t,x)), D\psi(x)\rangle_H\,.$$

Now we define

$$V_{\alpha,n}(t,x) = -DK_n(Du_{\alpha,n}(t,x)).$$

By arguing as above for the function $U_{\alpha,n}(t,x)$ defined in (A.4), we have that $V_{\alpha,n} : [0,\tau_\star]\times H \to H$ satisfies the hypotheses of Lemma A.1 so that the stochastic problem

$$\text{(A.9)}\quad dy(t) = (A_n y(t) + F_{\alpha,n}(y(t)) + V_{\alpha,n}(t,y(t)))\,dt + Q_n dw(t), \qquad y(r) = P_n x,$$

admits a unique strong solution $y_{\alpha,n}(t,r;x)$ for any $0 \leq r \leq \tau_\star$. If we denote by $S_{s,t}^{\alpha,n}$ the transition semigroup associated with (A.9), by arguing as before for the semigroup associated with the problem (A.5), we have that the solution of the problem

$$\frac{\partial v}{\partial t}(t,x) = \mathcal{M}_{\alpha,n} v(t,x) + cv(t,x) + |Dg_n(x)|_H^2, \qquad v(0,x) = |D\varphi_n(x)|_H^2,$$

is given by

$$v_{\alpha,n}(t,x) = e^{c\,t} S^{\alpha,n}_{\tau_\star - t,\tau_\star}|D\varphi_n|_H^2(x) + \int_0^t e^{c(t-s)} S^{\alpha,n}_{\tau_\star-t,\tau_\star-s}|Dg_n|_H^2(x)\,ds.$$

This yields

$$\sup_{t\in\,[0,\tau_\star]} \|v_{\alpha,n}(t,\cdot)\|_0^H \leq c\,(1+T)\,e^{c\,T}\left(\|\varphi\|_1^H + \|g\|_1^H\right)^2$$

so that by a comparison argument we conclude

$$\text{(A.10)}\quad \sup_{t\in\,[0,\tau_\star]} \|Du_{\alpha,n}(t,\cdot)\|_0^H \leq 2 \sup_{t\in\,[0,T]} \|z_{\alpha,n}(t,\cdot)\|_0^H \leq c\,(1+\sqrt{T})\,e^T\left(\|\varphi\|_1^H + \|g\|_1^H\right).$$

From (A.7) and (A.10), due to (A.3), we conclude that our statement holds. □

Acknowledgments. The author wishes to thank Arnaud Debussche and Fausto Gozzi for many useful discussions she had with them.

REFERENCES

[1] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand Math. Studies 2, Van Nostrand, Princeton, 1965.

[2] J.-M. Bismut, *Martingales, the Malliavin calculus and hypoellipticity under general Hörmander's conditions*, Z. Wahrsch. Verw. Gebiete, 56 (1981), pp. 469–505.

[3] P. Cannarsa and G. Da Prato, *Second-order Hamilton–Jacobi equations in infinite dimensions*, SIAM J. Control Optim., 29 (1991), pp. 474–492.

[4] P. Cannarsa and G. Da Prato, *Direct solutions of a second-order Hamilton-Jacobi equation in Hilbert spaces*, in Stochastic Partial Differential Equations and Applications, G. Da Prato and L. Tubaro, eds., Pitman Res. Notes Math. Ser. 208, Longman, Harlow, UK, pp. 72–85.

[5] S. Cerrai, *Differentiability with respect to initial datum for solutions of SPDE'S with no Fréchet differentiable drift term*, Commun. Appl. Anal., 2 (1998), pp. 249–270.

[6] S. Cerrai, *Smoothing properties of transition semigroups relative to SDE's with values in Banach spaces*, Probab. Theory Related Fields, 113 (1999), pp. 85–114.

[7] S. Cerrai, *Differentiability of Markov semigroups for stochastic reaction-diffusion equations and applications to control*, Stochastic Process. Appl., 83 (1999), pp. 15–37.

[8] S. Cerrai, *Ergodicity of stochastic reaction-diffusion systems with polynomial coefficients*, Stochastics Stochastics Rep., 67 (1999), pp. 17–51.

[9] S. Cerrai, *Second Order PDE's in Finite and Infinite Dimension. A Probabilistic Approach*, Lecture Notes in Math., Springer-Verlag, to appear.

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

[11] G. Da Prato, *Applications croissantes et équations d'évolutions dans les espaces de Banach*, Istituto Nazionale di Alta Matematica, Institutiones Mathematicae, Volume II, Academic Press, London, 1976.

[12] G. Da Prato and A. Debussche, *Control of the stochastic Burgers model of turbulence*, SIAM J. Control Optim., 37 (1999), pp. 1123–1149.

[13] G. Da Prato and A. Debussche, *Dynamic programming for the stochastic Burgers equation*, Ann. Mat. Pura Appl. (4), 178 (2000), pp. 143–174.

[14] G. Da Prato, K. D. Elworthy, and J. Zabczyk, *Strong Feller property for stochastic semilinear equations*, Stochastic Anal. Appl., 13 (1995), pp. 35–45.

[15] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.

[16] K. D. Elworthy and X. M. Li, *Formulae for the derivatives of heat semigroups*, J. Funct. Anal., 125 (1994), pp. 252–286.

[17] W. H. Fleming and M. Nisio, *On stochastic relaxed control for partially observed diffusions*, Nagoya Math. J., 93 (1984), pp. 71–108.

[18] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[19] M. Freidlin, *Markov Processes and Differential Equations: Asymptotic Problems*, Lectures in Mathematics EHT Zürich, Birkhäuser Verlag, Basel, 1996.

[20] F. Gozzi, *Regularity of solutions of a second order Hamilton-Jacobi equation and application to a control problem*, Comm. Partial Differential Equations, 20 (1995), pp. 775–826.

[21] F. Gozzi, *Global regular solutions of second order Hamilton-Jacobi equations in Hilbert spaces with locally Lipschitz nonlinearities*, J. Math. Anal. Appl., 198 (1996), pp. 399–443.

[22] J. L. Lions, *Quelques méthods de résolution des problèms aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[23] P.-L. Lions, *Viscosity solutions of fully nonlinear second order equations and optimal control in infinite dimensions.* I. *The case of bounded stochastic evolutions*, Acta Math., 161 (1998), pp. 243–278.

[24] P.-L. Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal control in infinite dimensions.* II. *Optimal control of Zakai's equation*, in Stochastic Partial Differential Equations, II (Trento, 1988), Lecture Notes in Math. 1390, G. Da Prato and L. Tubaro, eds., Springer-Verlag, New York, 1989, pp. 147–170.

[25] P.-L. Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal control in infinite dimensions.* III. *Uniqueness of viscosity solutions for general second-order equations*, J. Funct. Anal., 86 (1989), pp. 1–18.

[26] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser Verlag, Basel, 1995.

[27] S. Peszat and J. Zabczyk, *Strong Feller property and irreducibility for diffusion processes on Hilbert spaces*, Ann. Probab., 23 (1995), pp. 157–172.

[28] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1988.

[29] X. Y. Zhou, *On the existence of optimal relaxed controls of stochastic partial differential equations*, SIAM J. Control Optim., 30 (1992), pp. 247–261.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1817–1832

THE LATTICE STRUCTURE OF BEHAVIORS[*]

SHIVA SHANKAR[†]

For Professor J. C. Willems, on the occasion of his sixtieth birthday.

Abstract. If a linear, continuous, shift invariant distributed system is considered as a (dynamical) system converting input signals to output signals, then this information is encapsulated in the impulse response or the transfer function of the system. The set of all transfer functions has the structure of a ring, corresponding to the operations of parallel and cascade connections of two systems. However, in the behavioral theory of Willems, a system is not described in terms of its input-output transformation property. Indeed, the concept of a behavior does not even need the notions of inputs and outputs and is therefore more fundamental than the classical concept of a system given by its transfer function. The question then arises as to what is the structure of the set of all behaviors. This paper argues that the relevant structure here is that of a modular lattice.

Key words. distributed systems, systems of partial differential equations, modular lattice

AMS subject classifications. 93C20, 93C35, 35B37, 35E20, 06C05

PII. S0363012999358427

1. Introduction.

The purpose of this paper is to explain the algebraic structure of the set of all linear (distributed) behaviors.

In the previous algebraic theory, where the focus was more on the discrete version of a distributed system, i.e., linear multidimensional systems described by difference equations in several indeterminates (see Bose [2] for the 2-D case and Shankar and Sule [10] and Sule [12] for the general n-D case), the point of view adopted was to consider a plant as a (nonautonomous) dynamical system, transferring input signals into output signals. The system being linear, continuous, and shift invariant, the relationship between inputs and outputs could be encapsulated in the impulse response or the transfer function of the system. The transfer function of such a system was a rational function (in several indeterminates). Now two such multidimensional input-output systems could be connected either in parallel or in cascade. The transfer function of the resulting input-output system was then the sum or the product of the transfer functions of the individual systems. The set of all these systems, or more accurately their transfer functions, with these two operations had then the structure of a ring. It was this structure that was the foundation for the theory developed in [10], which described the set of all stabilizing controllers of a plant and which described the topology of robust stabilization, etc. It was this same algebraic structure that was crucial for the characterization of the obstruction to the problem of simultaneously stabilizing two plants by a single controller (or equivalently the problem of stabilizing a plant by a stable controller); see [6] and Ying [16]. This was of course hardly surprising, for any solution of a problem which required one to consider, a priori, the set of all plants should have to take into account, and indeed rely on, the structure of the set of all plants.

More recently, J. C. Willems has in a series of fundamental papers, summarized in [14, 15], transformed the stage in which control problems had been formulated and

[*] Received by the editors July 6, 1999; accepted for publication (in revised form) October 11, 2000; published electronically April 18, 2001.
http://www.siam.org/journals/sicon/39-6/35842.html

[†] Chennai Mathematical Institute, 92, G. N. Chetty Road, T. Nagar, Chennai (Madras) 600017, India (sshankar@smi.ernet.in).

solved. In his far-reaching theory, which goes beyond the description of a plant as something that transforms inputs to outputs, indeed which goes beyond notions of inputs and outputs altogether, the central object is the *behavior* of an autoregressive (AR) system (more generally the behavior of an autoregressive moving average (ARMA) system, but the "elimination theorem" allows one to reduce such a behavior to the behavior of an AR system)—formal definitions appear in the next section. This behavior is the set of all homogeneous solutions of a system of differential equations in some appropriate function or distribution space. (By a system I mean here a matrix, each entry of which is a differential operator.) While Willems and his coworkers have in the large concentrated in formulating and solving control problems, both old and new, for lumped systems (i.e., where the differential operators are ordinary)—see Trentelman [13] for a recent exposition—this theory has also been carried over effectively to distributed systems, i.e., the case of partial differential operators, in Oberst [4], in Pillai and Shankar [5], and in [7, 8, 9].

One of the important features of Willems's behavioral theory, and arguably the principal reason for its success, is the algebraic nature of the theory—in fact the algebraic structure of the set of homogeneous solutions of a matrix of differential operators is known in mathematics as a $\mathcal{D}$-module. This algebraic structure is especially crucial in the study of distributed systems for the following reason. Classically, one studied a distributed system in the framework of semigroups of transformations of some infinite dimensional space. These methods of study were function theoretic, necessarily of infinite dimensional objects, and hence rarely effective (in the sense of a finite computable procedure). On the other hand, in the behavioral theory, one associates to a distributed behavior a submodule of a free module of finite rank over the ring of constant coefficient partial differential operators (with coefficients in $\mathbb{R}$ or $\mathbb{C}$), and one studies properties of the behavior in terms of this submodule. As this ring is Noetherian, submodules of free modules of finite rank are *finitely generated.* This enables one, often, to reduce matters to the calculation of *finitely generated* objects such as torsion submodules, Ext modules, etc. (see [4, 5, 7]). Moreover, the availability of computational techniques from commutative algebra (Gröbner basis, etc.) makes the behavioral theory a *practical* one, and I foresee not long from now effective procedures to calculate stabilizing controllers for distributed systems and other such problems.

The question now arises, as in the previous theory of transfer functions, as to what is the structure underlying the set of all behaviors. As remarked earlier, a behavior does not come with a priori notions of inputs and outputs—it is possible to impose an input-output structure for hyperbolic systems (see [8])—but the general notion of a behavior does not depend upon this possibility. Thus it does not make sense, in general, to connect two behaviors in parallel or in cascade, and it is not the structure of a ring that is the right notion here. Instead, this paper argues, the relevant structure here is that of a lattice. The set of behaviors is partially ordered by inclusion, and any collection of behaviors has a least upper bound and a greatest lower bound. This much is straightforward. However, there are two other lattices that are in the picture here. One is the lattice of all submodules of the free module of fixed (finite) rank, each summand some function or distribution space, in which these behaviors are located. A behavior is such a submodule, and furthermore of a very special kind—viz., one that is a $\mathcal{D}$-module, i.e., the kernel of a system of differential operators—and the lattice of behaviors is a subset of this lattice of all submodules. The other is the lattice of all submodules of the free module (of the same rank as above) over the ring of differential operators. The question then arises as to what is the relationship between these three lattices, and the answer to it is the subject of this paper. In relating these structures there are several problems that have to be overcome. For one, distinct submodules of the free module over the ring of differential operators could determine the same behavior. For another, the sum of two behaviors need not be a behavior. It turns out that the notion of a Willems submodule (or the more general notion of closure introduced in this paper) is crucial in overcoming these problems. The importance of a Willems submodule has already been demonstrated in [5] and [7] from both the control as well as the "pure" mathematical points of view (it is the analogue of a radical ideal when a polynomial is considered as a partial differential operator). The results of those papers play a crucial role here, and in turn the results of this paper justify further this notion.

The paper is organized as follows. In the next section I establish a calculus whose final intent is to prove the differential analogue of the fact that the radical of a finite intersection of ideals is the intersection of the individual radicals (Corollary 2.3). To prove this I need results from [4] and [7]. This corollary is of crucial importance in understanding the relationship between the lattices described above, and this is the subject of section 3. The final section concerns the structure of stable and stabilizable behaviors. Stability here is a notion generalizing bounded input–bounded output (BIBO) stability of lumped behaviors that was introduced in [5] and [8], and is fundamental to control. I therefore also include a result describing necessary and sufficient conditions for the stabilizability of a behavior. The paper also includes several examples (or pathologies) which explain when things go wrong and why. It concludes with a result on determinantal ideals, an example of the reverse direction, where the theory of behaviors is used to conclude facts in algebra and in geometry.

2. The calculus of Willems submodules.

In this section I collect some preliminary definitions and results that I need in what follows.

As explained in the introduction, the central object of study in control theory, as it has emerged in recent years, is the behavior of an AR system, which I define first (see [14, 15, 4, 5]).

Let $\mathcal{D}'$ be the space of all distributions on $\mathbb{R}^n$. Let $\mathcal{A} = K[\partial_1, \dots, \partial_n]$, $K = \mathbb{R}$ or $\mathbb{C}$, be the polynomial ring in the n indeterminates $\partial_1, \dots, \partial_n$, with coefficients in K, where ∂_i is partial differentiation in the ith coordinate direction x_i (and thus an element $p(\partial)$ in $\mathcal{A}$ is a (constant coefficient) partial differential operator). The ring $\mathcal{A}$ acts on $\mathcal{D}'$ by differentiation and makes $\mathcal{D}'$ an $\mathcal{A}$-module. Given any $\mathcal{A}$-submodule $\mathcal{W}$ of $\mathcal{D}'$, the map

$$\begin{array}{rccc} p(\partial): & \mathcal{W} & \longrightarrow & \mathcal{W}, \\ & w & \mapsto & p(\partial)w \end{array}$$

is an $\mathcal{A}$-module morphism. The kernel of $p(\partial)$, which is an $\mathcal{A}$-submodule of $\mathcal{W}$, is the *behavior* of $p(\partial)$ in $\mathcal{W}$ (or the $\mathcal{W}$-behavior of $p(\partial)$), and is denoted by $\mathrm{Ker}_{\mathcal{W}}(p(\partial))$. Given an ideal $\mathcal{I}$ in $\mathcal{A}$, the $\mathcal{A}$-submodule $\bigcap_{p(\partial)\in\mathcal{I}} \mathrm{Ker}_{\mathcal{W}}(p(\partial))$ of $\mathcal{W}$ is called the $\mathcal{W}$-behavior of $\mathcal{I}$, and is denoted by $\mathrm{Ker}_{\mathcal{W}}(\mathcal{I})$.

More generally, given an element $r(\partial) = (p_1(\partial), \dots, p_k(\partial))$ in the free module $\mathcal{A}^k$, the map

$$\begin{array}{rccc} r(\partial): & \mathcal{W}^k & \longrightarrow & \mathcal{W}, \\ & w = (w_1, \dots, w_k) & \mapsto & r(\partial)w = \sum_{i=1}^k p_i(\partial)w_i \end{array}$$

is an $\mathcal{A}$-module morphism. Its kernel, denoted $\mathrm{Ker}_{\mathcal{W}}(r(\partial))$, is an $\mathcal{A}$-submodule of $\mathcal{W}^k$, and is called the $\mathcal{W}$-behavior of $r(\partial)$. Given a submodule $\mathcal{R}$ of $\mathcal{A}^k$, let $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$

denote the $\mathcal{A}$-submodule of $\mathcal{W}^k$ given by $\bigcap_{r(\partial)\in\mathcal{R}} \mathrm{Ker}_{\mathcal{W}}(r(\partial))$. Call it the $\mathcal{W}$-behavior of $\mathcal{R}$. As $\mathcal{A}$ is Noetherian, $\mathcal{R}$ is finitely generated, say, by $r_1(\partial),\ldots,r_l(\partial)$. Clearly then $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ is the kernel of the morphism

$$\begin{array}{cccc} R(\partial): & \mathcal{W}^k & \longrightarrow & \mathcal{W}^l, \\ & w & \mapsto & (r_1(\partial)w,\ldots,r_l(\partial)w). \end{array}$$

If $r_i(\partial) = (r_{i1}(\partial),\ldots,r_{ik}(\partial)), i = 1,\ldots,l$, then $R(\partial)$ can be represented, in the usual way, by the following matrix:

$$R(\partial) = \begin{pmatrix} r_{11}(\partial) & \ldots & r_{1k}(\partial) \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ r_{l1}(\partial) & \ldots & r_{lk}(\partial) \end{pmatrix}. \tag{1}$$

Thus the $\mathcal{W}$-behavior of a submodule of $\mathcal{A}^k$ is the set of homogeneous solutions in the space $\mathcal{W}^k$ of a system (i.e., matrix) of differential operators. If these differential operators are ordinary, i.e., if $n = 1$, then the behavior is said to be lumped; otherwise it is a distributed behavior. Formally, one has the following.

Definition. *Let $\mathcal{W}$ be any $\mathcal{A}$-submodule of $\mathcal{D}'$. A behavior in $\mathcal{W}^k$ (or a $\mathcal{W}$-behavior) is an $\mathcal{A}$-submodule of $\mathcal{W}^k$ of the type* $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$*, for some $\mathcal{A}$-submodule $\mathcal{R}$ of $\mathcal{A}^k$.*

It is an observation of Malgrange that

$$\mathrm{Hom}_{\mathcal{A}}(\mathcal{A}^k/\mathcal{R},\mathcal{W}) \simeq \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}),$$

where the isomorphism is given by

$$\phi \mapsto (\phi(\overline{e_1}),\ldots,\phi(\overline{e_k})),$$

$\overline{e_i}$ is the image of $e_i = (0,\ldots,1,\ldots,0)$ (1 in the ith position) in $\mathcal{A}^k/\mathcal{R}$. This isomorphism which was studied in [4] and [7] plays a central role in this paper.

On the other hand, given a behavior $\mathcal{B}$ in $\mathcal{W}^k$, i.e., an $\mathcal{A}$-submodule of $\mathcal{W}^k$ of the kind $\mathcal{B} = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ for some submodule $\mathcal{R}$ of $\mathcal{A}^k$, let $\mathcal{M}(\mathcal{B}) = \mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}))$ be the submodule of $\mathcal{A}^k$ consisting of all the elements in $\mathcal{A}^k$ that map to zero every element in $\mathcal{B}$. Clearly $\mathcal{R} \subset \mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}))$. Thus there are two assignments $\mathrm{Ker}_{\mathcal{W}}$ and $\mathcal{M}$ which are both inclusion reversing, i.e., $\mathcal{R}_1 \subset \mathcal{R}_2$ implies $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2) \subset \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1)$ and $\mathcal{B}_1 \subset \mathcal{B}_2$ implies $\mathcal{M}(\mathcal{B}_2) \subset \mathcal{M}(\mathcal{B}_1)$. In other words $\mathrm{Ker}_{\mathcal{W}}$ and $\mathcal{M}$ define a Galois connection between the partially ordered sets of submodules of $\mathcal{A}^k$ and behaviors in $\mathcal{W}^k$. The primary purpose of this paper is to study this Galois connection from the viewpoint of lattice theory.

Lemma 2.1. *Let $\{\mathcal{R}_i\}$ (respectively, $\{\mathcal{B}_i\}$) be any collection of submodules of $\mathcal{A}^k$ (respectively, behaviors in $\mathcal{W}^k$). Then*

(i) $\mathrm{Ker}_{\mathcal{W}}(\sum_i \mathcal{R}_i) = \bigcap_i \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$,

(ii) $\sum_i \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i) \subset \mathrm{Ker}_{\mathcal{W}}(\bigcap_i \mathcal{R}_i)$,

(iii) $\mathcal{M}(\sum_i \mathcal{B}_i) = \bigcap_i \mathcal{M}(\mathcal{B}_i)$,

(iv) $\sum_i \mathcal{M}(\mathcal{B}_i) \subset \mathcal{M}(\bigcap_i \mathcal{B}_i)$.

Proof. Elementary. □

Lemma 2.2. $\mathrm{Ker}_{\mathcal{W}} \circ \mathcal{M}$ *is the identity map on behaviors for any $\mathcal{A}$-submodule $\mathcal{W}$ of $\mathcal{D}'$, i.e.,* $\mathrm{Ker}_{\mathcal{W}} \circ \mathcal{M}(\mathcal{B}) = \mathcal{B}$ *for all $\mathcal{W}$-behaviors $\mathcal{B}$.*

Proof. Clearly $\mathcal{B} \subset \mathrm{Ker}_{\mathcal{W}} \circ \mathcal{M}(\mathcal{B})$. But $\mathcal{B}$ by definition is $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ for some submodule $\mathcal{R}$ of $\mathcal{A}^k$ (for some k). Hence $\mathcal{R} \subset \mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})) = \mathcal{M}(\mathcal{B})$. Then $\mathrm{Ker}_{\mathcal{W}} \circ \mathcal{M}(\mathcal{B}) \subset \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}) = \mathcal{B}$. □

COROLLARY 2.1. *The correspondence $\mathcal{B} \to \mathcal{M}(\mathcal{B})$ (between behaviors in $\mathcal{W}^k$ and submodules of $\mathcal{A}^k$) is injective for all $\mathcal{A}$-submodules $\mathcal{W}$ of $\mathcal{D}'$.*

Proof. Suppose $\mathcal{M}(\mathcal{B}_1) = \mathcal{M}(\mathcal{B}_2)$. Then $\mathcal{B}_1 = \mathrm{Ker}_{\mathcal{W}} \circ \mathcal{M}(\mathcal{B}_1) = \mathrm{Ker}_{\mathcal{W}} \circ \mathcal{M}(\mathcal{B}_2) = \mathcal{B}_2$. □

The correspondence $\mathcal{R} \to \mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ is not in general injective, i.e., in general $\mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}))$ may strictly contain $\mathcal{R}$ (for instance if $\mathcal{W}$ is the 0 submodule of $\mathcal{D}'$, then $\mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})) = \mathcal{A}^k$ for every submodule $\mathcal{R}$ of $\mathcal{A}^k$). This prompts the following definition (see also [5] and [7]).

DEFINITION. *The submodule $\mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}))$ is called the Willems closure of $\mathcal{R}$ with respect to $\mathcal{W}$. If $\mathcal{R}$ is equal to $\mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}))$, i.e., if $\mathcal{R}$ is equal to its Willems closure with respect to $\mathcal{W}$, then the submodule $\mathcal{R}$ is called a Willems submodule with respect to $\mathcal{W}$.*

The analogy between this definition and the classical definition of a radical ideal was the subject of [7]. This analogy further motivates the following results.

LEMMA 2.3. *The Willems closure of $\mathcal{R}$ with respect to $\mathcal{W}$ is Willems with respect to $\mathcal{W}$. It is maximal amongst all submodules with the same $\mathcal{W}$-behavior as $\mathcal{R}$.*

Proof. The first statement follows from

$$\mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}))) = \mathcal{M}(\mathrm{Ker}_{\mathcal{W}} \circ \mathcal{M})\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}) = \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}).$$

For the second, $\mathrm{Ker}_{\mathcal{W}}(\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$; thus the $\mathcal{W}$-behavior of $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ equals the $\mathcal{W}$-behavior of $\mathcal{R}$. If also $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$, then $\mathcal{R}_1 \subset \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) = \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$. □

LEMMA 2.4. *Let $\{\mathcal{B}_i\}$ be a collection of $\mathcal{W}$-behaviors. Then the Willems closure of $\sum_i \mathcal{M}(\mathcal{B}_i)$ with respect to $\mathcal{W}$ is $\mathcal{M}(\bigcap_i \mathcal{B}_i)$.*

Proof. Lemma 2.1 implies that $\mathrm{Ker}_{\mathcal{W}}(\sum_i \mathcal{M}(\mathcal{B}_i)) =\bigcap_i \mathrm{Ker}_{\mathcal{W}}(\mathcal{M}(\mathcal{B}_i))$, and this in turn equals $\bigcap_i \mathcal{B}_i = \mathrm{Ker}_{\mathcal{W}}\mathcal{M}(\bigcap_i \mathcal{B}_i)$. Hence the behavior of $\sum_i \mathcal{M}(\mathcal{B}_i)$ is equal to the behavior of the submodule $\mathcal{M}(\bigcap_i \mathcal{B}_i)$ which is Willems. □

LEMMA 2.5. *If $\{\mathcal{R}_i\}$ is any collection of submodules of $\mathcal{A}^k$, each Willems with respect to $\mathcal{W}$, then $\bigcap_i \mathcal{R}_i$ is also Willems with respect to $\mathcal{W}$.*

Proof. By assumption $\mathcal{M}(\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)) = \mathcal{R}_i$ for each i. Hence

$$\bigcap_i \mathcal{R}_i \subset \mathcal{M}\mathrm{Ker}_{\mathcal{W}}\left(\bigcap_i \mathcal{R}_i\right) \subset \mathcal{M}\left(\sum_i \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)\right) = \bigcap_i \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i) = \bigcap_i \mathcal{R}_i,$$

where the second inclusion follows from Lemma 2.1(ii) and the inclusion reversing nature of the assignment $\mathcal{M}$. This implies equality everywhere, and thus that $\bigcap_i \mathcal{R}_i = \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\bigcap_i \mathcal{R}_i)$. □

A question more general than the above lemma, which is of crucial importance in the next section, is whether $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\bigcap_i \mathcal{R}_i)$ equals $\bigcap_i \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$ for an arbitrary collection of submodules $\{\mathcal{R}_i\}$ of $\mathcal{A}^k$. In other words, is the Willems closure of an intersection equal to the intersection of the Willems closures? Of course as $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i) \subset \mathrm{Ker}_{\mathcal{W}}(\bigcap_i \mathcal{R}_i)$ it follows that, always,

$$\mathcal{M}\mathrm{Ker}_{\mathcal{W}}\left(\bigcap_i \mathcal{R}_i\right) \subset \bigcap_i \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i). \tag{2}$$

The other inclusion is not in general true for an arbitrary collection of submodules, as the following "pathology" demonstrates.

Pathology 1. Let $\mathcal{A} = \mathbb{C}[\frac{d}{dt}]$, and let $\mathcal{I}$ be any nonzero proper ideal of $\mathcal{A}$. Let $\mathcal{R}_i = \mathcal{I}^i, i \geq 0$, and consider the behaviors of $\mathcal{R}_i$ in $\mathcal{D}$, the space of compactly supported, complex valued, smooth functions on $\mathbb{R}$. As each $\mathcal{R}_i$ is a nonzero ideal, $\text{Ker}_{\mathcal{D}}(\mathcal{R}_i)$ consists of just the 0 function alone (no nonzero differential operator admits a compactly supported homogeneous solution by the Paley–Wiener theorem). Thus each $\mathcal{M}\text{Ker}_{\mathcal{D}}(\mathcal{R}_i) = \mathcal{A}$ and hence $\bigcap_i \mathcal{M}\text{Ker}_{\mathcal{D}}(\mathcal{R}_i) = \mathcal{A}$. On the other hand $\bigcap_i \mathcal{R}_i = 0$ by Krull's theorem ($\mathcal{A}$ is an integral domain, so its $\mathcal{I}$-adic completion is Hausdorff). Thus $\text{Ker}_{\mathcal{D}}(\bigcap_i \mathcal{R}_i) = \mathcal{D}$, and as the 0 ideal is the only ideal whose behavior is all of $\mathcal{D}$, it follows that $\mathcal{M}\text{Ker}_{\mathcal{D}}(\bigcap_i \mathcal{R}_i) = 0$. □

The question therefore is whether the above is true for finite intersections, i.e., does $\mathcal{M}\text{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i) = \bigcap_{i=1}^m \mathcal{M}\text{Ker}_{\mathcal{W}}(\mathcal{R}_i)$ hold? (This is the differential analogue of the fact that the radical of a finite intersection of ideals equals the intersection of the individual radicals.) While it may possibly be true for any $\mathcal{A}$-submodule $\mathcal{W}$ of $\mathcal{D}'$, I show below that it is indeed true for the "classical" spaces by using the characterization of Willems submodules with respect to these spaces established in [4] and [7]. As this paper restricts attention to these spaces, I highlight them in the following definition.

Definition. *An $\mathcal{A}$-submodule $\mathcal{W}$ of $\mathcal{D}'$ is said to be a classical space if it is one of the following:*

(i) *$\mathcal{D}'$, the space of all distributions on $\mathbb{R}^n$,*
(ii) *$\mathcal{C}^\infty$, the space of all (complex valued) smooth functions on $\mathbb{R}^n$,*
(iii) *$\mathcal{S}'$, the space of temperate distributions on $\mathbb{R}^n$,*
(iv) *$\mathcal{S}$, the Schwartz space of rapidly decreasing functions on $\mathbb{R}^n$,*
(v) *$\mathcal{E}'$, the space of compactly supported distributions on $\mathbb{R}^n$,*
(vi) *$\mathcal{D}$, the space of compactly supported smooth functions on $\mathbb{R}^n$.*

In the following theorem, I collect results on the Willems closure of a submodule $\mathcal{R}$ of $\mathcal{A}^k$ with respect to the classical spaces from [4] and [7]. Following this I prove a stronger result on the Willems closure with respect to $\mathcal{S}'$ than the one in [7] (this stronger version is more convenient for the purposes of this paper and is used below). To state them, I employ the following notation. Given the submodule $\mathcal{R}$ of $\mathcal{A}^k$, represent $\mathcal{R}$ by an $l \times k$ matrix (whose entries are elements of $\mathcal{A}$), where the l rows (as elements of $\mathcal{A}^k$) generate $\mathcal{R}$, viz., the matrix in (1) above. Consider the kth determinantal ideal of this matrix, i.e., the ideal generated by the determinants of all the $k \times k$ submatrices. Clearly this ideal depends only on the submodule $\mathcal{R}$ and not on the choice of the matrix representing it, i.e., it is independent of the choice of the generators of $\mathcal{R}$. Denote this determinantal ideal by $I_k(\mathcal{R})$ and call it the characteristic ideal of $\mathcal{R}$. (Note that if $l < k$, then $I_k(\mathcal{R}) = 0$.) The affine variety of this ideal in $\mathbb{C}^n$, $\mathcal{V}(I_k(\mathcal{R}))$, is called the characteristic variety of $\mathcal{R}$ (i.e., considering the elements of $\mathcal{A}$ as polynomials). Let $\Im\mathcal{V}(I_k(\mathcal{R}))$ denote the set of purely imaginary points on $\mathcal{V}(I_k(\mathcal{R}))$, i.e., $\Im\mathcal{V}(I_k(\mathcal{R})) = \mathcal{V}(I_k(\mathcal{R})) \cap \imath\mathbb{R}^n$.

Theorem 2.1. (i) *Every submodule $\mathcal{R}$ of $\mathcal{A}^k$ is Willems with respect to $\mathcal{D}'$ or $\mathcal{C}^\infty$.*

(ii) *Let $\mathcal{R} = \bigcap_{i=1}^t \mathcal{Q}_i$ be an irredundant primary decomposition of $\mathcal{R}$ in $\mathcal{A}^k$, where $\mathcal{Q}_i$ is $\mathcal{P}_i$-primary. Suppose that the affine varieties in $\mathbb{C}^n$ of $\mathcal{P}_1, \ldots, \mathcal{P}_r$ intersect $\Im\mathcal{V}(I_k(\mathcal{R}))$, and that of the varieties of $\mathcal{P}_{r+1}, \ldots, \mathcal{P}_t$ do not. Then the Willems closure $\mathcal{M}\text{Ker}_{\mathcal{S}'}(\mathcal{R})$ of $\mathcal{R}$ with respect to $\mathcal{S}'$ is $\bigcap_{i=1}^r \mathcal{Q}_i$, so that $\mathcal{R}$ is Willems with respect to $\mathcal{S}'$ if and only if the variety of every $\mathcal{P}_i$ intersects $\Im\mathcal{V}(I_k(\mathcal{R}))$. In other words $\mathcal{R}$ is*

Willems with respect to $\mathcal{S}'$ if and only if the variety of every associated prime of $\mathcal{A}^k/\mathcal{R}$ intersects $\Im\mathcal{V}(I_k(\mathcal{R}))$.

(iii) *Let $\pi : \mathcal{A}^k \to \mathcal{A}^k/\mathcal{R}$ be the canonical projection. Then the Willems closure of $\mathcal{R}$ with respect to $\mathcal{S}$, $\mathcal{E}'$, or $\mathcal{D}$ is $\pi^{-1}(T(\mathcal{A}^k/\mathcal{R}))$, where $T(\mathcal{A}^k/\mathcal{R})$ is the submodule of torsion elements of $\mathcal{A}^k/\mathcal{R}$, so that $\mathcal{R}$ is Willems with respect to $\mathcal{S}$, $\mathcal{E}'$, or $\mathcal{D}$ if and only if $\mathcal{A}^k/\mathcal{R}$ is torsion-free (or equivalently if and only if $\mathcal{R}$ is 0-primary).* □

Remark. (i) The first part of the above result is a consequence of two deep theorems about $\mathcal{D}'$ and $\mathcal{C}^\infty$. Ehrenpreis, Malgrange, and Palamodov prove that these two spaces are injective $\mathcal{A}$-modules, i.e., that $\mathrm{Hom}_{\mathcal{A}}(-, \mathcal{D}')$ and $\mathrm{Hom}_{\mathcal{A}}(-, \mathcal{C}^\infty)$ are exact (contravariant) functors (see Hörmander [3]). Furthermore, Oberst in [4] proves that $\mathcal{D}'$ and $\mathcal{C}^\infty$ are cogenerators, i.e., that $\mathrm{Hom}_{\mathcal{A}}(\mathcal{X}, \mathcal{D}')$ and $\mathrm{Hom}_{\mathcal{A}}(\mathcal{X}, \mathcal{C}^\infty)$ are equal to 0 if and only if the $\mathcal{A}$-module $\mathcal{X}$ is 0. I use these results in the next section.

The following is now immediate from Lemma 2.4.

COROLLARY 2.2. *Let $\mathcal{B}_i$ be any collection of $\mathcal{D}'$ or $\mathcal{C}^\infty$-behaviors. Then $\sum_i \mathcal{M}(\mathcal{B}_i) = \mathcal{M}(\bigcap_i \mathcal{B}_i)$.*

(ii) It is an easy fact that the Willems closure of $\mathcal{R}$ with respect to $\mathcal{S}'$ described above is independent of the primary decomposition of $\mathcal{R}$ in $\mathcal{A}^k$ (see [7]). Below I establish a stronger and more convenient version of this result.

(iii) The Willems closure of a submodule $\mathcal{R}$ of $\mathcal{A}^k$ with respect to $\mathcal{S}$ is thus equal to its Willems closure with respect to $\mathcal{E}'$ or $\mathcal{D}$ [7].

The following result is a strengthening of the second part of the above theorem.

THEOREM 2.2. *Let $\mathcal{R} = \bigcap_{i=1}^t \mathcal{Q}_i$ be an irredundant primary decomposition of the submodule $\mathcal{R}$ in $\mathcal{A}^k$, where $\mathcal{Q}_i$ is $\mathcal{P}_i$-primary. Suppose that the affine varieties in $\mathbb{C}^n$ of $\mathcal{P}_1, \ldots, \mathcal{P}_r$ contain purely imaginary points (i.e., intersect $\imath\mathbb{R}^n$) and that of $\mathcal{P}_{r+1}, \ldots, \mathcal{P}_t$ do not. Then the Willems closure $\mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R})$ of $\mathcal{R}$ with respect to $\mathcal{S}'$ is $\bigcap_{i=1}^r \mathcal{Q}_i$, so that $\mathcal{R}$ is Willems with respect to $\mathcal{S}'$ if and only if the variety of every associated prime of $\mathcal{A}^k/\mathcal{R}$ contains purely imaginary points.*

Proof. I show that if $\mathcal{R} = \bigcap_{i=1}^t \mathcal{Q}_i$ is an irredundant primary decomposition of $\mathcal{R}$ in $\mathcal{A}^k$, where $\mathcal{Q}_i$ is $\mathcal{P}_i$-primary, then $\bigcup_{i=1}^t \mathcal{V}(\mathcal{P}_i)$ is contained in $\mathcal{V}(I_k(\mathcal{R}))$. This theorem now follows from Theorem 2.1(ii) above, as then the purely imaginary points on $\mathcal{V}(\mathcal{P}_i)$ are exactly those that are the intersection of $\mathcal{V}(\mathcal{P}_i)$ with $\Im\mathcal{V}(I_k(\mathcal{R}))$.

Assume first that $\mathcal{R}$ itself is a $\mathcal{P}$-primary submodule of $\mathcal{A}^k$. I need to show that the determinantal ideal $I_k(\mathcal{R})$ is contained in $\mathcal{P}$, so that $\mathcal{V}(\mathcal{P})$ is then contained in $\mathcal{V}(I_k(\mathcal{R}))$. Thus let d in $I_k(\mathcal{R})$ be the determinant of some $k \times k$ matrix, say D, with entries in $\mathcal{A}$, all of whose rows, considered as elements of $\mathcal{A}^k$, are in $\mathcal{R}$—by definition every generator of $I_k(\mathcal{R})$ occurs in this way. Let D' be the matrix adjoint to D, so that $D'D$ is the $k \times k$ diagonal matrix with the diagonal elements each equal to d. But (matrix) multiplying D on the left by any row, i.e., by an element of $\mathcal{A}^k$, is equivalent to taking an $\mathcal{A}$-linear combination of the rows of D. Thus each row of $D'D$ is an element of $\mathcal{R}$ (as each row of D is an element of $\mathcal{R}$). But as these rows are $d \cdot e_i$, $i = 1, \ldots, k$, $e_i = (0, \ldots, 1, \ldots, 0)$—1 in the ith place—it follows that d is in $ann(\mathcal{A}^k/\mathcal{R})$. As $\mathcal{R}$ is $\mathcal{P}$-primary, $\mathcal{P}$ is the radical of $ann(\mathcal{A}^k/\mathcal{R})$, and thus it further follows that d is in $\mathcal{P}$. This shows that $I_k(\mathcal{R})$ is contained in $\mathcal{P}$, so that $\mathcal{V}(\mathcal{P})$ is contained in $\mathcal{V}(I_k(\mathcal{R}))$.

If now $\mathcal{R} = \bigcap_{i=1}^t \mathcal{Q}_i$, where $\mathcal{Q}_i$ is $\mathcal{P}_i$-primary, then $I_k(\mathcal{R})$ is contained in each $I_k(\mathcal{Q}_i)$ and hence, by the above, that it is in fact contained in each $\mathcal{P}_i$ and hence in $\bigcap_{i=1}^t \mathcal{P}_i$. Then again $\bigcup_{i=1}^t \mathcal{V}(\mathcal{P}_i)$ is contained in $\mathcal{V}(I_k(\mathcal{R}))$. This proves the theorem. □

Remark. The above description of the Willems closure of a submodule with respect to the classical spaces can also be described in terms of localization in a uniform manner. Recollect that if U is a multiplicatively closed subset of $\mathcal{A}$, then one can form rings and modules of fractions in the usual way. Thus $U^{-1}\mathcal{A}^k$ is a $U^{-1}\mathcal{A}$ module, and $\phi : \mathcal{A}^k \to U^{-1}\mathcal{A}^k$, mapping x in $\mathcal{A}^k$ to $\frac{x}{1}$, is an $\mathcal{A}$-module morphism. It is an easy fact that if $\mathcal{R}$ is a submodule of $\mathcal{A}^k$, and if $U^{-1}\mathcal{R}$ is its extension in $U^{-1}\mathcal{A}^k$, then its contraction $\phi^{-1}(U^{-1}\mathcal{R})$ is given by $\cup_{u\in U}(\mathcal{R} : u)$. In this language the above results can be described as follows.

(i) If U_1 is the set of units of $\mathcal{A}$, then $U_1^{-1}\mathcal{A}$ is equal to $\mathcal{A}$, the map ϕ above is the identity, and the Willems closure of $\mathcal{R}$ with respect to $\mathcal{D}'$ or $\mathcal{C}^\infty$ equals $\phi^{-1}(U_1^{-1}\mathcal{R})$, which is of course just equal to $\mathcal{R}$.

(ii) Let U_2 be the multiplicatively closed subset of $\mathcal{A}$ consisting of those polynomials whose affine varieties *do not* contain purely imaginary points. Then an easy check shows that $\mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R})$ is equal to $\phi^{-1}(U_2^{-1}\mathcal{R})$.

(iii) Let U_3 be the multiplicatively closed subset of $\mathcal{A}$ consisting of all the nonzero elements. Then the Willems closure of $\mathcal{R}$ with respect to $\mathcal{S}$, $\mathcal{E}'$, or $\mathcal{D}$ is $\phi^{-1}(U_3^{-1}\mathcal{R})$.

The description of the lattice structure of behaviors in the next section is based on the following corollary to the above theorems. This corollary is, as observed above, the differential analogue of the fact that the radical of a finite intersection of ideals is the intersection of the individual radicals.

COROLLARY 2.3. *Let $\mathcal{R}_i, i = 1, \dots, m$, be submodules of $\mathcal{A}^k$, and let $\mathcal{W}$ be any of the classical spaces. Then*

$$\mathcal{M}\mathrm{Ker}_{\mathcal{W}}\left(\bigcap_{i=1}^{m}\mathcal{R}_i\right) = \bigcap_{i=1}^{m}\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i). \tag{3}$$

Thus in a classical space, the Willems closure of a finite intersection equals the intersection of the individual Willems closures.

Proof. By induction it suffices to prove the above when $m = 2$.

(i) In the case when $\mathcal{W}$ is either $\mathcal{D}'$ or $\mathcal{C}^\infty$, the statement is trivial, as then every submodule is Willems. Thus each side of (3) equals $\mathcal{R}_1 \cap \mathcal{R}_2$.

(ii) Because of the inclusion (2) observed earlier, it is sufficient to prove the other inclusion, viz., that $\mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R}_1) \cap \mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R}_2) \subset \mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R}_1 \cap \mathcal{R}_2)$. Thus let $\mathcal{R}_1 = \bigcap_{i=1}^{t_1}\mathcal{Q}_i$ and $\mathcal{R}_2 = \bigcap_{j=1}^{t_2}\mathcal{Q}'_j$ be irredundant primary decompositions of $\mathcal{R}_1$ and $\mathcal{R}_2$ in $\mathcal{A}^k$, where $\mathcal{Q}_i$ and $\mathcal{Q}'_j$ are $\mathcal{P}_i$-primary and $\mathcal{P}'_j$-primary, respectively, $i = 1, \dots, t_1, j = 1, \dots, t_2$. Suppose that the varieties of $\mathcal{P}_1, \dots, \mathcal{P}_{r_1}$ contain purely imaginary points, whereas those of $\mathcal{P}_{r_1+1}, \dots, \mathcal{P}_{t_1}$ do not. Similarly suppose that the varieties of $\mathcal{P}'_1, \dots, \mathcal{P}'_{r_2}$, and not those of $\mathcal{P}'_{r_2+1}, \dots, \mathcal{P}'_{t_2}$, contain purely imaginary points. Then by the above theorem

$$\mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R}_1) \cap \mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R}_2) = \left(\bigcap_{i=1}^{r_1}\mathcal{Q}_i\right) \cap \left(\bigcap_{j=1}^{r_2}\mathcal{Q}'_j\right).$$

On the other hand $(\bigcap_{i=1}^{t_1}\mathcal{Q}_i) \cap (\bigcap_{j=1}^{t_2}\mathcal{Q}'_j)$ is a primary decomposition of $\mathcal{R}_1 \cap \mathcal{R}_2$ in $\mathcal{A}^k$, though perhaps not irredundant. An irredundant primary decomposition can, however, be obtained from it by omitting (if necessary) some of the $\mathcal{Q}_i$ or $\mathcal{Q}'_j$. Thus the set of associated primes of $\mathcal{R}_1 \cap \mathcal{R}_2$ is a subset of $\{\mathcal{P}_1, \dots, \mathcal{P}_{t_1}, \mathcal{P}'_1, \dots, \mathcal{P}'_{t_2}\}$. This implies that those associated primes of $\mathcal{R}_1 \cap \mathcal{R}_2$ whose varieties contain purely imaginary points are a subset of $\{\mathcal{P}_1, \dots, \mathcal{P}_{r_1},\ P'_1, \dots, \mathcal{P}'_{r_2}\}$. Clearly then this implies (by the above theorem) that $\mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R}_1) \cap \mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R}_2) \subset \mathcal{M}\mathrm{Ker}_{\mathcal{S}'}(\mathcal{R}_1 \cap \mathcal{R}_2)$, and thus that (3) holds when $\mathcal{W}$ is $\mathcal{S}'$.

(iii) Suppose $\mathcal{W}$ is $\mathcal{S}$, $\mathcal{E}'$, or $\mathcal{D}$. Then the Willems closure of $\mathcal{R}_1 \cap \mathcal{R}_2$ (with respect to such a $\mathcal{W}$) is by the above theorem $\pi^{-1}(T(\mathcal{A}^k/\mathcal{R}_1 \cap \mathcal{R}_2))$, where $\pi : \mathcal{A}^k \to \mathcal{A}^k/\mathcal{R}_1 \cap \mathcal{R}_2$ is the canonical projection. But an element x in $\mathcal{A}^k$ such that ax is in $\mathcal{R}_1 \cap \mathcal{R}_2$, for some nonzero a in $\mathcal{A}$, implies that the residue class of x in $\mathcal{A}^k/\mathcal{R}_1$ and in $\mathcal{A}^k/\mathcal{R}_2$ is a torsion element of $\mathcal{A}^k/\mathcal{R}_1$ and $\mathcal{A}^k/\mathcal{R}_2$, respectively. Conversely, let x be an element of $\mathcal{A}^k$ whose residue class in $\mathcal{A}^k/\mathcal{R}_1$ and in $\mathcal{A}^k/\mathcal{R}_2$ is a torsion element of $\mathcal{A}^k/\mathcal{R}_1$ and $\mathcal{A}^k/\mathcal{R}_2$, respectively. So let a_1x and a_2x (a_1 and a_2 nonzero) belong to $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. This implies that a_1a_2x belongs to $\mathcal{R}_1 \cap \mathcal{R}_2$. As $\mathcal{A}$ is an integral domain, a_1a_2 is nonzero, so that the residue class of x in $\mathcal{A}^k/\mathcal{R}_1 \cap \mathcal{R}_2$ is a torsion element. Thus when $\mathcal{W}$ is $\mathcal{S}$, $\mathcal{E}'$, or $\mathcal{D}$, it again follows that $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1 \cap \mathcal{R}_2) = \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) \cap \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2)$, so that (3) holds for these classical spaces as well. □

These results permit the description of the lattice structure of behaviors in the classical spaces in the next section. The calculus of behaviors is further developed in [11].

Remark. In view of the fact that inclusion (2) holds for any $\mathcal{A}$-submodule $\mathcal{W}$ of $\mathcal{D}'$, it is tempting to conjecture that the reverse inclusion (for finite intersections because of Pathology 1), and hence equality (3), also holds for any $\mathcal{W}$. It is not clear to me how to even proceed to prove this general statement. The proof of (3) for the classical spaces depends upon the description of the Willems closures of modules, a description that is not available for a general $\mathcal{W}$. Relevant here is the following warning. As observed earlier, the concept of a Willems closure is the generalization to analysis of the algebraic notion of the radical of an ideal (via the Hilbert Nullstellensatz). The proof of the algebraic fact corresponding to (3), viz., that the radical of a finite intersection equals the intersection of the individual radicals, also depends upon a concrete description of a radical ideal.

3. The structure of behaviors. The starting point of this paper, as explained in the introduction, is the elementary observation that the set $\mathbf{L}$ of all submodules of $\mathcal{A}^k$, partially ordered by inclusion, is a complete modular lattice under the operations of module sum and intersection. Recollect (from Birkhoff [1]) that to say $\mathbf{L}$ is a complete lattice is to say that any collection $\{\mathcal{R}_i\}$, not necessarily finite, of submodules of $\mathcal{A}^k$ has a least upper bound (l.u.b.), or join, given by $\sum_i \mathcal{R}_i$, and a greatest lower bound (g.l.b.) or meet given by $\bigcap_i \mathcal{R}_i$, and to say $\mathbf{L}$ is modular is to say that if $\mathcal{R}_2 \subset \mathcal{R}_1$, then $\mathcal{R}_1 \cap (\mathcal{R}_2 + \mathcal{R}_3) = \mathcal{R}_2 + (\mathcal{R}_1 \cap \mathcal{R}_3)$ for any submodule $\mathcal{R}_3$. Recollect also that $\mathbf{L}$ is not, however, a distributive lattice; standard examples show that $\mathcal{R}_1 \cap (\mathcal{R}_2 + \mathcal{R}_3)$ is *not* always equal to $(\mathcal{R}_1 \cap \mathcal{R}_2) + (\mathcal{R}_1 \cap \mathcal{R}_3)$.

Now let $\mathcal{W}$ be any $\mathcal{A}$-submodule of $\mathcal{D}'$. Each fact noted above about the set $\mathbf{L}$ is of course also true about the set $\mathbf{W}$ of all $\mathcal{A}$-submodules of $\mathcal{W}^k$. Thus $\mathbf{W}$, partially ordered by inclusion, is also a complete modular lattice. Consider now the contravariant functor $\mathrm{Ker}_{\mathcal{W}}$ described in the previous section. Let $\mathbf{B}(\mathcal{W}) = \{\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}) \mid \mathcal{R} \in \mathbf{L}\}$ be the set of all behaviors in $\mathcal{W}^k$, also partially ordered by inclusion. I claim that this partially ordered set is also, trivially, a lattice. For suppose that $\{\mathcal{B}_i\}$ is any collection of elements in $\mathbf{B}(\mathcal{W})$. Then its g.l.b. is $\bigcap_i \mathcal{B}_i$ (this intersection is a behavior by Lemma 2.1(i)). This collection also has an l.u.b.; in fact let $\{\mathcal{B}_\alpha\}$ be the collection of all behaviors that contain every $\mathcal{B}_i$—this collection is clearly nonempty as it contains the behavior $\mathcal{W}^k$. Then $\bigcap_\alpha \mathcal{B}_\alpha$ is a behavior (again by Lemma 2.1(i)) which is clearly the smallest behavior that contains every $\mathcal{B}_i$ and is therefore its l.u.b.

Thus there are three lattices in picture here, viz., $\mathbf{L}$, $\mathbf{W}$, and its subset $\mathbf{B}(\mathcal{W})$, and the question arises as to what is the relationship between them. I show below that if $\mathcal{W}$ is a classical space, then $\mathbf{B}(\mathcal{W})$ is anti-isomorphic to a homomorphic image of $\mathbf{L}$ (anti-isomorphic means that the map between these lattices is inclusion reversing, so that g.l.b.'s go to l.u.b.'s and vice versa). This implies then that $\mathbf{B}(\mathcal{W})$ is modular. Furthermore, if $\mathcal{W}$ is $\mathcal{D}'$, $\mathcal{C}^\infty$, or $\mathcal{S}'$, then $\mathbf{B}(\mathcal{W})$ is a sublattice of $\mathbf{W}$. These are the central results of this paper.

PROPOSITION 3.1. *Let $\mathcal{W}$ be a classical space and let $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i), i = 1,\dots,m$, be any finite set of behaviors in $\mathcal{W}^k$. Then $\mathrm{Ker}_{\mathcal{W}}(\sum_{i=1}^m \mathcal{R}_i)$ is the largest $\mathcal{W}$-behavior contained in every $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$, and $\mathrm{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i)$ is the smallest $\mathcal{W}$-behavior containing every $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$.*

Proof. The first part of the statement follows from Lemma 2.1(i). For the second part, suppose $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ contains $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i), i = 1,\dots,m$. Then $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ is contained in each $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$, and hence in $\bigcap_{i=1}^m \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$. But this intersection is by Corollary 2.3 (i.e., by (3)) equal to $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i)$ (as $\mathcal{W}$ is by assumption a classical space). It now follows that

$$\mathrm{Ker}_{\mathcal{W}}\left(\bigcap_{i=1}^m \mathcal{R}_i\right) = \mathrm{Ker}_{\mathcal{W}}\mathcal{M}\mathrm{Ker}_{\mathcal{W}}\left(\bigcap_{i=1}^m \mathcal{R}_i\right) \subset \mathrm{Ker}_{\mathcal{W}}\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}).$$

As clearly $\mathrm{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i)$ contains every $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$, it follows that it is in fact the smallest behavior containing every $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$. □

The converse is also true, viz., the following.

PROPOSITION 3.2. *Suppose that $\mathrm{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i)$ is the smallest $\mathcal{W}$-behavior containing every $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$. Then (3) holds.*

Proof. Suppose that $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ contains every $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$. Then by assumption $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ contains $\mathrm{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i)$. In other words, if $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ is contained in $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$ for each i, then $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ is contained in $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i)$. Hence as $\bigcap_{i=1}^m \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$ is contained in $\mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i)$ for each i, it follows that

$$\bigcap_{i=1}^m \mathcal{M}\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i) \subset \mathcal{M}\mathrm{Ker}_{\mathcal{W}}\left(\bigcap_{i=1}^m \mathcal{R}_i\right),$$

which together with (2) implies (3). □

In essence what Proposition 3.1 establishes (via Corollary 2.3) is that if $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R})$ contains $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i), i = 1,\dots,m$, then it also contains $\mathrm{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i)$. This statement is not true if i ranges over an infinite set for the same reasons that in this case Corollary 2.3 does not hold.

Pathology 2. Consider again the situation of Pathology 1. Let $\mathcal{R} = \mathcal{I}$ be any nonzero proper ideal of $\mathbb{C}[\frac{d}{dt}]$, and let $\mathcal{R}_i = \mathcal{I}^i, i \geq 1$. Then $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}) = \mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_i) = 0$, whereas $\mathrm{Ker}_{\mathcal{D}}(\bigcap_i \mathcal{R}_i) = \mathrm{Ker}_{\mathcal{D}}(0) = \mathcal{D}$. □

By Proposition 3.1 it follows that the lattice structure of $\mathbf{B}(\mathcal{W})$ described in the beginning of this section is in fact given by defining the g.l.b. of any finite set of behaviors $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i), i = 1,\dots,m$, to be $\mathrm{Ker}_{\mathcal{W}}(\sum_{i=1}^m \mathcal{R}_i)$ and its l.u.b. to be $\mathrm{Ker}_{\mathcal{W}}(\bigcap_{i=1}^m \mathcal{R}_i)$. This lattice structure is inherited from the lattice structure of $\mathbf{L}$ in the following manner. Consider the equivalence relation $\sim$ on $\mathbf{L}$ given by

$$\mathcal{R} \sim \mathcal{R}' \quad \text{if} \quad \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}').$$

In other words, $\mathcal{R}$ is equivalent to $\mathcal{R}'$ if their Willems closures in $\mathcal{W}$ are equal. Then if $\mathcal{R}_1 \sim \mathcal{R}_1'$ and $\mathcal{R}_2 \sim \mathcal{R}_2'$, it follows (by Lemma 2.1) that $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1+\mathcal{R}_2) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1)\cap$

$\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}'_1) \cap \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}'_2) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}'_1 + \mathcal{R}'_2)$, and thus that $\mathcal{R}_1 + \mathcal{R}_2 \sim \mathcal{R}'_1 + \mathcal{R}'_2$. It also follows that $\mathcal{R}_1 \cap \mathcal{R}_2 \sim \mathcal{R}'_1 \cap \mathcal{R}'_2$, that is, that $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1 \cap \mathcal{R}_2) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}'_1 \cap \mathcal{R}'_2)$, for these two behaviors are, by the above proposition, the l.u.b.'s of $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2)$ and $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}'_1) + \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}'_2)$, respectively, and these two sums are equal by assumption. Thus $\sim$ is a congruence relation on $\mathbf{L}$, and quotienting $\mathbf{L}$ by it gives a lattice homomorphic to $\mathbf{L}$. One therefore obtains the following theorem.

THEOREM 3.1. *Let $\mathcal{W}$ be a classical space. Then $\mathbf{B}(\mathcal{W})$, the lattice of all behaviors in $\mathcal{W}^k$, is anti-isomorphic to a homomorphic image of the lattice $\mathbf{L}$ of all submodules of $\mathcal{A}^k$. It follows therefore that $\mathbf{B}(\mathcal{W})$ is a modular lattice.*

Proof. It only remains to observe that a homomorphic image of a modular lattice is itself modular [1, Ex. 5, p. 66]). □

On the other hand, as observed earlier, the set $\mathbf{W}$ of all $\mathcal{A}$-submodules of $\mathcal{W}^k$, partially ordered by inclusion, is a complete modular lattice. The lattice $\mathbf{B}(\mathcal{W})$ of all behaviors in $\mathcal{W}^k$ is a subset of $\mathbf{W}$, and the question then arises as to whether $\mathbf{B}(\mathcal{W})$ is a sublattice of $\mathbf{W}$. This will indeed be so if the sum of two behaviors (and hence of a finite number of behaviors) is not just an $\mathcal{A}$-submodule of $\mathcal{W}^k$, but is itself a behavior—the intersection of behaviors is of course always a behavior (Lemma 2.1(i)). As $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1 \cap \mathcal{R}_2)$ is the smallest $\mathcal{W}$-behavior containing $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2)$ (by Proposition 3.1), the question therefore reduces to whether $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2)$ equals $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1 \cap \mathcal{R}_2)$. More generally, when is the inclusion in Lemma 2.1(ii) an equality?

Pathology 3. Consider once more the situation of Pathologies 1 and 2; i.e., let $\mathcal{A} = \mathbb{C}[\frac{d}{dt}]$. Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be cyclic submodules of $\mathcal{A}^2$ generated by $(1, 0)$ and $(1, -\frac{d}{dt})$, respectively. Then $\mathcal{R}_1 \cap \mathcal{R}_2$ is the 0 submodule, so that $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_1 \cap \mathcal{R}_2)$ is all of $\mathcal{D}^2$. On the other hand, $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_1) = \{(0, f) \mid f \in \mathcal{D}\}$ and $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_2) = \{(\frac{dg}{dt}, g) \mid g \in \mathcal{D}\}$. Thus an element (u, v) in $\mathcal{D}^2$ is in $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_2)$ only if $u = \frac{dg}{dt}, v = f + g$, where f and g are arbitrary elements in $\mathcal{D}$. Let now u be any (nonzero) *nonnegative* compactly supported smooth function. Then $(u, 0)$, which is in $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_1 \cap \mathcal{R}_2)$, is, however, not in $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_2)$, as $g(t) = \int_{-\infty}^{t} dg = \int_{-\infty}^{t} u dt$ is not compactly supported.

Hence $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_2)$ can be strictly contained in $\mathrm{Ker}_{\mathcal{D}}(\mathcal{R}_1 \cap \mathcal{R}_2)$, and thus the lattice structure of $\mathbf{B}(\mathcal{D})$ described by the theorem above does not make it a sublattice of the lattice of all $\mathbb{C}[\frac{d}{dt}]$-submodules of $\mathcal{D}^2$. □

The situation is, however, different if $\mathcal{W}$ is $\mathcal{D}'$, $\mathcal{C}^\infty$, or $\mathcal{S}'$.

THEOREM 3.2. *Let $\mathcal{A} = K[\partial_1, \ldots, \partial_n]$, $K = \mathbb{R}$ or $\mathbb{C}$, and let $\mathcal{R}_1$ and $\mathcal{R}_2$ be submodules of $\mathcal{A}^k$. Then $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1 \cap \mathcal{R}_2)$ if $\mathcal{W}$ is an injective $\mathcal{A}$-module.*

Proof. Because of Lemma 2.1(ii), it suffices to prove that $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1 \cap \mathcal{R}_2) \subset \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2)$. Thus given an element f in $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1 \cap \mathcal{R}_2)$, one needs to show that it can be written as $f_1 + f_2$, f_i in $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_i), i = 1, 2$, when $\mathcal{W}$ is an injective $\mathcal{A}$-module.

Recollect from the previous section the isomorphism of Malgrange, viz., that $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}) \simeq \mathrm{Hom}_{\mathcal{A}}(\mathcal{A}^k/\mathcal{R}, \mathcal{W})$ for any submodule $\mathcal{R}$ of $\mathcal{A}^k$. Thus given an element f in $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1 \cap \mathcal{R}_2)$, consider it as an $\mathcal{A}$-module morphism

$$f : \mathcal{A}^k/\mathcal{R}_1 \cap \mathcal{R}_2 \longrightarrow \mathcal{W}.$$

Define

$$\begin{array}{cccccc} g: & \mathcal{R}_1 & + & \mathcal{R}_2 & \longrightarrow & \mathcal{W}, \\ & x_1 & + & x_2 & \mapsto & g(x_1 + x_2) = f([x_2]), \end{array}$$

where $[x_2]$ is the residue class of x_2 in $\mathcal{A}^k/\mathcal{R}_1\cap\mathcal{R}_2$. This morphism g is well defined, for if $x_1+x_2 = x_1'+x_2'$ (x_i, x_i' in $\mathcal{R}_i$), then $x_1-x_1' = x_2'-x_2$ is in $\mathcal{R}_1\cap\mathcal{R}_2$. This implies that $f([x_2'-x_2]) = 0$, i.e., that $f([x_2]) = f([x_2'])$, and hence that $g(x_1+x_2) = g(x_1'+x_2')$. By definition, g restricted to $\mathcal{R}_1$ is the zero morphism, and so it induces a morphism

$$h : (\mathcal{R}_1 + \mathcal{R}_2)/\mathcal{R}_1 \longrightarrow \mathcal{W}.$$

Let $\pi : (\mathcal{R}_1+\mathcal{R}_2)/\mathcal{R}_1\cap\mathcal{R}_2 \to (\mathcal{R}_1+\mathcal{R}_2)/\mathcal{R}_1$ be the canonical projection. Let h_1 be the morphism obtained by composing π with h, i.e., let

$$h_1 = h\circ\pi : (\mathcal{R}_1+\mathcal{R}_2)/\mathcal{R}_1\cap\mathcal{R}_2 \longrightarrow \mathcal{W}.$$

Now $(\mathcal{R}_1+\mathcal{R}_2/\mathcal{R}_1)$ is canonically isomorphic to $\mathcal{R}_2/\mathcal{R}_1\cap\mathcal{R}_2$, a submodule of $(\mathcal{R}_1+\mathcal{R}_2)/\mathcal{R}_1\cap\mathcal{R}_2$. By construction h_1 is equal to f on this submodule and equal to the zero morphism on the submodule $\mathcal{R}_1/\mathcal{R}_1\cap\mathcal{R}_2$. As $\mathcal{W}$ is an injective $\mathcal{A}$-module (by assumption), there exists a morphism $f_1 : \mathcal{A}^k/\mathcal{R}_1\cap\mathcal{R}_2 \to \mathcal{W}$ so that the following diagram commutes:

$$\begin{tikzcd}
0 \arrow[r] & (\mathcal{R}_1+\mathcal{R}_2)/\mathcal{R}_1\cap\mathcal{R}_2 \arrow[r] \arrow[d, "h_1"'] & \mathcal{A}^k/\mathcal{R}_1\cap\mathcal{R}_2 \arrow[dl, "f_1"] \\
 & \mathcal{W} &
\end{tikzcd}.$$

As f_1 restricts to the zero morphism on $\mathcal{R}_1/\mathcal{R}_1\cap\mathcal{R}_2$, one can consider f_1 as an element in $\mathrm{Hom}_{\mathcal{A}}(\mathcal{A}^k/\mathcal{R}_1, \mathcal{W})$, i.e., f_1 is in the behavior $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1)$.

Next consider the morphism $f_2 = f - f_1 : \mathcal{A}^k/\mathcal{R}_1\cap\mathcal{R}_2 \to \mathcal{W}$. As f_1 equals f on $\mathcal{R}_2/\mathcal{R}_1\cap\mathcal{R}_2$, f_2 is identically zero on this submodule and can therefore be considered as an element in $\mathrm{Hom}_{\mathcal{A}}(\mathcal{A}^k/\mathcal{R}_2, \mathcal{W})$, i.e., in $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2)$.

Thus $\mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1\cap\mathcal{R}_2) = \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_1) + \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}_2)$. □

COROLLARY 3.1. *Let $\mathcal{W}$ be $\mathcal{D}'$, $\mathcal{C}^\infty$, or $\mathcal{S}'$. Then $\mathbf{B}(\mathcal{W})$, the lattice of all behaviors in $\mathcal{W}^k$, is a sublattice of $\mathbf{W}$, the lattice of all $\mathcal{A}$-submodules of $\mathcal{W}^k$.*

Proof. By the theorem of Ehrenpreis–Malgrange–Palamodov (Remark (i) following Theorem 2.1), $\mathcal{D}'$ and $\mathcal{C}^\infty$ are injective $\mathcal{A}$-modules. It is a result in [9] that $\mathcal{S}'$ is also an injective $\mathcal{A}$-module. □

COROLLARY 3.2. *The lattice $\mathbf{W}$, when $\mathcal{W}$ is either $\mathcal{D}'$ or $\mathcal{C}^\infty$, contains a sublattice anti-isomorphic to $\mathbf{L}$.*

Proof. Consider the map

$$\begin{array}{cccc} \phi : & \mathbf{L} & \longrightarrow & \mathbf{W}, \\ & \mathcal{R} & \mapsto & \mathrm{Ker}_{\mathcal{W}}(\mathcal{R}), \end{array}$$

where $\mathcal{W}$ is either $\mathcal{D}'$ or $\mathcal{C}^\infty$. By Oberst (again Remark (i) following Theorem 2.1), these two $\mathcal{A}$-modules are cogenerators, hence ϕ is an injective map. The corollary now follows from Theorem 3.1 and Corollary 3.1 above. □

While $\mathbf{W}$ is a complete lattice (i.e., arbitrary collections of elements in $\mathbf{W}$ have l.u.b.'s and g.l.b.'s), the lattices $\mathbf{B}(\mathcal{D}')$ and $\mathbf{B}(\mathcal{C}^\infty)$ (which are also complete) are, however, not complete as sublattices of $\mathbf{W}$ as the following pathology demonstrates. Thus the map ϕ in the proof of the above corollary is not a morphism of *complete* lattices.

Pathology 4. Let $\mathcal{A} = \mathbb{C}[\frac{d}{dt}]$, $\mathcal{I} = (\frac{d}{dt})$, and $\mathcal{R}_i = \mathcal{I}^i, i \geq 0$. Consider the collection of $\mathcal{C}^\infty$-behaviors $\{\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i)\}$. As elements of $\mathbf{W}$, i.e., as $\mathbb{C}[\frac{d}{dt}]$-submodules of $\mathcal{C}^\infty$, this collection has an l.u.b., viz., $\sum_i \mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i)$. I claim, however, that $\sum_i \mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i)$

is not a behavior, so that $\sum_i \mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i)$ is not the l.u.b. of $\{\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i)\}$ in $\mathbf{B}(\mathcal{C}^\infty)$. For suppose that this sum is a behavior, say, equal to $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$. Then as every submodule now is Willems (Theorem 2.1(i)), it follows that

$$\mathcal{R} = \mathcal{M}\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}) = \mathcal{M}\left(\sum_i \mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i)\right) = \bigcap_i \mathcal{M}\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i) = \bigcap_i \mathcal{R}_i,$$

where the third equality follows from Lemma 2.1(iii). But as before, by Krull's theorem, $\bigcap_i \mathcal{R}_i = 0$, so that $\sum_i \mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i)$ must equal all of $\mathcal{C}^\infty$. This is of course absurd since $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_i)$ equals the $\mathbb{C}$-vector space of polynomials of degree less than i, and which therefore implies that the above sum of behaviors is only the vector space of all polynomials. This shows that $\mathbf{B}(\mathcal{C}^\infty)$ is not complete as a sublattice of $\mathbf{W}$. □

In the next section I study how this structure is inherited by the class of stable behaviors.

4. The structure of stable behaviors. I consider now, in the context of the lattice structure developed in the previous section, an important subclass of (smooth) behaviors, viz., the stable ones. Indeed, the fundamental problem in control theory is to stabilize an a priori unstable mechanism using a controller. The notion of stability that I consider here is a generalization of BIBO stability of lumped systems introduced in [5] and [8], where the growth of the elements in a behavior is specified along certain directions. For the sake of completeness, I quickly review this notion below.

DEFINITION. *The directions of stability is a proper closed convex cone C in $\mathbb{R}^n$ (with vertex at the origin). A $\mathcal{C}^\infty$-behavior is said to be stable with respect to C if every element in it tends to* 0 *along every half line in C.*

Given the cone C of stable directions, define the subset $C_<$ of $\mathbb{R}^n$ as consisting of those points x in $\mathbb{R}^n$ such that $\langle x, y\rangle < 0$ for every nonzero y in C. As the cone C is proper, i.e., as it does not contain a full line, $C_<$ is nonempty and in fact has nonempty interior.

The stability of a behavior is determined by its characteristic variety $\mathcal{V}(I_k(\mathcal{R}))$ (defined above Theorem 2.1), The reason for this is the following. If $R(\partial)$ is any matrix representing $\mathcal{R}$ (as in (1)), then let $R(x)$ be the matrix obtained by replacing the entries of $R(\partial)$ by corresponding polynomials (in the n indeterminates $x_1, \dots, x_n$). Substituting for x any point ξ in $\mathcal{V}(I_k(\mathcal{R}))$ results in a matrix $R(\xi)$ with entries in $\mathbb{C}$ whose column rank is less than k. This implies that there is a nonzero element, say, $c = (c_1, \dots, c_k)$ in $\mathbb{C}^k$, which is in the kernel of the linear map determined by $R(\xi)$. An easy check now shows that then $(c_1 e^{\langle x,\xi\rangle}, \dots, c_k e^{\langle x,\xi\rangle})$ is in the $\mathcal{C}^\infty$-behavior of $\mathcal{R}$. (Here $\langle x, \xi\rangle = \sum_{i=1}^n x_i \xi_i$.) Thus there are such exponential elements in the behavior corresponding to every point in $\mathcal{V}(I_k(\mathcal{R}))$.

Conversely, suppose that $w = (c_1 e^{\langle x,\xi\rangle}, \dots, c_k e^{\langle x,\xi\rangle})$ is in the $\mathcal{C}^\infty$-behavior of $\mathcal{R}$ or equivalently in the kernel of the morphism given by the matrix $R(\partial)$ of (1). As in the proof of Theorem 2.2, let $D(\partial)$ be any $k \times k$ submatrix of $R(\partial)$, so that the w above is also in the behavior of the submodule generated by the rows of $D(\partial)$. Let $d(\partial)$ be the determinant of $D(\partial)$. This $d(\partial)$ is in $I_k(\mathcal{R})$; in fact $I_k(\mathcal{R})$ is generated by such $d(\partial)$. Multiplying $D(\partial)$ on the left by its adjoint $D'(\partial)$ results in a diagonal matrix all of whose entries are $d(\partial)$. It now follows that every component of w is a homogeneous solution of $d(\partial)$. But it is an easy check that $e^{\langle x,\xi\rangle}$ is a homogeneous solution of $d(\partial)$ if and only if ξ lies in the variety of d (now considered as a polynomial). As this is true for every generator of $I_k(\mathcal{R})$, it follows that ξ must lie in $\mathcal{V}(I_k(\mathcal{R}))$. More generally, call an element $(p_1(x)e^{\langle x,\xi\rangle}, \dots, p_k(x)e^{\langle x,\xi\rangle})$ in a behavior, where $p_1, \dots, p_k$

are polynomials, an *exponential element.* Just as above, these exponential elements in $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$ arise from exactly all the points of $\mathcal{V}(I_k(\mathcal{R}))$. It is a theorem of Ehrenpreis, Malgrange, and Palamodov that the closed linear hull of such exponential elements in the $\mathcal{C}^\infty$-behavior of a submodule equals its entire $\mathcal{C}^\infty$-behavior (viz., Theorem 7.6.14 in Hörmander [3]). This enables one to reduce the study of the stability of a behavior to algebraic criteria.

I quote from [5] and [8] a result on stable behaviors that I need at the very end of this section. I use in it, and further below, the following terminology from [3]. Given a subset X of $\mathbb{R}^n$, the *tube* over X is the set of points z in $\mathbb{C}^n$ such that $\Re z$, the real part of z, is in X.

THEOREM 4.1. *If the $\mathcal{C}^\infty$-behavior of a submodule $\mathcal{R}$ is stable with respect to C, then the characteristic variety $\mathcal{V}(I_k(\mathcal{R}))$ is contained in the tube over $C_<$. On the other hand, suppose that the characteristic ideal $I_k(\mathcal{R})$ of $\mathcal{R}$ contains a polynomial without multiple factors (for instance, if this ideal were a radical ideal). Then if $\mathcal{V}(I_k(\mathcal{R}))$ is contained in the tube over $C_<$ and if its distance from the boundary of this tube is strictly positive, then* $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$ *is stable with respect to $C_<$.*

DEFINITION. *A behavior* $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$ *is said to be stabilizable with respect to the cone C if it contains a nontrivial sub-behavior stable with respect to C.*

Restricting such a behavior to a stable sub-behavior is the process of *control.* Suppose that $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}')$ is a sub-behavior of $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$ stable with respect to C. As every submodule of $\mathcal{A}^k$ is Willems with respect to $\mathcal{C}^\infty$, it follows that $\mathcal{R} \subset \mathcal{R}'$. Let $\mathcal{R}_1$ be any submodule of $\mathcal{A}^k$ such that $\mathcal{R} + \mathcal{R}_1 = \mathcal{R}'$. Then as $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}') = \mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}) \cap \mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_1)$ is stable with respect to C (by supposition), the behavior $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_1)$ is said to be a stabilizing controller. If $\mathcal{R}$ and $\mathcal{R}_1$ are represented by matrices $R(\partial)$ and $R_1(\partial)$ (as in (1)), then the above process amounts to appending the rows of $R_1(\partial)$ to the rows of $R(\partial)$. This formulation of control, which does not require any notions of inputs and outputs, is one of Willems' fundamental contributions.

THEOREM 4.2. *A behavior* $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$ *is stabilizable with respect to the cone C if and only if $\mathcal{V}(I_k(\mathcal{R}))$, the characteristic variety of $\mathcal{R}$, intersects the tube over $C_<$.*

Proof. Suppose that $\mathcal{V}(I_k(\mathcal{R}))$ does not intersect the tube over $C_<$. Then any exponential element in $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$, corresponding to some point $\mathcal{V}(I_k(\mathcal{R}))$, will not tend to 0 along some half line in C. As $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$, as also every nontrivial sub-behavior of it, is a closed linear hull of such exponentials (Theorem 7.6.14 in [3]), it follows that $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$ is not stabilizable with respect to C.

Conversely suppose that $\mathcal{V}(I_k(\mathcal{R}))$ intersects the tube over $C_<$. Let ξ be some point in this intersection. Then the element $w = (c_1 e^{\langle x,\xi\rangle}, \dots, c_k e^{\langle x,\xi\rangle}), c_1, \dots, c_k$ arbitrary, is in the behavior of the submodule $\mathcal{R}_1$ of $\mathcal{A}^k$ generated by the rows of

$$R_1(\partial) = \begin{pmatrix} \partial_1 - \xi_1 & 0 & \dots & 0 \\ \partial_2 - \xi_2 & 0 & \dots & 0 \\ . & . & \dots & . \\ \partial_n - \xi_n & 0 & \dots & 0 \\ 0 & \partial_1 - \xi_1 & \dots & 0 \\ . & . & \dots & . \\ 0 & \partial_n - \xi_n & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & \partial_1 - \xi_1 \\ . & . & \dots & . \\ 0 & 0 & \dots & \partial_n - \xi_n \end{pmatrix}.$$

In fact every exponential solution of this behavior is of the above form. However, as observed earlier, there is a choice of the $c_1, \ldots, c_k$ such that the corresponding exponential element w, as above, is also in $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$. Indeed, these are the only kind of exponential elements that are in $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}) \cap \mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_1)$. As such exponentials tend to 0 uniformly along every half line in C, it follows, once again by Theorem 7.6.14 in [3], that $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$ is stabilizable with respect to C, and that $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_1)$ is a stabilizing controller. $\square$

Recollect from [1] the definitions of an ideal in a lattice and that of a meet homomorphism. An ideal in a lattice is a subset such that (i) the l.u.b. of any two elements of the subset is in it, and (ii) the g.l.b. of an element in the subset and any element in the lattice is also in the subset. A meet homomorphism between two lattices is an order preserving map such that the image of the g.l.b. of any two elements equals the g.l.b. of the images of the two elements.

THEOREM 4.3. *The set of $\mathcal{C}^\infty$-behaviors, stable with respect to a cone C, is an ideal in the lattice of all behaviors. Given a stabilizable behavior, the set of its stabilizing controllers is the inverse image of this ideal under a meet homomorphism.*

Proof. Let $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_1)$ and $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R}_2)$ be two behaviors stable with respect to C. Then by Theorem 3.2, the l.u.b. of these two behaviors is their sum, which is clearly also stable with respect to C. Moreover, if $\mathrm{Ker}_{\mathcal{C}^\infty}(\mathcal{R})$ is any behavior, the g.l.b. of this behavior and a stable behavior, being a sub-behavior of both, is also stable. This shows that this set of behaviors, stable with respect to C, is an ideal, say $\mathcal{J}$ of $\mathbf{B}(\mathcal{C}^\infty)$.

Let $\mathcal{B}_0$ be a stabilizable behavior. Consider the following map from the lattice of $\mathcal{C}^\infty$-behaviors to itself:

$$\begin{array}{rccc} \phi: & \mathbf{B}(\mathcal{C}^\infty) & \longrightarrow & \mathbf{B}(\mathcal{C}^\infty), \\ & \mathcal{B} & \mapsto & \mathcal{B} \cap \mathcal{B}_0. \end{array}$$

This map is clearly order preserving. As $\phi(\mathcal{B}_1 \cap \mathcal{B}_2) = (\mathcal{B}_1 \cap \mathcal{B}_2) \cap \mathcal{B}_0 = \phi(\mathcal{B}_1) \cap \phi(\mathcal{B}_2)$, it is a meet homomorphism. The set of behaviors that stabilize $\mathcal{B}_0$ are those that are mapped by ϕ into the ideal $\mathcal{J}$ of stable behaviors, i.e., $\phi^{-1}(\mathcal{J})$. $\square$

I conclude with a curious geometric consequence of the above results.

COROLLARY 4.1. *Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be two submodules of $\mathcal{A}^k$ such that both the determinantal ideals $I_k(\mathcal{R}_1)$ and $I_k(\mathcal{R}_2)$ contain elements without multiple factors. Suppose also that both the varieties $\mathcal{V}(I_k(\mathcal{R}_1))$ and $\mathcal{V}(I_k(\mathcal{R}_2))$ are contained in the tube over $C_<$. Then the variety $\mathcal{V}(I_k(\mathcal{R}_1 \cap \mathcal{R}_2))$ is also contained in the tube over $C_<$.*

Proof. Let C' be any cone contained in the interior of C, so that $C_<$ is contained in the interior of $C'_<$. By Theorem 4.1, the $\mathcal{C}^\infty$-behaviors of $\mathcal{R}_1$ and $\mathcal{R}_2$ are both stable with respect to C'. Thus by the above, the sum of these behaviors, which is the behavior of $\mathcal{R}_1 \cap \mathcal{R}_2$, is also stable with respect to C'. Theorem 4.1 now implies that $\mathcal{V}(I_k(\mathcal{R}_1 \cap \mathcal{R}_2))$ is contained in the tube over $C'_<$. As the intersection of all such $C'_<$ is the cone $C_<$, the corollary follows. $\square$

REFERENCES

[1] G. BIRKHOFF, *Lattice Theory*, AMS, Providence, RI, 1948.

[2] N.K. BOSE, *Applied Multidimensional Systems Theory*, Van Nostrand Reinhold Company, New York, 1982.

[3] L. HÖRMANDER, *An Introduction to Complex Analysis in Several Variables*, 3rd ed., North-Holland, Amsterdam, 1990.

[4] U. OBERST, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.

[5] H. Pillai and S. Shankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1999), pp. 388–408.

[6] S. Shankar, *An obstruction to the simultaneous stabilization of two $n-D$ plants*, Acta Appl. Math., 36 (1994), pp. 289–301.

[7] S. Shankar, *The Nullstellensatz for systems of PDE*, Adv. Appl. Math., 23 (1999), pp. 360–374.

[8] S. Shankar, *Can one control the vibrations of a drum? Recent progress in multidimensional theory and applications*, Multidimens. Systems Signal Process., 11 (2000), pp. 67–81.

[9] S. Shankar, *The Fundamental Principle over $\mathcal{S}'$*, preprint.

[10] S. Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.

[11] S. Shankar and J. C. Willems, *Behaviours of n-D systems*, in Proceedings of the 2nd International Workshop on Multidimensional Systems, Lower Silesia, Poland, 2000, pp. 23–30.

[12] V. R. Sule, *Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control Optim., 32 (1994), pp. 1675–1695.

[13] H. L. Trentelman, *Behaviours of Linear Differential Systems*, The Advanced Lecture Series at the Indian Institute of Technology, Bombay, 1998.

[14] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[15] J. C. Willems, *Open dynamical systems and their control*, in Proceedings of the International Congress of Mathematicians, Vol. III, Berlin, 1998, Doc. Math., 1998, pp. 697–706.

[16] J. Q. Ying, *On the strong stabilizability of MIMO n-dimensional linear systems*, SIAM J. Control Optim., 38 (1999), pp. 313–335.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1833–1851

STABILIZABILITY OF SYSTEMS OF ONE-DIMENSIONAL WAVE EQUATIONS BY ONE INTERNAL OR BOUNDARY CONTROL FORCE*

FARID AMMAR KHODJA† AND AHMED BADER†

Abstract. We study the internal and boundary stabilizability of a system of wave equations by one control force. We prove that the "classical" internal damping applied to only one of the equations never gives exponential stability if the wave speeds are different and, if the wave speeds are the same, we give explicit necessary and sufficient conditions for the stability to occur. We also study the simultaneous boundary stabilization of the same system.

Key words. stabilization, semigroups, essential spectrum

AMS subject classifications. 93C20, 93D99, 37L15

PII. S0363012900366613

1. Introduction.

The starting point of this work was the study of the stabilizability of two coupled abstract second order equations in Hilbert spaces using only one control force. More precisely, let us consider the system

$$\begin{cases} u''(t) = & -Au(t) + Bv'(t) \qquad \text{in } H, \\ v''(t) = & -B^*u'(t) - Cv(t) - Dv'(t) \quad \text{in } G, \end{cases}$$

where H and G are Hilbert spaces. The question then is to characterize the widest classes of operators A, B, C, and D for which the uniform stability of the semigroup associated with this system (once the conditions for its existence are ensured) holds. A general answer seems to be difficult but some results are given, with rather restrictive assumptions, by Afilal and Khodja [1] (see also [3], [5], and [7] for abstract thermoelastic systems which correspond to this system by neglecting v''). In this last paper, it was pointed out that there was a "gap" between the cases $A = C$ and $A \neq C$ in that it is easier to find a stabilizing operator D in the first case ($A = C$) than in the second ($A \neq C$); see [1] for more details.

In this paper, keeping in mind the abstract approach we have just described, we will confine ourselves to the study of one-dimensional hyperbolic systems which are close to the Timoshenko beam equations (see, for instance, [11]).

The first problem we consider is the following.

$$(1.1) \quad \begin{cases} u_{tt} = u_{xx} + b(x)v_t \ + f \qquad\qquad \text{in } (0,\infty)\times(0,1), \\[2ex] v_{tt} = \eta^2 v_{xx} - b(x)u_t + g \quad \text{in } (0,\infty)\times(0,1), \\[2ex] u(t,0) = v(t,0) = u(t,1) = v(t,1) = 0, \quad t \in (0,\infty), \\[2ex] u(0,x) = u_0(x),\ u_t(0,x) = u_1(x),\ v(0,x) = v_0(x),\ v_t(0,x) = v_1(x), \end{cases}$$

*Received by the editors January 26, 2000; accepted for publication (in revised form) November 4, 2000; published electronically April 18, 2001.
http://www.siam.org/journals/sicon/39-6/36661.html

†Université de Franche-Comté, Département de Mathématiques et UMR 6623, 16 route de Gray, 25030 Besançon cedex, France (ammar@math.univ-fcomte.fr, bader@math.univ-fcomte.fr).

where $\eta \in \mathbb{R}$, $b \in C([0,1])$ and f and g are two control forces. Our aim here is to study the exponential stability of (1.1) whenever we choose

$$f \equiv 0,\ g = -a(x)v_t, \tag{1.2}$$

where $a \in C([0,1])$. Our system then writes

$$\begin{cases} u_{tt} = u_{xx} + b(x)v_t & \text{in } (0,\infty)\times(0,1), \\ v_{tt} = \eta^2 v_{xx} - b(x)u_t \ - a(x)v_t & \text{in } (0,\infty)\times(0,1), \\ u(t,0) = v(t,0) = u(t,1) = v(t,1) = 0, \quad t \in (0,\infty), \\ u(0,x) = u_0(x),\ u_t(0,x) = u_1(x), v(0,x) = v_0(x), v_t(0,x) = v_1(x). \end{cases} \tag{1.3}$$

The natural energy associated with (1.3) is

$$E(t) = \int_0^1 \left(\mid u_t \mid^2 + \mid u_x \mid^2 + \mid v_t \mid^2 + \eta^2 \mid v_x \mid^2\right) dx. \tag{1.4}$$

Let us recall that (1.3) is *exponentially stable* if there exist $\omega > 0$ and $M > 0$ such that

$$E(t) \leq Me^{-\omega t}E(0) \qquad \forall t > 0 \tag{1.5}$$

holds for any initial data (u_0, u_1, v_0, v_1) with finite energy. It is said to be *strongly stable* if for any initial data (u_0, u_1, v_0, v_1) with finite energy

$$\lim_{t\to\infty} E(t) = 0. \tag{1.6}$$

Our result is then the following.

THEOREM 1.1. *Assume that $a, b \in C([0,1])$.*

(i) *If $\eta \neq 1$, then* (1.5) *does not hold.*

(ii) *If $\eta = 1$, assume moreover that a and b have* disjoint supports. *Then* (1.5) *holds if and only if* (1.3) *is strongly stable and*

$$\overline{a} := \int_0^1 a(x)dx > 0,\ \overline{b} := \int_0^1 b(x)dx \notin \pi\mathbb{Z}. \tag{1.7}$$

Remark 1.2. We will see that under the condition (1.7), the exponential stability holds up in an invariant subspace of the energy space of finite codimension.

If a and b have *disjoint supports,* we are not able to prove strong stability even in the dissipative case ($a \geq 0$ on $(0,1)$). If a and b have *the same support,* the strong stability follows from Kapitonov's result [6] in the dissipative case (see the next remark).

Remark 1.3. The technique we use allows us to deal with systems with more than two wave equations. This is what is done by Kapitonov [6] in higher dimensions. His assumptions amount to taking a, b with the same support assuming that $a > 0$ on its support, and using the multiplier technique, he proves the exponential stability under additional geometrical assumptions (which are easily verified in the one-dimensional case).

With a slight modification of Kapitonov's proof, it is possible, in the one-dimensional case, to prove the exponential stability by assuming only that the supports of a and b contain a common interval.

N. Burq (lecture given at an international conference on control theory, Nancy, France, March 1999), has generalized our result to higher dimension using the microlocal defect measures of P. Gérard and L. Tartar.

The second problem we deal with is

$$\begin{cases} u_{tt} = u_{xx} + b(x)v_t & \text{in } (0,\infty)\times(0,1), \\ v_{tt} = \eta^2 v_{xx} - b(x)u_t & \text{in } (0,\infty)\times(0,1), \\ u(t,0) = v(t,0) = 0, \\ u(t,1) = f(t), \quad \eta^2 v_x(t,1) = g(t), \quad t\in(0,\infty), \\ u(0,x) = u_0(x),\ u_t(0,x) = u_1(x),\ v(0,x) = v_0(x),\ v_t(0,x) = v_1(x). \end{cases}$$

Following Lions [8], we would like to stabilize *simultaneously* this system by two boundary control forces which are related by the relation

$$f' = g \qquad \text{on } (0,\infty).$$

A natural choice of the force g which makes our system dissipative (i.e., $E'(t) \le 0$ for $t > 0$) is

$$g(t) = -\alpha\left(u_x(t,1) + v_t(t,1)\right), \ \ \alpha > 0.$$

We then get the system

$$(1.8) \quad \begin{cases} u_{tt} = u_{xx} + b(x)v_t & \text{in } (0,\infty)\times(0,1), \\ v_{tt} = \eta^2 v_{xx} - b(x)u_t & \text{in } (0,\infty)\times(0,1), \\ u(t,0) = v(t,0) = 0, \\ \quad \eta^2 v_x(t,1) = u_t(t,1) = -\alpha\left(u_x(t,1) + v_t(t,1)\right), \quad t\in(0,\infty), \\ u(0,x) = u_0(x),\ u_t(0,x) = u_1(x),\ v(0,x) = v_0(x),\ v_t(0,x) = v_1(x). \end{cases}$$

The energy associated with this system is again given by (1.4). Our result for system (1.8) is the following.

THEOREM 1.4. *Assume that* $b \in C([0,1])$ *and* $\alpha > 0$. *Then*

(i) *if* $\eta \neq 1$, (1.5) *holds if and only if* (1.8) *is strongly stable and*

$$\eta = \frac{2p+1}{q} \ \ \textit{for some } (p,q) \in \mathbb{Z}\times\mathbb{Z}^*;$$

(ii) *if* $\eta = 1$, (1.5) *holds if and only if* (1.8) *is strongly stable and*

$$\bar{b} := \int_0^1 b(x)dx \neq (2k+1)\frac{\pi}{2} \qquad \textit{for any} \ \ k\in\mathbb{Z}.$$

A related work which deals with simultaneous controllability of a system of one-dimensional wave equations can be found in Avodin and Tucsnak [4].

Remark 1.5. As for the previous system, we will show that exponential stability holds up in an invariant subspace of the energy space of finite codimension.

If $b \equiv 0$, it is easy to verify that strong stability holds. However, if $b \neq 0$, it seems to be difficult to find conditions on b which imply the strong stability.

The paper is organized as follows. In the second section, we begin by stating and proving a lemma extending a result of Neves, Ribeiro, and Lopes [9]. We prove Theorems 1.1 and 1.4 in the third section.

Some of the results of this paper were already announced in [2].

2. A lemma. We consider a one-dimensional hyperbolic system written in the form

$$(2.1)\quad \begin{cases} \dfrac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = -M(x)\dfrac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix} - N(x)\begin{pmatrix} u \\ v \end{pmatrix} \qquad \text{on } [0,T]\times\,]0,l[, \\[2ex] \dfrac{d}{dt}\left[v(t,l) - Du(t,l)\right] = Fu(t,l) + Gv(t,l), \\[2ex] u(t,0) = Ev(t,0), \end{cases}$$

where

(i) *$N(x)$ is an $n \times n$ matrix whose entries n_{ij} are continuous complex valued functions of x in $[0,l]$,*

(ii) *$M(x)$ is a diagonal matrix satisfying*

$$M(x) = \operatorname{diag}\left([M_{ii}(x)]_{i=1}^{r}, [M_{jj}(x)]_{j=r+1}^{q}\right),$$

where M_{ii} (resp., M_{jj}) are diagonal matrices such that

$$\begin{aligned} M_{ii}(t,x) &= \lambda_i(x) I_{m_i}, \qquad & i &= 1,\dots,r, \\ M_{jj}(t,x) &= \mu_j(x) I_{m_j}, \qquad & j &= r+1,\dots,q, \end{aligned}$$

and where I_{m_i} is the identity matrix of size m_i and

$$\sum_{i=1}^{r} m_i = p; \quad \sum_{j=r+1}^{q} m_j = n - p.$$

We suppose also that the entries of $M(x)$ are real valued C^1 functions in x with

$$\lambda_i(x) > 0 \text{ and } \mu_j(x) < 0 \quad \forall\, i,j \text{ and } \forall\, x \in [0,l],$$

(iii) *$u(t,x) = (u_i(t,x))_{i=1}^{p}$ and $v(t,x) = (v_i(t,x))_{j=p+1}^{n}$,*

(iv) *$D, E, F,$ and G are matrices of appropriate sizes.*

With system (2.1), we consider the reduced system

$$(2.2)\quad \begin{cases} \dfrac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = -M(x)\dfrac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix} - N_0(x)\begin{pmatrix} u \\ v \end{pmatrix} \qquad \text{on } (0,T)\times\,]0,l[, \\[2ex] u(t,0) = Ev(t,0)\ \ ,\ \ v(t,l) = Du(t,l), \end{cases}$$

where

(v) $N_0(x) = \mathrm{diag}(N_{11}(x), N_{22}(x), \ldots, N_{qq}(x))$ *is a diagonal matrix per block whose elements* $N_{\eta\eta}(x)$ $(1 \leq \eta \leq q)$ *are* $m_\eta \times m_\eta$ *matrices* (m_η *is the algebraic multiplicity of the eigenvalue* $\lambda_\eta(x)$ (*or* $\mu_\eta(x)$)) *and each matrix* $N_{\eta\eta}(x)$ *is a block of the matrix* $N(x)$ *such that*

$$N_{\eta\eta}(x) = (n_{k,\,l})_{S_{\eta-1} \leq k,l \leq S_\eta} \qquad \textit{with } \; S_0 = 1, \; \textit{and} \;\; S_\eta = \sum_{d=1}^{\eta} m_d.$$

To illustrate this last assumption, if, for example, $M = \mathrm{diag}(1, 1, 2, -3, -3)$ and $N = (n_{ij})_{1 \leq i,j \leq 5}$, then the corresponding reduced matrix N_0 is given by

$$N_0 = \begin{pmatrix} n_{11} & n_{12} & 0 & 0 & 0 \\ n_{21} & n_{22} & 0 & 0 & 0 \\ 0 & 0 & n_{33} & 0 & 0 \\ 0 & 0 & 0 & n_{44} & n_{45} \\ 0 & 0 & 0 & n_{54} & n_{55} \end{pmatrix}.$$

In this study we prove that the two systems (2.1) and (2.2) have the same essential spectral radius.

Before giving the main result of this section, we need to recall some definitions and properties that may be found in Van Neerven [12, pp. 106–111].

Definition 2.1. (i) *If* L *is a linear bounded operator in a Banach space, the essential spectral radius of* L *is*

$$r_{ess}(L) := \inf \{ r > 0 : \lambda \in \sigma(L), |\lambda| \geq r; \; \textit{implies} \\ \lambda \textit{ is an isolated eigenvalue of finite multiplicity} \},$$

where $\sigma(L)$ *is the spectrum of* L.

(ii) *The type (or growth bound)* $\omega(T)$ *of a* C_0*-semigroup* $(T(t))$ *generated by* A *is*

$$\omega(T) := \inf \left\{ \omega \in \mathbb{R}, \, \exists M_\omega > 0, \; \|T(t)\| \leq M_\omega e^{\omega t} \quad \forall \, t \geq 0 \right\}.$$

A well-known result is that there exists a real number $\omega_{ess}(T)$ (the *essential type* of $T(t)$) such that

$$\omega_{ess}(T) := \frac{\ln [r_{ess}(T(t))]}{t}, \;\; t > 0.$$

The property below will play a significant role in our study:

$$r_{ess}(L + K) = r_{ess}(L) \quad \text{for any compact operator } K. \tag{2.3}$$

We recall also (see [12, pp. 106–111]) that the type $\omega(T)$ of a semigroup $(T(t))$ is given by

$$\omega(T) = \max \left[s(A)\, ; \, \omega_{ess}(T) \right],$$

where $s(A)$ is the spectral abscissa of A:

$$s(A) := \sup \{ \mathrm{Re}\, \lambda ; \lambda \in \sigma(A) \}.$$

Returning to our problem, let us introduce the new variable:

$$z(t) = v(t, l) - Du(t, l)$$

and define, on the energy space $H = \left[L^2([0,l])\right]^n \times \mathbb{C}^{n-p}$, the following operators:

$$A_1 \begin{pmatrix} u \\ v \\ z \end{pmatrix} = \left(\left[-M(x)\frac{\partial}{\partial x} - N(x)\right]\begin{pmatrix} u \\ v \end{pmatrix}, Fu(t,l) + Gv(t,l)\right) \tag{2.4}$$

and

$$A_4 \begin{pmatrix} u \\ v \end{pmatrix} = \left[-M(x)\frac{\partial}{\partial x} - N_0(x)\right]\begin{pmatrix} u \\ v \end{pmatrix} \tag{2.5}$$

whose domains are, respectively,

$$\begin{aligned} D(A_1) = \{(u,v,z) \in H\ ; (u,v) \in \left[W^{1,2}(0,l)\right]^n\ ; \\ u(0) = E\,v(0),\ z = v(l) - Du(l)\} \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} D(A_4) = \{(u,v) \in H\ ; (u,v) \in \left[W^{1,2}(0,l)\right]^n\ ; \\ u(0) = E\,v(0),\ v(l) = Du(l)\}\,. \end{aligned} \tag{2.7}$$

Equations (2.1) and (2.2) can be viewed as abstract systems

$$Y_t = A_1 Y \quad \text{in } H,$$

$$Z_t = A_4 Z \quad \text{in } \tilde{H},$$

where $\tilde{H} := \{(u,v,z) \in H;\ z \equiv 0\}$ and

$$A_1 = \begin{pmatrix} -M\dfrac{\partial}{\partial x} - N & 0 \\ 0 & R \end{pmatrix}, \qquad A_4 := -M\frac{\partial}{\partial x} - N_0$$

with $R : \mathbb{C}^{n-p} \longrightarrow \mathbb{C}$ and $Rz(t) = Fu(t,l) + Gv(t,l)$.

As in [9], under the assumptions (i)–(v), A_1 (resp., A_4) defined by (2.4) and (2.6) (resp., by (2.5) and (2.7)) generates a C_0-semigroup $T_1(t)$ on H (resp., ($T_4(t)$ on $\tilde{H}$).

Our main ingredient is the following.

LEMMA 2.2. *Suppose that the assumptions* (i)–(v) *hold; then the difference of the two semigroups $T_1(t)$ and $T_4(t)$ is a compact operator. In particular,*

$$r_{ess}(T_1(t)) = r_{ess}(T_4(t)).$$

Consequently,

$$\omega_{ess}(T_1) = \omega_{ess}(T_4).$$

Remark 2.3. This lemma has been proved by Neves, Ribeiro, and Lopes [9] in the case where the eigenvalues $\lambda_i(x)$ and $\mu_j(x)$ of the diagonal matrix $M(x)$ are all distinct (i.e., $m_i = 1$ for $i = 1, \dots, q$). In this case, these authors showed also that $\omega_{ess}(T_4) = s(A_4)$.

For the proof of Lemma 2.2, we use the same techniques as Neves, Ribeiro, and Lopes [9]. Note that the proof we propose for Lemma 2.2 works in the nonautonomous case as well.

Proof of Lemma 2.2. To simplify, we will prove Lemma 2.2 in the following situation:

- $M(x) = \mathrm{diag}(\lambda(x), \lambda(x), \mu(x))$; with $\lambda(x) > 0$ and $\mu(x) < 0$ for any x in$[0, l]$.
- $N(x) = (n_{ij}(x))_{1\leq i,j\leq 3}$.
- $D = \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}$, $F = \begin{pmatrix} f_1 & 0 \\ 0 & f_2 \end{pmatrix}$, $G = g$, and $E = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$, where d_1, d_2, f_1, f_2, g, e_1, and e_2 are real or complex constants.

We introduce two intermediate reduced systems:

$$(2.8) \qquad \begin{cases} \dfrac{\partial}{\partial t}\begin{pmatrix} u \\ v \\ w \end{pmatrix} = -M(x)\dfrac{\partial}{\partial x}\begin{pmatrix} u \\ v \\ w \end{pmatrix} - N_0(x)\begin{pmatrix} u \\ v \\ w \end{pmatrix} \qquad \text{on }]0, l[, \\ \\ \dfrac{d}{dt}\left[w(t,l) - d_1u(t,l) - d_2v(t,l)\right] = f_1u(t,l) + f_2v(t,l) + gw(t,l), \\ \\ u(t,0) = e_1w(t,0), \quad v(t,0) = e_2w(t,0), \end{cases}$$

$$(2.9) \qquad \begin{cases} \dfrac{\partial}{\partial t}\begin{pmatrix} u \\ v \\ w \end{pmatrix} = -M(x)\dfrac{\partial}{\partial x}\begin{pmatrix} u \\ v \\ w \end{pmatrix} - N_0(x)\begin{pmatrix} u \\ v \\ w \end{pmatrix} \qquad \text{on }]0, l[, \\ \\ \dfrac{d}{dt}\left[w(t,l) - d_1u(t,l) - d_2v(t,l)\right] = 0, \\ \\ u(t,0) = e_1w(t,0), \; v(t,0) = e_2w(t,0). \end{cases}$$

Denote by A_2 (resp., A_3) the associated operator of system (2.8) (resp., (2.9)). These two operators are defined on the same energy space and have the same domain as A_1. A_2 and A_3 generate, respectively, two C_0-semigroups which we will denote by $T_2(t)$ and $T_3(t)$.

We introduce the space $\bar{H} := \{(u, v, w, z) \in H \; ; \; z = 0\}$ and we define the orthogonal projection P of H on $\bar{H}$:

$$\begin{aligned} P : \qquad & H \longrightarrow \bar{H}, \\ (u, v, w, z) & \longmapsto (u, v, w, 0). \end{aligned}$$

We can identify $T_4(t)$ with $T_4(t)P$ which is an operator on H. It is enough to show that : $(T_1(t) - T_4(t)P)$ is compact.

We write

$$T_1(t) - T_4(t)P = (T_1(t) - T_2(t)) + (T_2(t) - T_3(t)) + (T_3(t) - T_4(t)P)\,.$$

We will show that each term of the right-hand side member is a compact operator.

Step 1. $(T_2(t) - T_3(t))$ *is a compact operator.*

To compute explicitly the semigroups $T_2(t)$ and $T_3(t)$, we follow [9] using the characteristics method. Taking the initial data $Y_0 = ((u_0, v_0), w_0, z_0)$ in H, we put $U = (u, v)$ and $(u_0, v_0) = U_0$. Denote by $z(t) = w(t,l) - d_1u(t,l) - d_2v(t,l)$ the new variable and by $[(u, v), w, z]$ the unique solution of system (2.2) with initial data Y_0.

Given a fixed point (t, x) in $(0, T)\times]0, l[$, let $\varphi(., t, x)$ (resp., $\psi(., t, x)$) be the unique solution of

$$\begin{cases} \dfrac{dx}{ds} = \lambda(x(s)), \\ x(t) = x, \end{cases} \qquad \text{resp.,} \begin{cases} \dfrac{dx}{ds} = \mu(x(s)), \\ x(t) = x. \end{cases}$$

We define the maps

$$\begin{aligned} \tau_i : [0,T] \times [0,l] &\longrightarrow \mathbb{R}, \\ (t,x) &\longrightarrow \tau_i(t,x) \quad i = 1,2, \\ \text{such that } \ \varphi(\tau_1(t,x),t,x) = 0 \ &\text{ and } \ \psi(\tau_2(t,x),t,x) = l. \end{aligned}$$

Denote by $R(t,y,x)$ the fundamental matrix associated with

$$\frac{d}{dt} Y(t) = N_0^{11}(x(t))Y(t),$$

where

$$N_0^{11}(x) = \begin{pmatrix} n_{11}(x) & n_{12}(x) \\ n_{21}(x) & n_{22}(x) \end{pmatrix} \ \text{ and } \ N_0(x) = \begin{pmatrix} N_0^{11}(x) & 0 \\ 0 & n_{33}(x) \end{pmatrix}.$$

Using the corresponding boundary conditions, the solution (U,w,z) is given by the following:

- If $0 \le x \le \varphi(t,0,0)$, then

$$\begin{aligned} U(t,x) = \exp\left[- \int_0^{\tau_1(t,x)} n_{33}(\psi(s,\tau_1(t,x),0))ds \right] R(t,\tau_1(t,x),x) \\ \times \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} w_0(\psi(0,\tau_1(t,x),0)). \end{aligned}$$

- If $\varphi(t,0,0) \le x \le l$, then

$$U(t,x) = R(t,0,x) \times U_0(\varphi(0,t,x)).$$

- If $0 \le x \le \psi(t,0,l)$, then

$$w(t,x) = \exp\left[- \int_0^t n_{33}(\psi(s,t,x))ds \right] w_0(\psi(0,t,x)).$$

- If $\psi(t,0,l) \le x \le l$, then

$$\begin{aligned} w(t,x) = \exp\left(- \int_{\tau_2(t,x)}^t n_{33}(\psi(s,t,x))ds \right) \\ \times \left[z(\tau_2(t,x)) \ + \ [d_1,d_2]R(\tau_2(t,x),0,l)U_0(\varphi(0,\tau_2(t,x),l)) \right]. \end{aligned}$$

- For any t in the interval $J = [0,T]$, we have

$$\begin{aligned} z(t) &= \int_0^t e^{gt-s}(gd_1 + f_1)u(t,l) + (gd_2 + f_2)v(t,l)ds + e^{tg}z_0 \\ &= \int_0^t e^{gt-s}(GD + F)U(s,l)ds + e^{tg}z_0. \end{aligned}$$

Consequently, $T_3(t)Y_0$ is obtained by cancelling the constants f_1, f_2, and g.

Thus

$$(T_2(t) - T_3(t))\left((u_0,v_0),w_0,z_0\right) = (0,\bar{w}(t,x),\bar{z}(t)),$$

where $\bar{w}$ and $\bar{z}$ are given by the following:

- If $0 \leq x \leq \psi(t,0,l)$, then

$$\bar{w}(t,x) = 0.$$

- If $\psi(t,0,l) \leq x \leq l$, then

$$\bar{w}(t,x) = \exp\left(- \int_{\tau_2(t,x)}^{t} n_{33}(\psi(s,t,x))ds \right) \bar{z}(\tau_2(t,x)).$$

- And for all $t \in J$, we have

$$\bar{z}(t) = \int_0^t e^{tg-s}(GD+F)R(s,0,l)U_0(\varphi(0,s,l))ds + (e^{tg}-1)z_0.$$

To prove the compactness of the difference $T_2(t) - T_3(t)$, it is sufficient to show that each component of $(0, \overline{w}, \overline{z}) = (T_2(t) - T_3(t))\, Y_0$ is a compact operator. Remark that the first component is null and that the operator defined by $z_0 \longrightarrow (e^{gt}-1)z_0$ is compact for any $t \in J = [0,T]$. In addition, in view of Lemma 4 in [9], the operator

$$\begin{aligned} U_0 \longrightarrow & \int_0^t \underbrace{e^{tg-s}\left[GD+F\right]R(s,0,l)}_{L(t,s)} U_0(\varphi(0,s,l))ds \\ = & \int_0^t L(t,s)U_0(\varphi(0,s,l))ds \end{aligned}$$

is compact since $L = (L_{ij})_{1 \leq i,j \leq 2}$ is such that the L_{ij} $(1 \leq i,j \leq 2)$ are continuous functions of their arguments and

$$U_0(\varphi(0,s,l)) = \begin{pmatrix} u_0(\varphi(0,s,l)) \\ \\ v_0(\varphi(0,s,l)) \end{pmatrix}$$

$$\text{since} \quad : \frac{d}{ds}(\varphi(0,s,l)) = -\frac{\partial \varphi}{\partial x}(0,s,l)\,\lambda(l) \neq 0.$$

Consequently, $T_2(t) - T_3(t)$ is a compact operator for all t in the compact interval J.

Step 2. *$(T_1(t) - T_2(t))$ is a compact operator.*

For this we write, for a given initial data $Y_0 = ((u_0, v_0), w_0, z_0)$, the differential equation corresponding with system (2.1) in the form

$$\frac{\partial}{\partial t}Y(t) = A_1 Y(t) = A_2 Y(t) + BY(t),$$

where

$$B = \begin{pmatrix} N - N_0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Then we have

$$Y(t) = T_1(t)Y_0 = T_2(t)Y_0 + \int_0^t T_2(t-s)BY(s)ds.$$

Consequently,

$$\begin{aligned}(T_1(t)-T_2(t))Y_0 &= \int_0^t T_2(t-s)BT_1(s)Y_0ds\\ &= \int_0^t T_2(t-s)BT_2(s)Y_0ds + \int_0^t T_2(t-s)B(T_1(s)-T_2(s))Y_0ds.\end{aligned}$$

Lemma 2.4 (see [9, Lemma 3]). *Suppose that $\int_0^t T_2(t-s)BT_2(s)ds$ is a compact operator for all $t \in J$; then $(T_1(t)-T_2(t))$ is a compact operator for all t in J. In fact, $\{(T_1(t)-T_2(t))\,Y;\ \|Y\|_H \le 1,\ t\in J\}$ is precompact.*

To show that $\{(T_1(t)-T_2(t))\,Y; \|Y\|_H \le 1, t\in J\}$ is precompact, it is enough, according to Lemma 2.4, to prove that $\int_0^t T_2(t-s)BT_2(s)ds$ is a compact operator. However, we know that

$$\begin{aligned}T_2(t-s)BT_2(s) = T_3(t-s)BT_3(s) + \left[T_2(t-s)-T_3(t-s)\right]BT_3(s)\\ +T_2(t-s)B\left[T_2(s)-T_3(s)\right]\end{aligned}$$

As $T_2(t)-T_3(t)$ is a compact operator, $\int_0^t T_2(t-s)BT_2(s)ds$ is compact if $\int_0^t T_3(t-s)BT_3(s)ds$ is. For this, we are going to use Lemma 4 of [9]. Let $Y_0 = ((u_0,v_0),w_0,z_0) = (U_0,w_0,z_0)$ be in H. We wish to prove that each integral component of $\int_0^t T_3(t-s)BT_3(s)Y_0ds$ defines a compact operator. Note that the third component $\bar{z}$ defines an operator of finite dimensional range; thus it is compact. We can write

$$\int_0^t T_3(t-s)BT_3(s)Y_0ds = \int_0^t \begin{pmatrix}\bar{U}(t,s,x)\\ \bar{w}(t,s,x)\\ \bar{z}(t,s,x)\end{pmatrix} ds.$$

We deduce that $\bar{U}(t,s)$ is the first component of $\bar{Y}(t,s)$ which is the unique solution of the following Cauchy problem:

$$\begin{cases}\dfrac{d}{dt}Y(t) = A_3Y(t),\\ \\ Y(s)=\tilde{Y}(s).\end{cases}$$

For $\boldsymbol{0 \le x \le \varphi(t,0,0)}$.

We have

$$\int_0^t \bar{U}(t,s,x)ds = \int_0^{\tau_1(t,x)} \bar{U}(t,s,x)ds + \int_{\tau_1(t,x)}^t \bar{U}(t,s,x)ds$$

In the same way as [9], we find

$$\begin{aligned}\int_0^{\tau_1(t,x)} \bar{U}(t,s,x)ds = \int_0^{s_1} \xi_1(t,s,x)\ U_0(\varphi(0,s,x_\psi))ds + \int_{s_1}^{\tau_1(t,x)}\\ \left[\xi_2(t,s,x)\begin{pmatrix}e_1\\ e_2\end{pmatrix} w_0(\psi(0,\tau_1(s,x_\psi),0))ds\right].\end{aligned}$$

And

$$\begin{aligned}\int_{\tau_1(t,x)}^t \bar{U}(t,s,x)ds = \int_{\tau_1(t,x)}^{s_2} \xi_3(t,s,x)\begin{pmatrix}n_{13}\\ n_{23}\end{pmatrix} w_0(\psi(0,s,x_\varphi))ds\\ +\int_{s_2}^t \xi_4(t,s,x)U_0(\varphi(0,\tau_2(s,x_\varphi),l))ds,\end{aligned}$$

where s_1 and s_2 are the two reals satisfying

$$\begin{cases} \varphi(s_1, 0, 0) = \psi(s_1, \tau_1(t, x), 0) := x_\psi, \\ \\ \psi(s_2, 0, l) = \varphi(s_2, t, x) := x_\varphi \end{cases}$$

and $(\xi_i)_{i=1}^4$ are continuous functions of their arguments.

For $\boldsymbol{\varphi(t, 0, 0) \leq x \leq l}$.

It is easy to see that it is a particular case of that previously treated ($0 \leq x \leq \varphi(t, 0, 0)$).

To conclude, we have

$$\begin{aligned} \frac{d\varphi}{ds}(0, s, x_\psi) &= \frac{\partial\varphi}{\partial s}(0, s, x_\psi) + \frac{\partial\varphi}{\partial x}(0, s, x_\psi)\frac{\partial x_\psi}{\partial s} \\ &= \frac{\partial\varphi}{\partial x}(0, s, x_\psi)\ [\mu(x_\psi) - \lambda(x_\psi)]\,. \end{aligned}$$

However, $\mu(x_\psi) < 0$ and $\lambda(x_\psi) > 0$, thus

$$\frac{d\varphi}{ds}(0, s, x_\psi) \neq 0.$$

In addition we have

$$\frac{d\psi}{ds}(0, \tau_1(s, x_\psi), 0) = \frac{\partial\psi}{\partial\tau_1}(0, \tau_1(s, x_\psi), 0) \times \left[\frac{\partial\tau_1}{\partial s}(s, x_\psi) + \frac{\partial\tau_1}{\partial s}(s, x_\psi)\frac{\partial x_\psi}{\partial s}\right].$$

However,

$$\frac{\partial\psi}{\partial\tau_1}(0, \tau_1(s, x_\psi), 0) = -\frac{\partial\psi}{\partial x}(0, \tau_1(s, x_\psi), 0)\,\mu(0) \quad \text{and} \quad \frac{\partial x_\psi}{\partial s} = \mu(x_\psi).$$

Finally,

$$\begin{aligned} \frac{d\psi}{ds}(0, \tau_1(s, x_\psi), 0) = &-\frac{\mu(0)}{\lambda(x_\psi)}\frac{\partial\psi}{\partial x}(0, \tau_1(s, x_\psi), 0) \times \frac{\partial\varphi}{\partial x}(\tau_1(s, x_\psi), s, x_\psi) \\ &\times [\lambda(x_\psi) - \mu(x_\psi)]\,. \end{aligned}$$

In particular,

$$\frac{d\psi}{ds}(0, \tau_1(s, x_\psi), 0) \neq 0.$$

In the same way, we get

$$\frac{d\psi}{ds}(0, s, x_\varphi) \neq 0 \quad \text{and} \quad \frac{d\varphi}{ds}(0, \tau_2(s, x_\varphi), l) \neq 0.$$

We can, according to Lemma 4 of [9], conclude that $\int_0^t \bar{U}(t, s, x)ds$ defines a compact operator.

The second integral component $\int_0^t \bar{w}(t, s, x)ds$ can be treated by similar techniques. Hence $\int_0^t T_2(t-s)BT_2(s)ds$ is compact for any $t \in J$. Thus, from Lemma 2.4, we deduce that $(T_1(t) - T_2(t))$ is a compact operator for any $t \in J$ and that $\{(T_1(t) - T_2(t))Y,\ \|Y\|_H \leq 1,\ t \in J\}$ is precompact.

Finally, $(T_3(t) - T_4(t)P)$ is compact. Indeed, we have

$$T_4(t)P = T_3(t)P \text{ since } A_3 = (A_4, 0).$$

Then $(T_3(t) - T_4(t)P) = T_3(t)(I - P)$ is an operator of finite dimensional range, thus it is compact. Moreover, $\{(T_3(t) - T_4(t)P)Y,\ \|Y\|_H \leq 1,\ t \in J\}$ is precompact.

Hence, Lemma 2.2 is proved. □

Remark 2.5. Lemma 2.2 holds also in the case where the matrices M, N, D, E, F, and G are dependent on the time.

3. Proofs. To prove our two theorems, we introduce the following new variables:

$$\left\{ \begin{array}{ll} p = u_t - u_x; & q = u_t + u_x, \\ & \\ r = v_t - \eta v_x; & s = v_t + \eta v_x. \end{array} \right. \tag{3.1}$$

Our system becomes

$$\frac{\partial}{\partial t}\begin{pmatrix} U \\ V \end{pmatrix} + M\frac{\partial}{\partial x}\begin{pmatrix} U \\ V \end{pmatrix} + N(x)\begin{pmatrix} U \\ V \end{pmatrix} = 0, \tag{3.2}$$

where

$$U = \begin{pmatrix} p \\ r \end{pmatrix}; V = \begin{pmatrix} q \\ s \end{pmatrix} \text{ and } M = \text{diag}(1, \eta; -1, -\eta) \text{ with}$$

$$N(x) = \frac{1}{2}\begin{pmatrix} 0 & -b(x) & 0 & -b(x) \\ b(x) & a(x) & b(x) & a(x) \\ 0 & -b(x) & 0 & -b(x) \\ b(x) & a(x) & b(x) & a(x) \end{pmatrix}. \tag{3.3}$$

Now the boundary conditions in (1.3) transform into

$$U(t,0) = -V(t,0), \quad U(t,1) = -V(t,1), \qquad t > 0, \tag{3.4}$$

and the boundary conditions in (1.8) into

$$U(t,0) = -V(t,0),$$

$$\begin{pmatrix} 1 & \eta \\ \alpha & \eta - \alpha \end{pmatrix} U(t,1) = \begin{pmatrix} -1 & \eta \\ \alpha & \eta - \alpha \end{pmatrix} V(t,1), \tag{3.5}$$

and (3.2) will represent system (1.8) if we set $a \equiv 0$ in (3.3). The equivalence of the transformed systems with our initial systems clearly holds.

In view of the proof of Lemma 2.2, we will set

$$A_1 = -M\frac{\partial}{\partial x} - N(x)$$

with, for the proof of Theorem 1.1, the associated boundary conditions

$$U(0) = -V(0), \quad U(1) = -V(1) \tag{3.6}$$

and, for the proof of Theorem 1.4, the associated boundary conditions

$$U(0) = -V(0),$$

$$\begin{pmatrix} 1 & \eta \\ \alpha & \eta - \alpha \end{pmatrix} U(1) = \begin{pmatrix} -1 & \eta \\ \alpha & \eta - \alpha \end{pmatrix} V(1). \tag{3.7}$$

(Note that in this system, $a \equiv 0$.)

Proof of Theorem 1.1. Assume first that $\eta \neq 1$. According to Lemma 2.2 (see also [9]), the semigroup associated with (3.2), (3.4) has the same essential type as the semigroup associated with the system

$$\begin{gathered} \frac{\partial}{\partial t}\begin{pmatrix} U \\ V \end{pmatrix} + M\frac{\partial}{\partial x}\begin{pmatrix} U \\ V \end{pmatrix} + \widetilde{N}(x)\begin{pmatrix} U \\ V \end{pmatrix} = 0, \\ U(t,0) = -V(t,0), \quad U(t,1) = -V(t,1), \qquad t > 0, \end{gathered} \tag{3.8}$$

with $\widetilde{N}(x) = \operatorname{diag}(0, a(x), 0, a(x))$. We set

$$A_4 = -M\frac{\partial}{\partial x} - \widetilde{N}(x)$$

with the associated boundary conditions (3.6). It is sufficient (see Remark 2.3) to prove that the spectral abscissa $s(A_4)$ is 0. So, let us consider the system

$$(\lambda - A_4)\begin{pmatrix} U \\ V \end{pmatrix} = 0,$$

$$U(0) = -V(0), \qquad U(1) = -V(1).$$

Computing the solutions of this last system, it is easy to prove that λ is an eigenvalue if and only if it satisfies

$$\left(e^{2\lambda} - 1\right)\left(e^{\frac{2}{\eta}(\lambda + \overline{a})} - 1\right) = 0.$$

Thus, the eigenvalues are

$$\begin{aligned} &\lambda = ik\pi, \ k \in \mathbb{Z}^*, \\ &\lambda = -\overline{a} + ik\eta\pi, \qquad k \in \mathbb{Z}, \end{aligned}$$

where $\overline{a} = \int_0^1 a(x)dx$. This proves that $s(A_4) = 0$ and concludes the proof of (i).

Let us now consider the case $\eta = 1$. We set for simplicity $\{x, a(x) \neq 0\} = [\alpha, \beta]$ and $\{x, b(x) \neq 0\} = [\delta, \gamma]$ with $\beta < \delta$ and $(\alpha, \beta, \delta, \gamma) \in ([0,1])^4$.

According to Lemma 2.2, let $A_4 = -M\frac{\partial}{\partial x} - N_0(x)$ be the reduced operator associated with A_1, where

$$N_0(x) = \frac{1}{2}\begin{pmatrix} 0 & -b(x) & 0 & 0 \\ b(x) & a(x) & 0 & 0 \\ 0 & 0 & 0 & -b(x) \\ 0 & 0 & b(x) & a(x) \end{pmatrix}. \tag{3.9}$$

We denote by $T_4(t)$ the C_0-semigroup generated by A_4.

Lemma 2.2 implies that

$$\omega_{ess}(T_1) = \omega_{ess}(T_4).$$

Now we compute $s(A_4)$. Given $Y = (U, V) \in D(A_4)$,

$$(\lambda I - A_4)\begin{pmatrix} U \\ V \end{pmatrix} = 0$$

or

$$\begin{cases} \dfrac{dU}{dx}(x) = -(\lambda I + N_{11}(x))U(x), \\[2ex] \dfrac{dV}{dx}(x) = (\lambda I + N_{11}(x))V(x), \end{cases}$$

where

$$N_{11}(x) = \frac{1}{2}\begin{pmatrix} 0 & -b(x \\ b(x) & a(x) \end{pmatrix}.$$

Solving the differential system above and taking into account the boundary conditions leads to

$$P(\lambda)V(0) := \begin{pmatrix} (e^{-\lambda} - e^{\lambda})\cos\frac{\bar{b}}{2} & (e^{-\lambda-\frac{\bar{a}}{2}} + e^{\lambda+\frac{\bar{a}}{2}})\sin\frac{\bar{b}}{2} \\[2ex] -(e^{-\lambda} + e^{\lambda})\sin\frac{\bar{b}}{2} & (e^{-\lambda-\frac{\bar{a}}{2}} + e^{\lambda+\frac{\bar{a}}{2}})\cos\frac{\bar{b}}{2} \end{pmatrix} V(0) = 0.$$

Consequently, it is clear that

$$\begin{aligned} &\lambda \in \sigma(A_4) \Leftrightarrow \det P(\lambda) = 0, \\ &\quad \Leftrightarrow e^{4\lambda} - (e^{-\bar{a}} + 1)\cos\left(\bar{b}\right)\, e^{2\lambda} + e^{-\bar{a}} = 0. \end{aligned}$$

Let $\delta = (e^{-\bar{a}} + 1)^2 \cos^2\left(\bar{b}\right) - 4e^{-\bar{a}}$. Thus, if we put

$$\begin{cases} x_1 = \frac{1}{2}(e^{-\bar{a}} + 1)\cos\left(\bar{b}\right) + \dfrac{\sqrt{\delta}}{2}, \\[2ex] x_2 = \frac{1}{2}(e^{-\bar{a}} + 1)\cos\left(\bar{b}\right) - \dfrac{\sqrt{\delta}}{2}, \end{cases}$$

then we have

$$s(A_4) = \begin{cases} \left.\begin{array}{ll} \frac{1}{2}\ln x_1 & \text{if } \cos\left(\bar{b}\right) \geq 0, \\[2ex] \frac{1}{2}\ln(-x_2) & \text{if } \cos\left(\bar{b}\right) \leq 0 \end{array}\right\} \text{ if } \delta \geq 0, \\[3ex] -\frac{1}{4}\bar{a} = -\frac{1}{4}\int_0^1 a(t)dt \qquad\qquad \text{if } \delta < 0. \end{cases}$$

Remark that if $\bar{a} \leq 0$, then $s(A_4) \geq 0$ and we deduce the following.

If $\bar{a} > 0$, then $s(A_4) \leq 0$ and $s(A_4) = 0$ if and only if $\cos\left(\bar{b}\right) = \pm 1$, that is, if and only if $\bar{b} \in \pi\mathbb{Z}$.

A simple computation shows that the eigenvalues of A_4 are distributed on at most two vertical axes.

We conclude with the help of the following result.

LEMMA 3.1 (Renardy [10, Theorem 1, p. 1300]). *Let H be a Hilbert space and let $L = L_0 + B$ be the infinitesimal generator of a C_0-semigroup of operators in H. Assume that L_0 is normal and B is bounded. Assume that there exists a number $M > 0$ and an integer n such that the following hold:*

(a) *If $\lambda \in \sigma(L_0)$ and $\mid \lambda \mid > M - 1$, then λ is an isolated eigenvalue of finite multiplicity.*

(b) *If $\mid z \mid > M$, then the number of eigenvalues of L_0 in the unit disk centered at z (counted by multiplicity) does not exceed n.*

Then $\omega_{ess}(e^{Lt}) \leq s(L)$.

We apply this lemma with $L = A_4$ (defined by (2.5) and (2.7)), $L_0 = -M\frac{\partial}{\partial x}$ with $D(L_0) = D(A_4)$ and $B = N_0$. Since M is a constant matrix, it is easy to see that L_0 is normal. Its eigenvalues are $\lambda_k = ik\pi, k \in \mathbb{Z}$, and their (algebraic and geometric) multiplicities are equal to 2. We can then take $n = 2$ and since $\mid \lambda_{k+1} - \lambda_k \mid = \pi$, assertion (b) in Lemma 3.1 is satisfied. Thus, $\omega_{ess}(T_4) \leq s(A_4) < 0$. Note that, according to the definition of the essential spectral radius and the previous inequality, one deduces that $s(A_4) = \omega_{ess}(T_4)$. □

Proof of Theorem 1.4. (i) Assume first that $\eta = \frac{2p+1}{q}$ for some $(p, q) \in \mathbb{Z} \times \mathbb{Z}^*$ with $\eta \neq 1$.

As in the previous proof, computing the essential type amounts to computing the eigenvalues λ of the reduced system, namely,

$$\lambda \begin{pmatrix} U \\ V \end{pmatrix} + M\frac{\partial}{\partial x}\begin{pmatrix} U \\ V \end{pmatrix} = 0$$

with the boundary conditions

$$U(0) = -V(0),$$

$$\begin{pmatrix} 1 & \eta \\ \alpha & \eta - \alpha \end{pmatrix} U(1) = \begin{pmatrix} -1 & \eta \\ \alpha & \eta - \alpha \end{pmatrix} V(1).$$

It is then easy to see that λ is an eigenvalue if and only if it satisfies the equation

$$\alpha \sinh \lambda \sinh \frac{\lambda}{\eta} + \eta \sinh \lambda \cosh \frac{\lambda}{\eta} + \alpha\eta \cosh \lambda \cosh \frac{\lambda}{\eta} = 0. \tag{3.10}$$

To prove the assertion of the theorem, we proceed by contradiction. Assume that there exists a sequence (λ_n) of eigenvalues such that

$$\lim_{n\to\infty} \operatorname{Re} \lambda_n = 0. \tag{3.11}$$

Let us set $\lambda_n = x_n + iy_n$. Using the relations

$$\sinh(a + ib) = \sinh a \cos b + i \cosh a \sin b,$$
$$\cosh(a + ib) = \cosh a \cos b + i \sinh a \sin b,$$

(3.10), and (3.11), we get

$$-\alpha \sin y_n \sin \frac{y_n}{\eta} + i\eta \sin y_n \cos \frac{y_n}{\eta} + \alpha\eta \cos y_n \cos \frac{y_n}{\eta} \underset{n\to\infty}{\to} 0.$$

This last condition is equivalent to

$$\sin y_n \cos \frac{y_n}{\eta} \underset{n\to\infty}{\longrightarrow} 0 \tag{3.12}$$

and

$$\eta \cos y_n \cos \frac{y_n}{\eta} - \sin y_n \sin \frac{y_n}{\eta} \underset{n\to\infty}{\longrightarrow} 0. \tag{3.13}$$

From (3.12), it follows that either $\underset{n\to\infty}{\longrightarrow} 0$ or $\cos \frac{y_n}{\eta} \underset{n\to\infty}{\longrightarrow} 0$. If one of the alternatives holds, (3.13) will imply the second one. We deduce that (3.12)–(3.13) are equivalent to

$$\sin y_n \underset{n\to\infty}{\longrightarrow} 0 \text{ and } \cos \frac{y_n}{\eta} \underset{n\to\infty}{\longrightarrow} 0. \tag{3.14}$$

Now, if (y_n) is bounded, there exists a subsequence that converges to y such that

$$\sin y = 0 \text{ and } \cos \frac{y}{\eta} = 0.$$

But this is possible if and only if there exists $(k, j) \in \mathbb{Z}^2$ such that $\eta = \frac{2k}{2j+1}$. This contradicts our assumption on η. It follows that (y_n) is unbounded and we may assume that $\mid y_n \mid \underset{n\to\infty}{\longrightarrow} \infty$. In this case, (3.14) is equivalent to the existence of a sequence $(k_n, j_n)_n \subset \mathbb{Z} \times \mathbb{Z}$ such that

$$k_n - \eta \left(j_n + \frac{1}{2} \right) \to 0 \text{ as } n \to \infty$$

or equivalently

$$k_n - \eta j_n \to \frac{\eta}{2} \text{ as } n \to \infty. \tag{3.15}$$

Let us consider the set

$$G = \left\{ k - \eta j, (k, j) \in \mathbb{Z}^2 \right\}. \tag{3.16}$$

It is an additive subgroup of $\mathbb{R}$. A well-known result of algebra asserts that either there exists a real number $a > 0$ such that $G = a\mathbb{Z}$ or $\overline{G} = \mathbb{R}$.

The first alternative holds if and only if $a \in \mathbb{Q}$ since $\mathbb{Z} \subset G$. On the other hand, G is closed in $\mathbb{R}$ and (3.15) holds if and only if $\frac{\eta}{2} \in G$. This means that there exists $j \in \mathbb{Z}$ such that $\eta = 2aj = \frac{2p}{q}$ for some $(p, q) \in \mathbb{Z} \times \mathbb{Z}^*$. But this contradicts the form of η.

The second alternative holds if and only if $\eta \in \mathbb{R} \backslash \mathbb{Q}$. So the sufficiency part is proved.

Assume now that $\eta \neq \frac{2p+1}{q}$ for all $(p, q) \in \mathbb{Z} \times \mathbb{Z}^*$. Equation (3.10) rewrites

$$(\alpha + \eta + \alpha\eta) e^{2\lambda(1+\frac{1}{\eta})} - (\alpha + \eta - \alpha\eta) e^{2\frac{\lambda}{\eta}} - (\alpha - \eta - \alpha\eta) e^{2\lambda} + \alpha - \eta + \alpha\eta = 0.$$

We set

$$\begin{aligned} f(\lambda) &= a_0 e^{2\lambda(1+\frac{1}{\eta})} - a_1 e^{2\frac{\lambda}{\eta}} - a_2 e^{2\lambda} + a_3, \quad \lambda \in \mathbb{C}, \\ a_0 &= \alpha + \eta + \alpha\eta, \quad a_1 = \alpha + \eta - \alpha\eta, \\ a_2 &= \alpha - \eta - \alpha\eta, \quad a_3 = \alpha - \eta + \alpha\eta. \end{aligned}$$

We will apply Rouché's theorem to f to prove that there exists a sequence of eigenvalues such that (3.11) holds true. According to the previous computations, there exists a sequence $(k_n, j_n) \in \mathbb{Z}^2$ such that

$$\varepsilon_n := (2k_n - (2j_n + 1)\eta)\,\frac{\pi}{2} \underset{n\to\infty}{\to} 0. \tag{3.17}$$

Let us set $\lambda_n = ik_n\pi$. To achieve our goal, it suffices to prove that there exists a sequence of positive numbers r_n such that $\lim_{n\to\infty} r_n = 0$ and

$$|f(\lambda) - f'(\lambda_n)(\lambda - \lambda_n)| < |f'(\lambda_n)(\lambda - \lambda_n)| \text{ if } |\lambda - \lambda_n| = r_n. \tag{3.18}$$

To estimate the left-hand term in the previous inequality, we have from Taylor's formula

$$|f(\lambda) - f'(\lambda_n)(\lambda - \lambda_n)| \le |f(\lambda_n)| + \sum_{p\ge 2} \frac{\left|f^{(p)}(\lambda_n)\right|}{p!}\,|\lambda - \lambda_n|^p\,. \tag{3.19}$$

We first have, using (3.17) and noting that $a_0 - a_1 = -\ a_2 + a_3 = 2\alpha\eta$

$$\begin{aligned} |f(\lambda_n)| &= \left|(a_0 - a_1)e^{2i\frac{\varepsilon_n}{\eta}} + a_2 - a_3\right| \\ &= 2\alpha\eta\left|e^{2i\frac{\varepsilon_n}{\eta}} - 1\right| \le 4\alpha\,|\varepsilon_n|\,. \end{aligned} \tag{3.20}$$

Moreover, for $p \ge 1$

$$f^{(p)}(\lambda_n) = -2^p\left[\left(a_0\left(1 + \frac{1}{\eta}\right)^p - \frac{a_1}{\eta^p}\right)e^{2i\frac{\varepsilon_n}{\eta}\pi} + a_2\right] \tag{3.21}$$

and for any $p \ge 2$

$$\begin{aligned} \left|f^{(p)}(\lambda_n)\right| &= 2^p\left|\left(a_0\left(1 + \frac{1}{\eta}\right)^p - \frac{a_1}{\eta^p}\right)e^{2i\frac{\varepsilon_n}{\eta}\pi} + a_2\right| \\ &\le 2^p\left(|a_0|\left(1 + \frac{1}{\eta}\right)^p + \frac{|a_1|}{\eta^p} + |a_2|\right). \end{aligned} \tag{3.22}$$

Thus (3.19), (3.20), and (3.22) imply

$$\begin{aligned} |f(\lambda) - f'(\lambda_n)(\lambda - \lambda_n)| &\le |a_0|\left(e^{2r_n(1+\frac{1}{\eta})} - 2r_n\left(1 + \frac{1}{\eta}\right) - 1\right) \\ &\quad + |a_1|\left(e^{2\frac{r_n}{\eta}} - 2\frac{r_n}{\eta} - 1\right) + |a_2|\left(e^{2r_n} - 2r_n - 1\right) \\ &\le 4\left(\left(1 + \frac{1}{\eta}\right)^2|a_0| + \frac{|a_1|}{\eta^2} + |a_2|\right)r_n^2 \quad \text{for } n \ge n_0. \end{aligned}$$

On the other hand, it is easy to see, using the definition of the a_i, that

$$\begin{aligned} |f'(\lambda_n)(\lambda - \lambda_n)| &= 2\Bigg[\left(\frac{a_1 - a_0}{\eta} - a_0\right)^2 - 2a_2\left(\frac{a_1 - a_0}{\eta} - a_0\right)\cos\left(2\frac{\varepsilon_n}{\eta}\right) \\ &\qquad + a_2^2\Bigg]^{1/2} r_n \\ &\ge 2\alpha r_n. \end{aligned}$$

We then need, in order to satisfy (3.18), to find r_n such that, for n sufficiently large

$$4\alpha\left|\varepsilon_n\right| + 4\left(\left(1+\frac{1}{\eta}\right)^2 |a_0| + \frac{|a_1|}{\eta^2} + |a_2|\right) r_n^2 < 2\alpha r_n.$$

Choosing $r_n = \sqrt{|\varepsilon_n|}$, there will exist $N \in \mathbb{N}$ such that for all $n \geq N$

$$|f(\lambda) - f'(\lambda_n)(\lambda-\lambda_n)| < |f'(\lambda_n)(\lambda-\lambda_n)| \quad \text{for} \quad |\,\lambda - \lambda_n| = \sqrt{|\varepsilon_n|}\,.$$

This ends the proof of the point (i) in Theorem 1.4.

(ii) Assume that $\eta = 1$.

As in the previous proof, computing the essential type amounts to computing the eigenvalues λ of the reduced system, namely,

$$\lambda\begin{pmatrix} U \\ V \end{pmatrix} + M\frac{\partial}{\partial x}\begin{pmatrix} U \\ V \end{pmatrix} + N_0\begin{pmatrix} U \\ V \end{pmatrix} = 0$$

with the boundary conditions

$$U(0) = -V(0),$$
$$\begin{pmatrix} 1 & 1 \\ \alpha & 1-\alpha \end{pmatrix} U(1) = \begin{pmatrix} -1 & 1 \\ \alpha & 1-\alpha \end{pmatrix} V(1).$$

Here

$$N_0(x) = \frac{1}{2}\begin{pmatrix} 0 & -b(x) & 0 & 0 \\ b(x) & 0 & 0 & 0 \\ 0 & 0 & 0 & -b(x) \\ 0 & 0 & b(x) & 0 \end{pmatrix}.$$

The eigenvalues satisfy the equation

$$e^{4\lambda} + \frac{4\alpha}{2\alpha+1}\sin(\bar{b})e^{2\lambda} + \frac{2\alpha-1}{2\alpha+1} = 0.$$

Let $\delta = \frac{4\alpha^2}{(1+2\alpha)^2}\sin^2(\bar{b}) - \frac{2\alpha-1}{2\alpha+1}$. Then we have the following.

- If $\delta < 0$, then all solutions satisfy $e^{2\operatorname{Re}\lambda} = \left|e^{2\lambda}\right| = \frac{2\alpha-1}{2\alpha+1}$. It follows that $\operatorname{Re}\lambda = \frac{1}{2}\ln\frac{2\alpha-1}{2\alpha+1} < 0$.
- If $\delta \geq 0$, then $e^{2\operatorname{Re}\lambda} = |e^{2\lambda}| = |-\frac{2\alpha}{2\alpha+1}\sin(\bar{b}) \pm \sqrt{\delta}|$. Thus

$$\operatorname{Re}\lambda = \frac{1}{2}\ln\left|-\frac{2\alpha}{2\alpha+1}\sin(\bar{b}) \pm \sqrt{\delta}\right| \leq 0.$$

Consequently,

$$s(A_4) = \begin{cases} \left.\begin{array}{ll} \frac{1}{2}\ln(-\frac{2\alpha}{2\alpha+1}\sin(\bar{b}) + \sqrt{\delta}) & \text{if } \sin(\bar{b}) \leq 0 \\ \frac{1}{2}\ln(\frac{2\alpha}{2\alpha+1}\sin(\bar{b}) + \sqrt{\delta}) & \text{if } \sin(\bar{b}) \geq 0 \end{array}\right. & \text{if } \delta \geq 0, \\ \frac{1}{2}\ln\frac{2\alpha-1}{2\alpha+1} & \text{if } \delta < 0. \end{cases}$$

A simple computation shows that

$$s(A_4) = 0 \text{ if and only if } \sin(\bar{b}) = \pm 1.$$

We deduce that

$$s(A_4) < 0 \text{ if and only if } \bar{b} \neq (2k+1)\frac{\pi}{2} \quad \forall\ k \in \mathbb{Z}.$$

Next we use Lemma 3.1 and the definition of essential spectral radius; we get (in the same way as in the proof of Theorem 1.1)

$$\omega_{ess}(T_1) = \omega_{ess}(T_4) = s(A_4).$$

We conclude that the semigroup $T_1(t)$ is exponentially stable in an invariant subspace of finite codimension.

REFERENCES

[1] M. Afilal and F. Ammar Khodja, *Stability of coupled second order equations*, Comput. Appl. Math., to appear.

[2] F. Ammar Khodja and A. Bader, *Sur le comportement asymptotique de solutions de systèmes hyperboliques linéaires*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 957–960.

[3] F. Ammar Khodja, A. Bader, and A. Benabdallah, *Dynamical stabilization of systems via decoupling techniques*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 577–594.

[4] S. Avodin and M. Tucsnak, *Simultaneous Controllability in Sharp Time for Elastic Strings*, Department of Mathematics, University of Nancy-I, France, preprint.

[5] D. B. Henry, O. Lopes, and A. Perissinitto, *On the essential spectrum of semigroup of thermoelasticity*, Nonlinear Anal., 21 (1993), pp. 65–75.

[6] B. Kapitonov, *Uniform stabilization and exact controllability for a class of coupled hyperbolic systems*, Comput. Appl. Math., 15 (1996), pp. 199–212.

[7] Z. Liu and J. Yong, *Qualitative properties of certain C_0 semigroups arising in elastic systems with various damping*, Adv. Differential Equations, 3 (1998), pp. 643–686.

[8] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*, Tome 1, Masson, Paris, 1988.

[9] A. F. Neves, H. De Souza Ribeiro, and O. Lopes, *On the spectrum of evolution operators generated by hyperbolic systems*, J. Funct. Anal., 67 (1986), pp. 320–344.

[10] M. Renardy, *On the type of certain C_0-semigroups*, Comm. Partial Differential Equations, 18 (1993), pp. 1299–1307.

[11] A. Soufyane, *Stabilisation de la poutre de Timoshenko*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 731–734.

[12] J. Van Neerven, *The Asymptotic Behaviour of Semigroups of Linear Operators*, Oper. Theory Adv. Appl. 88, Birkhaüser Verlag, Basel, 1996.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1852–1873

VARIATIONAL INEQUALITY PROBLEMS WITH A CONTINUUM OF SOLUTIONS: EXISTENCE AND COMPUTATION*

P. JEAN-JACQUES HERINGS†, DOLF TALMAN‡, AND ZAIFU YANG§

Abstract. In this paper three sufficient conditions are provided under each of which an upper semicontinuous point-to-set mapping defined on an arbitrary polytope has a connected set of zero points that connect two distinct faces of the polytope. Furthermore, we obtain an existence theorem of a connected set of solutions to a nonlinear variational inequality problem over arbitrary polytopes. These results follow in a constructive way by designing a new simplicial algorithm. The algorithm operates on a triangulation of the polytope and generates a piecewise linear path of points connecting two distinct faces of the polytope. Each point on the path is an approximate zero point. As the mesh size of the triangulation goes to zero, the path converges to a connected set of zero points linking the two distinct faces. As a consequence, our results generalize Browder's fixed point theorem [*Summa Brasiliensis Mathematicae*, 4 (1960), pp. 183–191] and an earlier result by the authors [*Math. Oper. Res.*, 21 (1996), pp. 675–696] on the n-dimensional unit cube. An application in economics and some numerical examples are also discussed.

Key words. polytope, simplicial algorithm, continuum of zero points, system of nonlinear equations, variational inequality, economic equilibrium model

AMS subject classifications. Primary, 49D35, 90A14; Secondary, 90C30, 90C33

PII. S0363012999360592

1. Introduction.

Whenever a mathematical model of some phenomenon is constructed, for instance, in engineering or in economics, the first question to ask is whether a solution to the model exists. A very powerful tool that is used to this end is Brouwer's fixed point theorem; see Brouwer [3]. When the model is not a system of equations but a system of correspondences, Kakutani's fixed point theorem [16] is invoked. An alternative to fixed point theorems consists of using intersection theorems on polytopes, with the KKM theorem of Knaster, Kuratowski, and Mazurkiewicz [17] perhaps the most prominent example. It is well known that there is a close relationship between fixed point theorems and intersection theorems. Yet another alternative consists of results that claim the existence of solutions to variational inequality problems, the existence of stationary points, or the existence of zero points.

For certain models, it is not only important to know that there exists at least one solution, but one would like to show the existence of a continuum of solutions. In economics the existence of a continuum of solutions leads to difficulties in expectation formation of agents, and as a consequence provides scope for endogenously generated fluctuations. A particular example comes from general equilibrium theory with price rigidities, where a continuum of solutions on the unit cube as a polytope is shown to

*Received by the editors August 17, 1999; accepted for publication (in revised form) October 18, 2000; published electronically April 18, 2001. This research is part of the Research Program "Competition and Cooperation" of Tilburg University, The Netherlands.

http://www.siam.org/journals/sicon/39-6/36059.html

†Department of Economics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands (P.Herings@algec.unimaas.nl; www.2.unimaas.nl/fdewbae/herings.htm). The research of this author was supported by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

‡Department of Econometrics and CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (Talman@kub.nl).

§Faculty of Business Administration, Yokohama National University, Yokohama 240-8501, Japan (zyang@business.ynu.ac.jp). The research of this author was supported by Grants-in-Aid of the Ministry of Education, Science, Sports and Culture of Japan, and by a grant of The Netherlands Organization for Scientific Research (NWO).

exist in Herings [13]. It is therefore important to have generally applicable tools that guarantee the existence of a continuum of zero points to a certain system of equations.

This leads us to the following problem: Given a point-to-set mapping $\varphi : P \Longrightarrow \mathbb{R}^n$, with P an arbitrary polytope, *what reasonable conditions can guarantee the existence of a continuum of solutions x to the system*

$$0^n \in \varphi(x),$$

where 0^n denotes the n-dimensional vector of zeros? Our approach to show the existence of a continuum of solutions is to show that there is a connected subset of solutions that links together at least two distinct points, thereby guaranteeing the continuum. In this paper we will show that any upper semicontinuous point-to-set mapping with some mild (boundary) conditions will have a connected set of zero points linking together two distinct faces of the polytope P.

It is well known that under certain conditions a point-to-set mapping defined on a polytope has a solution to the variational inequality problem. We generalize the variational inequality problem and define a parametric variational inequality problem. In this paper we show that under similar conditions a point-to-set mapping defined on a polytope P has a connected set of solutions to the parametric variational inequality problem, called parametrized stationary points. The set of parametrized stationary points connects two distinct faces of P. With respect to some given nonzero vector c, on one of these faces, denoted by F^-, the value $c^\top x$ is minimized for $x \in P$, while on the other face, denoted by F^+, the value $c^\top x$ is maximized for $x \in P$. A special case occurs when both F^- and F^+ are vertices of P and the set of parametrized stationary points contains both these vertices. Under the three different conditions the set of parametrized stationary points is a connected set of zero points linking the two distinct faces F^- and F^+ of P.

We prove the existence results by designing a simplicial variable dimension algorithm on a polytope. This type of algorithm was initiated by Scarf [22]. Simplicial homotopy methods were developed by Eaves [8]. The simplicial restart variable dimension algorithm was introduced by van der Laan and Talman [18] to compute a fixed point of a continuous function from the unit simplex into itself. Such an algorithm generates a unique sequence of simplices of varying dimension in a simplicial subdivision of the set and connects the arbitrarily chosen starting point with an approximate solution. For other recent developments, we refer to Talman and Yamamoto [24], Yamamoto [26], Brown, DeMarzo, and Eaves [5], DeMarzo and Eaves [6], Yang [27, 28], and van der Laan, Talman, and Yang [19]. Allgower and Georg [1], Todd [25], and Yang [28] provide comprehensive treatments of simplicial algorithms.

In this paper a simplicial algorithm is proposed which generates within a simplicial subdivision of P a finite sequence of simplices of varying dimension. This sequence connects two different simplices, one simplex lying in the face F^- of P and the other lying in the face F^+ of P. The sequence of simplices connecting these two simplices is generated by the algorithm through a sequence of semilexicographic pivot steps in a linear system of equations. In case the face F^- is not a vertex of P, the algorithm starts by finding a suitable simplex in F^-. Next it generates a sequence of simplices of varying dimension in P. It is possible that the algorithm returns to F^-. Then it generates a third simplex in F^- from where a sequence of adjacent simplices in P is generated. It is shown that the algorithm eventually finds a simplex in F^- from which a sequence of adjacent simplices reaching F^+ is generated.

Induced by the sequence of adjacent simplices, the algorithm yields a piecewise linear path of parametrized stationary points of a piecewise linear approximation of the underlying mapping. When the mesh size of the simplicial subdivision of P goes to zero the sequence (or at least a subsequence) of piecewise linear paths converges to a connected set of parametrized stationary points of the original mapping. This set has points in common with both faces F^- and F^+.

The results in the paper generalize earlier results of Browder [4], Mas-Colell [20], and Herings, Talman, and Yang [15]. In the case of Browder's theorem the polytope is the Cartesian product of a polytope of one dimension less and the unit interval $[0,1]$, while c is the unit vector with the one on the last position. Mas-Colell's result is an extension of Browder's result to deal with correspondences. Both Browder and Mas-Colell proved their results via a rather sophisticated machinery. Since our approach here is constructive, we therefore also obtain an alternative but constructive proof for their results. In Herings, Talman, and Yang [15] the polytope equals the unit cube and the vector c is the vector of ones. In both Browder's and Mas-Colell's theorems, a connected set of fixed points is obtained connecting the levels 0 and 1, whereas the result on the unit cube yields a connected set of zero points connecting the vector of zeros and the vector of ones.

Intersection theorems with a continuum of intersection points can be found in Freidenfelds [9] and Herings and Talman [14]. Although Freidenfelds's intersection theorem typically has a continuum of intersection points, this is not necessarily the case. Freidenfelds's result generalizes the intersection theorem of Scarf [22]. Herings and Talman's results generalize a number of intersection theorems on the unit simplex, including the Scarf-result and the KKM-result, to intersection theorems on the unit cube and show the existence of a continuum of intersection points. The reader should be aware that compared with a large amount of existence results for a single fixed or zero point, existence results for a continuum of fixed or zero points are very rare.

This paper is organized as follows. In section 2 we state the problem and in section 3 we give three sufficient conditions for the existence of a connected set of zero points of an upper semicontinuous point-to-set mapping over an arbitrary polytope which link together two distinct faces of the polytope. An example is also given. In section 4 we introduce the algorithm, prove its convergence, and illustrate the algorithm by an example. In section 5 we analyze the accuracy of the approximation of zero points and prove the existence theorems. In section 6 we discuss a more general case. In section 7 we derive as special cases Browder's and Mas-Colell's theorems and an earlier result of the authors on the unit cube, and we also give an interesting economic application.

2. The problem.

Let I_m denote the set of the first m positive integers. Consider an arbitrary full-dimensional polytope P that has the following representation as a polyhedron:

$$P = \{x \in \mathbb{R}^n \mid {a^i}^\top x \leq b_i \, \forall i \in I_m\}.$$

For each subset I of I_m, define

$$F(I) = \{x \in P \mid {a^i}^\top x = b_i \, \forall i \in I\}.$$

Then $F(I)$ is called a *face* of P unless it is empty. Note that $F(\emptyset) = P$. Let

$$\mathcal{I} = \{I \subset I_m \mid F(I) \text{ is a face of } P\}.$$

A face F of the set $F(I)$ of dimension one lower than the dimension of $F(I)$ is called a *facet* of $F(I)$. The polytope P is said to be *simple* if the dimension of any face $F(I)$ of P is equal to $n - |I|$. Throughout the paper, whenever we use a polytope

P, it is assumed that P is simple and that its representation as a polyhedron has no redundant constraints.

We have the following observations.

(i) For each face F of P, the set $I \in \mathcal{I}$ with $F = F(I)$ is unique and is given by the set $\{i \in I_m \mid a^{i\top} x = b_i \ \forall x \in F\}$.

(ii) The set $F(I)$ is a vertex of P if and only if $I \in \mathcal{I}$ with $|I| = n$.

(iii) If $I \in \mathcal{I}$, then $I \setminus \{i\} \in \mathcal{I}$ for any $i \in I$.

(iv) For some $I \in \mathcal{I}$, F is a facet of $F(I)$ if and only if $F = F(I \cup \{i\})$ for some $i \notin I$ with $I \cup \{i\} \in \mathcal{I}$.

(v) For any $I \in \mathcal{I}$, the vectors a^i with $i \in I$ are linearly independent.

Let c be an arbitrary nonzero vector in $\mathbb{R}^n$. Then F^+ will denote the face of P such that for each $x^+ \in F^+$ it holds that $c^\top x^+ = \max_{x \in P} c^\top x$, and F^- will denote the face of P such that for each $x^- \in F^-$ it holds that $c^\top x^- = \min_{x \in P} c^\top x$. Let $t^+ = c^\top x^+$ for $x^+ \in F^+$ and $t^- = c^\top x^-$ for $x^- \in F^-$. Since P is full-dimensional, it follows that $t^- < t^+$ and therefore $F^- \cap F^+ = \emptyset$. We define

$$I^+ = \{i \in I_m \mid a^{i\top} x = b_i \ \forall x \in F^+\},$$
$$I^- = \{i \in I_m \mid a^{i\top} x = b_i \ \forall x \in F^-\}.$$

So $F^+ = F(I^+)$ and $F^- = F(I^-)$.

We need some further notation. For each $I \in \mathcal{I}$, we define

$$A(I) = \left\{ y \in \mathbb{R}^n \mid y = \sum_{i \in I} \mu_i a^i + \beta c,\ \mu_i \geq 0 \ \forall i \in I, \text{ and } \beta \in \mathbb{R} \right\},$$

$$A_0(I) = \left\{ y \in \mathbb{R}^n \mid y = \sum_{i \in I} \mu_i a^i,\ \mu_i \geq 0 \ \forall i \in I \right\},$$

$$A^*(I) = \{x \in \mathbb{R}^n \mid x^\top y \leq 0 \ \forall y \in A(I)\},$$
$$A_0^*(I) = \{x \in \mathbb{R}^n \mid x^\top y \leq 0 \ \forall y \in A_0(I)\}.$$

Note that $A(\emptyset) = \{y \in \mathbb{R}^n \mid y = \beta c,\ \beta \in \mathbb{R}\}$, $A_0(\emptyset) = \{0^n\}$, $A^*(\emptyset) = \{x \in \mathbb{R}^n \mid x^\top c = 0\}$, and $A_0^*(\emptyset) = \mathbb{R}^n$. Moreover, for any $I \in \mathcal{I}$ we have that $A_0(I) \subset A(I)$, $A^*(I) \subset A_0^*(I)$, $A^*(I) \cap A(I) = \{0^n\}$, and $A_0^*(I) \cap A_0(I) = \{0^n\}$. We may interpret $A(I)$ and $A_0(I)$ as normal cones to the boundary of P and $A^*(I)$ and $A_0^*(I)$ as tangent cones. The pairs $(A(I), A^*(I))$ and $(A_0(I), A_0^*(I))$ may be viewed as primal and dual pairs as well; see, e.g., Aubin [2].

Let $\varphi : P \Longrightarrow \mathbb{R}^n$ be a correspondence that satisfies the following assumption.

ASSUMPTION 2.1. *The correspondence $\varphi : P \Longrightarrow \mathbb{R}^n$ is nonempty valued, compact valued, convex valued, and upper semicontinuous.*

For an arbitrary function $f : P \to \mathbb{R}^n$, the *stationary point* (or *variational inequality*) *problem* for f on the polytope P is to find a point $x^* \in P$ such that

$$(x^* - x)^\top f(x^*) \geq 0 \ \forall x \in P.$$

Such a point x^* is called a *stationary point* of f on P. It is well known that a continuous function on a convex compact set has a stationary point; see Hartman and Stampacchia [11] and Eaves [8].

In the following we give the definition of a stationary point of a correspondence on the polytope P with respect to the vector c.

Definition 2.2. *A point $x^* \in P$ is a stationary point of a correspondence φ on the polytope P with respect to the nonzero vector c if there exists $f \in \varphi(x^*)$ such that $(x^* - x)^\top f \geq 0 \forall x \in P$ satisfying $c^\top x = c^\top x^*$.*

We call the problem of finding a stationary point with respect to a nonzero vector c a *parametric variational inequality problem.* A solution to it is called a *parametrized stationary point.* Now we have the following simple but important observation.

Lemma 2.3. *A point $x^* \in P$ is a parametrized stationary point of the mapping φ on P with respect to c if and only if for some $I \in \mathcal{I}$ it holds that $x^* \in F(I)$ and $\varphi(x^*) \cap A(I) \neq \emptyset$.*

Proof. Let x^* be a parametrized stationary point of φ on P with respect to c. Hence there exists $f \in \varphi(x^*)$ such that $(x^* - x)^\top f \geq 0 \forall x \in P$ satisfying $c^\top x = c^\top x^*$; i.e., x^* maximizes $x^\top f$ subject to ${a^i}^\top x \leq b_i, i \in I_m$, and $c^\top x = c^\top x^*$. Therefore, there exist $\mu_i \geq 0, i \in I_m$, and $\beta \in \mathbb{R}$ satisfying

$$f = \sum_{i \in I_m} \mu_i a^i + \beta c,$$

and $\mu_i = 0$ if ${a^i}^\top x^* < b_i$. Let $I = \{i \in I_m \mid {a^i}^\top x^* = b_i\}$. Then $x^* \in F(I)$ and $f \in A(I) \cap \varphi(x^*)$.

Next let $x^* \in P$ be such that $x^* \in F(I)$ and $A(I) \cap \varphi(x^*) \neq \emptyset$ for some $I \in \mathcal{I}$. Take any $f \in A(I) \cap \varphi(x^*)$. Then $f \in \varphi(x^*)$ and there exists $\mu_i \geq 0, i \in I$, and $\beta \in \mathbb{R}$ satisfying

$$f = \sum_{i \in I} \mu_i a^i + \beta c.$$

Note that ${a^i}^\top x^* = b_i \forall i \in I$, and for any $x \in P$, ${a^i}^\top x \leq b_i \forall i \in I$. Thus ${x^*}^\top f = \sum_{i \in I} \mu_i b_i + \beta c^\top x^* \geq x^\top f$ for all $x \in P$ with $x^\top c = {x^*}^\top c$. By definition, x^* is a stationary point of φ on P with respect to c. □

We show the existence of a continuum of parametrized stationary points of the correspondence φ with respect to the vector c. A point $x^* \in P$ is called a *zero point* of the mapping φ if $0^n \in \varphi(x^*)$. Any zero point of φ is a parametrized stationary point. The main purpose of this paper is to find conditions on φ which enable us to guarantee the existence and the computation of a continuum of zero points of a correspondence φ on P such that it contains points of both F^- and F^+.

3. Existence conditions.

The first result is on the existence of parametrized stationary points.

Theorem 3.1. *Let $\varphi : P \Longrightarrow \mathbb{R}^n$ be any correspondence satisfying Assumption* 2.1 *and let $c \in \mathbb{R}^n \setminus \{0^n\}$ be given. Then there exists a connected set C of parametrized stationary points of φ on P with respect to c such that $C \cap F^- \neq \emptyset$ and $C \cap F^+ \neq \emptyset$.*

The theorem makes clear that there are many parametrized stationary points. To be more precise, there exists a continuum of them with a special topological structure. The set of parametrized stationary points has a connected subset that links the two faces F^- and F^+.

In order to obtain the existence of a continuum of zero points, we need to impose certain conditions on the correspondence φ. The following three results list sufficient conditions for the existence of a continuum of zero points of φ.

Theorem 3.2. *Let $\varphi : P \Longrightarrow \mathbb{R}^n$ be any correspondence satisfying Assumption* 2.1 *and let $c \in \mathbb{R}^n \setminus \{0^n\}$. If for any $x \in F(I), I \in \mathcal{I}$, it holds that $\varphi(x) \cap A(I) \subset \{0^n\}$, then there exists a connected set C of zero points of φ such that $C \cap F^- \neq \emptyset$ and $C \cap F^+ \neq \emptyset$.*

The condition in the theorem states that for any x in the face $F(I)$ of P the image $\varphi(x)$ may not have nonzero elements in common with $A(I)$.

THEOREM 3.3. *Let $\varphi : P \Longrightarrow \mathbb{R}^n$ be any correspondence satisfying Assumption* 2.1 *and let $c \in \mathbb{R}^n \setminus \{0^n\}$. If for any $x \in F(I), I \in \mathcal{I}$, it holds that*

(i) $\varphi(x) \cap A_0^*(I) \cap A(I) \subset \{0^n\}$,

(ii) $\varphi(x) \cap A_0^*(I) \neq \emptyset$,

then there exists a connected set C of zero points of φ such that $C \cap F^- \neq \emptyset$ and $C \cap F^+ \neq \emptyset$.

The conditions in the theorem imply that for any x in the face $F(I)$ of P the image $\varphi(x)$ may not have nonzero elements in common with $A(I) \cap A_0^*(I)$ and that at least one element of $\varphi(x)$ lies in $A_0^*(I)$.

THEOREM 3.4. *Let $\varphi : P \Longrightarrow \mathbb{R}^n$ be any correspondence satisfying Assumption* 2.1*, and let $c \in \mathbb{R}^n \setminus \{0^n\}$. If for any $x \in F(I), I \in \mathcal{I}$, it holds that*

$$\varphi(x) \cap A^*(I) \neq \emptyset,$$

then there exists a connected set C of zero points of φ such that $C \cap F^- \neq \emptyset$ and $C \cap F^+ \neq \emptyset$.

The condition in the theorem says that for any x in the face $F(I)$ of P at least one element of $\varphi(x)$ lies in $A^*(I)$.

The three theorems state different conditions for which a continuum of zero points can be shown to exist. Moreover, there is a logical order in these sufficient conditions. Theorem 3.2 states a weak condition, but one that holds for all elements of $\varphi(x)$. Theorem 3.4 gives a strong condition but requires this condition to hold for only one element in $\varphi(x)$. Theorem 3.3 is in between: it gives a very weak condition for all elements in $\varphi(x)$ together with a rather weak condition for some element in $\varphi(x)$.

The following result claims that for the special case where φ is a function f, Theorem 3.2 is the strongest and Theorem 3.4 the weakest. Note that Assumption 2.1 implies that the function f is continuous.

THEOREM 3.5. *Let P be any polytope and let $c \in \mathbb{R}^n \setminus \{0^n\}$. The collection of functions satisfying the conditions of Theorem* 3.2 *contains the collection of functions satisfying the conditions of Theorem* 3.3*, which contains the collection of functions satisfying the conditions of Theorem* 3.4.

Proof. Suppose that a function f from P to $\mathbb{R}^n$ satisfies the conditions of Theorem 3.4. Take any x in $F(I)$, so $f(x) \in A^*(I)$. Since $A^*(I) \subset A_0^*(I)$ and $A(I) \cap A^*(I) = \{0^n\}$ we obtain that $f(x) \in A_0^*(I)$ and $f(x)$ not in $A(I)$ unless $f(x) = 0^n$. Hence the conditions of Theorem 3.3 are satisfied. Suppose now that a function satisfies the conditions of Theorem 3.3. Again take any x in $F(I)$, so $f(x) \in A_0^*(I)$ and $f(x)$ not in $A(I) \cap A_0^*(I)$ unless $f(x) = 0^n$. Hence $f(x)$ not in $A(I)$ unless $f(x) = 0^n$. □

That Theorems 3.2, 3.3, and 3.4 are mutually exclusive for correspondences follows from the fact that in case of correspondences the image $\varphi(x)$ of any point x in P might consist of more than one element. For example, for a point x in the face $F(I)$ of P it is required in Theorem 3.2 that no nonzero element of $\varphi(x)$ lies in $A(I)$ which does not imply that at least one such element should lie in $A^*(I)$ as required in the conditions of Theorem 3.4. On the other hand if for an x in $F(I)$ it holds that some nonzero element $f \in \varphi(x)$ lies in $A^*(I)$, as in the conditions of Theorem 3.4, then this implies that f indeed does not lie in $A(I)$ but not necessarily that all the other elements of $\varphi(x)$ also do not lie in $A(I)$ as required in Theorem 3.2. Similar remarks can be made when comparing the conditions of Theorem 3.3 with the conditions in the other two theorems.

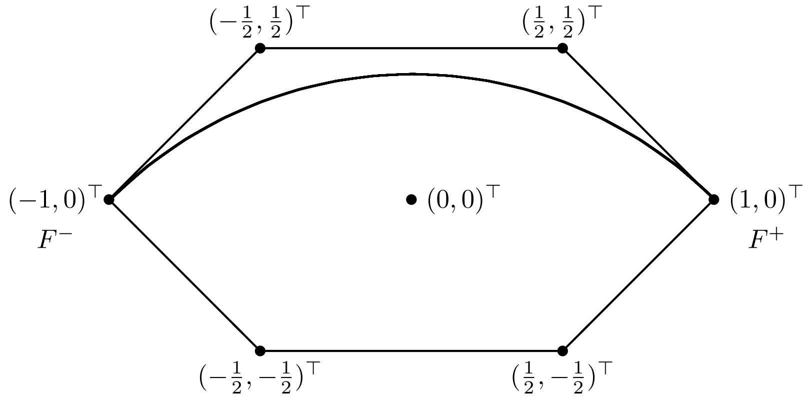

$(-\frac{1}{2},\frac{1}{2})^\top$
$(\frac{1}{2},\frac{1}{2})^\top$
$(-1,0)^\top$
F^-
$(0,0)^\top$
$(1,0)^\top$
F^+
$(-\frac{1}{2},-\frac{1}{2})^\top$
$(\frac{1}{2},-\frac{1}{2})^\top$

FIG. 3.1. *The set of zero points of f on P in Example* 3.1 *linking F^- and F^+.*

The theorems above will be proved in a constructive manner in the following sections. Their applications will be discussed in section 7. To conclude this section we will give a numerical example to illustrate the existence of a connected set of zero points on a two-dimensional diamond and then we will illustrate the algorithm with this example in the next section.

Example 3.1. Let the polytope P, illustrated in Figure 3.1, be given by

$$P = \{x \in \mathbb{R}^2 \mid |x_2| \leq 1/2,\ |x_1| + |x_2| \leq 1\}.$$

The vector c and the function f on P are given by $c = (1,0)^\top$ and

$$f(x) = \left(2 - x_1^2 - (x_2+1)^2, \sqrt{2 - x_1^2} - x_2 - 1\right)^\top.$$

Clearly, F^- and F^+ correspond to $(-1,0)^\top$, $(1,0)^\top$, vertices of P, respectively. Furthermore, if x lies on the boundary of P, then $f_1(x) > 0$ and $f_2(x) > 0$ when $x_2 < 0$, and $f_1(x) < 0$ and $f_2(x) < 0$ when $x_2 > 0$. Also, $f_1(x) = 0$ whenever $f_2(x) = 0$. Note that P can be rewritten as

$$\begin{aligned} P = \{x \in \mathbb{R}^2 \quad | \quad & x_1 + x_2 \leq 1,\ x_2 \leq 1/2,\ -x_1 + x_2 \leq 1, \\ & -x_1 - x_2 \leq 1,\ -x_2 \leq 1/2,\ x_1 - x_2 \leq 1\}. \end{aligned}$$

One can easily verify that f is continuous and satisfies the condition of Theorem 3.2. So it follows that there exists a connected set of zero points of f on P linking both $(-1,0)^\top$ and $(1,0)^\top$. In fact this set equals the circle segment $\{x \in P \mid x_1^2+(x_2+1)^2 = 2\}$; see Figure 3.1.

4. The algorithm and its convergence proof. In this section we describe an algorithm on the polytope P that generates a piecewise linear path of parametrized stationary points, with respect to the given nonzero vector c, of a piecewise linear approximation of the mapping φ. The piecewise linear approximation is taken with respect to some simplicial subdivision of the set P. In this and the following sections we assume that the faces F^- and F^+ are vertices of P, denoted by x^- and x^+, respectively. In that case the piecewise linear path generated by the algorithm contains both x^- and x^+ and can be traced by a sequence of semilexicographic pivot steps in a system of linear equations. The general case is discussed in section 6.

For a nonnegative integer t, a t-dimensional simplex or t-simplex, denoted by σ, is defined by the convex hull of $t+1$ affinely independent points $x^1, \ldots, x^{t+1}$ in $\mathbb{R}^n$. We often write $\sigma = \sigma(x^1, \ldots, x^{t+1})$ and call $x^1, \ldots, x^{t+1}$ the vertices of σ. A $(t-1)$-simplex being the convex hull of t vertices of σ is said to be a facet of σ. The facet $\tau(x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^{t+1})$ is called the facet of $\sigma(x^1, \ldots, x^{t+1})$ opposite to the vertex x^i. For k, $0 \le k \le t$, a k-simplex being the convex hull of $k+1$ vertices of σ is said to be a k-face or face of σ. A finite collection $\mathcal{T}$ of n-simplices is a triangulation of the polytope P if

(i) P is the union of all simplices in $\mathcal{T}$;

(ii) the intersection of any two simplices of $\mathcal{T}$ is either the empty set or a common face of both.

Let $\mathcal{T}$ be any triangulation of P. Then every face $F(I)$ of P is subdivided into t-simplices, where $t = n - |I|$. For example we can take the V-triangulation of Talman and Yamamoto [24]. Since $\mathcal{T}$ is finite and P is compact, every facet τ of an $(n-|I|)$-simplex σ in $F(I)$ either lies in the boundary of $F(I)$ and is only a facet of σ or is a facet of exactly one other $(n-|I|)$-simplex in $F(I)$. Let f be a simplicial approximation of φ with respect to $\mathcal{T}$. This means that $f(x) \in \varphi(x)$ for each vertex of $\mathcal{T}$ and f is affine on each simplex of $\mathcal{T}$.

A row vector is *lexicopositive* if it is a nonzero vector and its first nonzero entry is positive. A matrix is said to be *lexicopositive* if all its rows are lexicopositive. A matrix is said to be *semilexicopositive* if each row except possibly the last row is lexicopositive.

Definition 4.1. *Let $\tau(x^1, \ldots, x^t)$ be a $(t-1)$-simplex in $F(I)$, where $I \in \mathcal{I}$ with $I = \{i_{t+1}, \ldots, i_n\}$, $t = n - |I|$. The $(n+1) \times (n+1)$ matrix*

$$A_{\tau,I} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ -f(x^1) & \cdots & -f(x^t) & a^{i_{t+1}} & \cdots & a^{i_n} & c \end{bmatrix}$$

is the label matrix of τ with respect to I. The simplex τ is I-complete if $A_{\tau,I}^{-1}$ exists and is semilexicopositive; i.e., the first nonzero entry in every row, except possibly the last one, is positive.

Notice that if for an I-complete simplex τ we change the ordering of the first n columns of the matrix $A_{\tau,I}$, the inverse of the resulting matrix still exists and is semilexicopositive. Clearly, if, for some $I \in \mathcal{I}$, a $(t-1)$-simplex $\tau(x^1, \ldots, x^t)$ is an I-complete facet of a simplex $\sigma(x^1, \ldots, x^{t+1})$ in some face $F(I)$, then the system of $n+1$ linear equations with $n+2$ variables

$$\sum_{j=1}^{t+1} \lambda_j \begin{pmatrix} 1 \\ -f(x^j) \end{pmatrix} + \sum_{i \in I} \mu_i \begin{pmatrix} 0 \\ a^i \end{pmatrix} + \beta \begin{pmatrix} 0 \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0^n \end{pmatrix} \tag{$*$}$$

has a solution $(\lambda, \mu, \beta) = (\lambda_1, \ldots, \lambda_{t+1}, (\mu_i)_{i \in I}, \beta)$ satisfying $\lambda_j \ge 0$ for $j \in I_{t+1}$ and $\mu_i \ge 0$ for $i \in I$, with $\lambda_{t+1} = 0$. Let x be defined by $x = \sum_{j=1}^{t+1} \lambda_j x^j$ at a solution (λ, μ, β) of $(*)$; then x lies in σ and is a parametrized stationary point of f with respect to c.

The following result is a special case of Theorem 2.6 in Fujishige and Yang [10] and will be used later. This result has been proved in a constructive way.

Theorem 4.2. *Consider any polytope* Q *given by* $\mathrm{Q} = \{x \in \mathbb{R}^n \mid c^{i\top} x \le d_i,\ i \in I_n$ *and* $c^{0\top} x = d_0\}$. *Assume that* Q *is an $(n-1)$-dimensional simple polytope with no redundant constraints. For any $g \in \mathbb{R}^n$, there exists a unique subset $I = \{j_1, \ldots, j_{n-1}\}$*

of I_n with $|I| = n-1$ such that the matrix

$$B^{-1} = \left[\begin{array}{cccccc} 1 & 0 & 0 & \cdots & 0 & 0 \\ g & c^{j_1} & c^{j_2} & \cdots & c^{j_{n-1}} & c^0 \end{array} \right]^{-1}$$

exists and is semilexicopositive.

We now show that each of the two 0-simplices $\{x^-\}$ and $\{x^+\}$ is I-complete in the face $F(I)$ for a unique index set $I \in \mathcal{I}$ containing $n-1$ indices.

Lemma 4.3. *Let $x^1 = x^-$ and $\tau = \{x^1\}$. Then there exists a unique subset $I = \{j_1, \ldots, j_{n-1}\}$ of I^- with $|I| = n-1$ such that τ is an I-complete 0-simplex in $F(I)$.*

Proof. Recall that x^- is a unique solution to the problem

$$\min c^\top x \text{ s.t. } x \in P.$$

By duality theory there exists a unique solution $\lambda_i > 0 \forall i \in I^-$ such that $-c = \sum_{i \in I^-} \lambda_i a^i$. In other words, the vectors c and a^i, $i \in I^-$, are affinely independent. Consider the following polyhedron:

$$W = \{x \in \mathbb{R}^n \mid {a^i}^\top x \leq 1, i \in I^-, c^\top x \leq 1\}.$$

It is easy to see that W is bounded and contains 0^n in its interior and therefore is an n-dimensional polytope. Then the set

$$Q = \{x \in \mathbb{R}^n \mid {a^i}^\top x \leq 1, i \in I^-, c^\top x = 1\}$$

is an $(n-1)$-dimensional polytope. Let $g = -f(x^-)$. Now all the conditions of Theorem 4.2 are satisfied. So there exists a unique subset $I = \{j_1, \ldots, j_{n-1}\}$ of I^- with $|I| = n-1$ such that the matrix

$$B^{-1} = \left[\begin{array}{cccccc} 1 & 0 & 0 & \cdots & 0 & 0 \\ -f(x^-) & a^{j_1} & a^{j_2} & \cdots & a^{j_{n-1}} & c \end{array} \right]^{-1}$$

exists and is semilexicopositive. This means that τ is I-complete. Clearly, τ lies in $F(I)$ since $F(I^-)$ is a subset of $F(I)$ and $\tau = F(I^-)$. □

Lemma 4.4. *Let $x^1 = x^+$ and $\tau = \{x^1\}$. Then there exists a unique subset $I = \{j_1, \ldots, j_{n-1}\}$ of I^+ with $|I| = n-1$ such that τ is an I-complete 0-simplex in $F(I)$.*

Proof. Notice that x^+ is the unique solution to the problem

$$\max c^\top x \text{ s.t. } x \in P.$$

By duality theory there exists a unique solution $\lambda_i > 0 \forall i \in I^+$ such that $c = \sum_{i \in I^+} \lambda_i a^i$. In other words, the vectors $-c$ and a^i, $i \in I^+$, are affinely independent. Consider the following polyhedron:

$$W = \{x \in \mathbb{R}^n \mid {a^i}^\top x \leq 1, i \in I^+, -c^\top x \leq 1\}.$$

It is easy to see that W is an n-dimensional polytope. Then the set

$$Q = \{x \in \mathbb{R}^n \mid {a^i}^\top x \leq 1, i \in I^+, c^\top x = -1\}$$

is an $(n-1)$-dimensional polytope. Let $g = -f(x^+)$. Now all the conditions of Theorem 4.2 are satisfied. So there exists a unique subset $I = \{j_1, \ldots, j_{n-1}\}$ of I^+ with $|I| = n-1$ such that the matrix

$$B^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -f(x^+) & a^{j_1} & a^{j_2} & \cdots & a^{j_{n-1}} & c \end{bmatrix}^{-1}$$

exists and is semilexicopositive. This means that τ is I-complete. Clearly, τ lies in $F(I)$ since $F(I^+)$ is a subset of $F(I)$ and τ lies in $F(I^+)$. □

The following lemma is well known in linear programming theory and can easily be proved. We will invoke it later. Let B be a matrix. We denote its ith row by $B_{i.}$ and its jth column by $B_{.j}$.

Lemma 4.5. *Let $B = (B_{.1}, \ldots, B_{.n+1})$ be any nonsingular $(n+1)\times(n+1)$ matrix and let x be any vector in $\mathbb{R}^{n+1}$. Let $k \in I_{n+1}$ and*

$$\bar{B} = (B_{.1}, \ldots, B_{.k-1}, x, B_{.k+1}, \ldots, B_{.n+1}).$$

Then either $(B^{-1}x)_k = 0$ and $\bar{B}$ is singular, or $(B^{-1}x)_k \neq 0$, $\bar{B}$ is nonsingular, and $\bar{B}^{-1}$ is given by

$$\bar{B}^{-1} = \begin{bmatrix} (B^{-1})_{1.} - \frac{(B^{-1}x)_1}{(B^{-1}x)_k}(B^{-1})_{k.} \\ \vdots \\ (B^{-1})_{k-1.} - \frac{(B^{-1}x)_{k-1}}{(B^{-1}x)_k}(B^{-1})_{k.} \\ \frac{1}{(B^{-1}x)_k}(B^{-1})_{k.} \\ (B^{-1})_{k+1.} - \frac{(B^{-1}x)_{k+1}}{(B^{-1}x)_k}(B^{-1})_{k.} \\ \vdots \\ (B^{-1})_{n+1.} - \frac{(B^{-1}x)_{n+1}}{(B^{-1}x)_k}(B^{-1})_{k.} \end{bmatrix}.$$

Lemma 4.6. *Let σ be a t-simplex in $F(I)$, where $I \in \mathcal{I}$, $t = n - |I|$, and $I = \{i_{t+1}, \ldots, i_n\}$. If σ has an I-complete facet τ, then exactly one of the following two cases occurs:*

1. *The simplex σ is an $\bar{I}$-complete simplex in $F(\bar{I})$, where $\bar{I} = I\backslash\{i\}$ for precisely one index $i \in I$.*
2. *The simplex σ has exactly one other I-complete facet $\bar{\tau}$.*

Proof. Let x^{t+1} be the vertex of σ opposite to τ, and let $y = A_{\tau,I}^{-1}(1, -f(x^{t+1})^\top)^\top$. Notice that $y \neq 0^{n+1}$. Let $K = \{i \in I_n \mid y_i > 0\}$. We first prove $|K| > 0$. Since $A_{\tau,I}y = (1, -f(x^{t+1})^\top)^\top$, we have $\sum_{i=1}^t y_i = 1$. This implies that there exists at least one index $i \in I_t$ such that $y_i > 0$. Hence K is nonempty.

Consider the ratio vectors $(1/y_j)(A_{\tau,I}^{-1})_{j.} \forall j \in K$. Choose $k \in K$ such that the kth ratio vector is the minimum in the lexicographic order over all such ratio vectors. Since $A_{\tau,I}^{-1}$ is regular, k is uniquely determined. Now, we consider the following two cases.

(1) If $k \in I_n \backslash I_t$, then let $l = i_k$ and $\bar{I} = I\backslash\{l\}$. Clearly, $\bar{I} \in \mathcal{I}$ and σ is in $F(\bar{I})$. Let B be the matrix obtained from $A_{\tau,I}$ by replacing its kth column by $(1, -f(x^{t+1})^\top)^\top$. It follows from Lemma 4.5 that B^{-1} exists and is semilexicopositive. By reordering the columns of B we get $A_{\sigma,\bar{I}}$ whose inverse exists and is semilexicopositive. So σ is $\bar{I}$-complete.

(2) If $k \in I_t$, then let $\bar{\tau}$ be the facet of σ opposite to the vertex x^k. Using Lemma 4.5, it follows from the choice of k that $A^{-1}_{\bar{\tau},I}$ exists and is semilexicopositive. Hence $\bar{\tau}$ is an I-complete $(t-1)$-simplex in $F(I)$.

It follows immediately from Lemma 4.5 that if any column other than the kth column is replaced, then the inverse of the resulting matrix is not semilexicopositive. □

In Lemma 4.8 we consider the analogous case of making a lexicographic pivot step with a column $(1, a^{i\top})^\top$. First we need the next lemma.

LEMMA 4.7. *For any set $I \in \mathcal{I}$ with $I \neq I^-$ and $I \neq I^+$, there exist no solutions λ_0, λ_i, $i \in I$, to $\sum_{i\in I} \lambda_i a^i = \lambda_0 c$ such that $\lambda_i \leq 0 \forall i \in I$ and $\sum_{i\in I} \lambda_i < 0$.*

Proof. We need to consider the following three cases:

Case (1). If $\lambda_0 = 0$, then $\sum_{i\in I} \lambda_i a^i = \lambda_0 c = 0^n$ contradicts the fact that all vectors a^i, $i \in I$, are linearly independent.

Case (2). If $\lambda_0 < 0$, then by duality theory $\sum_{i\in I} \lambda_i a^i = \lambda_0 c$ and $I \neq I^+$ contradicts the fact that

$$c^\top x^+ = \max_{x\in P} c^\top x.$$

Case (3). If $\lambda_0 > 0$, then by duality theory $\sum_{i\in I} \lambda_i a^i = \lambda_0 c$ and $I \neq I^-$ contradicts the fact that

$$c^\top x^- = \min_{x\in P} c^\top x. \qquad \square$$

LEMMA 4.8. *Let σ be an I-complete $(t-1)$-simplex in $F(I)$, where $I \in \mathcal{I}$, $t = n - |I|$, and $I = \{i_{t+1}, \ldots, i_n\}$. If σ is in $F(\bar{I})$ and $\bar{I} \neq I^-$ or $\bar{I} \neq I^+$, where $\bar{I} = I \cup \{l\} \in \mathcal{I}$ for some $l \in I_m \setminus I$, then exactly one of the following two cases occurs:*

(1) *There exists a unique set $J \in \mathcal{I}$ with $|J| = |I|$ and $J \neq I$ so that σ is in $F(J)$ and is J-complete.*

(2) *There exists exactly one facet τ of σ which is in $F(\bar{I})$ and is $\bar{I}$-complete.*

Proof. Let $x = (0, a^{l\top})^\top$ and $y = A^{-1}_{\sigma,I} x$. Note that $y \neq 0^{n+1}$. Let $K = \{i \in I_n \mid y_i > 0\}$. Note that $A_{\sigma,I} y = (0, a^{l\top})^\top$. We need to consider the following two cases.

Case (i). If there exists an index $j \in I_t$ such that $y_j < 0$, then there must exist an index $i \in I_t$ such that $y_i > 0$ since $\sum_{k=1}^t y_i = 0$. Hence K is nonempty.

Case (ii). Suppose that $y_i = 0 \forall i \in I_t$. If $y_i \leq 0 \forall i = t+1, t+2, \ldots, n$, then we have that $a^l = \sum_{i=t+1}^n y_i a^{j_{i-t}} + y_{n+1} c$. By Lemma 4.7 it is impossible since $\bar{I}$ is neither equal to I^- nor equal to I^+. Hence there exists at least one index $i \in I_n \setminus I_t$ such that $y_i > 0$. Again K is nonempty.

Consider the ratio vectors $(1/y_j)(A^{-1}_{\sigma,I})_{j\cdot} \forall j \in K$. Choose $k \in K$ such that the kth ratio vector is the minimum in the lexicographic order over all such ratio vectors. Since $A^{-1}_{\tau,I}$ is regular, k is uniquely determined. Now, we consider the following two cases.

(1) If $k \in I_n \setminus I_t$, then let $p = i_k$ and $J = I \cup \{l\} \setminus \{p\}$. Clearly, $J \in \mathcal{I}$, $J \neq I$, $|J| = |I|$, and σ is in $F(J)$. Let B be the matrix obtained from $A_{\sigma,I}$ by replacing its kth column by x. It follows from Lemma 4.5 that B^{-1} exists and is semilexicopositive. It is clear that $A_{\sigma,J} = B$. Thus σ is a J-complete $(t-1)$-simplex in $F(J)$.

(2) If $k \in I_t$, then let τ be the facet of σ opposite to the vertex x^k. Clearly, τ is a $(t-2)$-simplex in $F(\bar{I})$. Let B be the matrix obtained from $A_{\sigma,I}$ by replacing its kth column by x. It follows from Lemma 4.5 that B^{-1} exists and is semilexicopositive. By reordering the columns of B we get $A_{\tau,\bar{I}}$, whose inverse also exists and is semilexicopositive. So τ is an $\bar{I}$-complete $(t-2)$-simplex in $F(\bar{I})$.

Again it follows from Lemma 4.5 that if any other column is replaced, then the new matrix is no longer semilexicopositive. ∎

We construct a graph $G = (\mathcal{V}, \mathcal{A})$, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{A}$ denotes the set of edges. Each I-complete $(n - |I| - 1)$-simplex is a *node* in $\mathcal{V}$. An I-complete $(n - |I| - 1)$-simplex τ^1 in $F(I)$ and a J-complete $(n - |J| - 1)$-simplex τ^2 in $F(J)$ are said to be *adjacent* complete simplices if $I = J = L$ and τ^1 and τ^2 are both facets of an $(n - |L|)$-simplex σ in $F(L)$, or τ^1 is a facet of τ^2 and τ^2 is an $(n - |I|)$-simplex in $F(I)$, or τ^2 is a facet of τ^1 and τ^1 is an $(n - |J|)$-simplex in $F(J)$. Two adjacent complete simplices τ^1 and τ^2 are connected by an edge $e = \{\tau^1, \tau^2\} \in \mathcal{A}$. The degree of a node τ in G is defined to be the number of nodes connected with it, denoted by $deg(\tau)$. A path in G from node $\tau^0 = \{x^-\}$ to node τ^l is defined as a sequence of the form $(\tau^0, e_1, \tau^1, \dots, e_l, \tau^l)$, where τ^0, τ^1,…, τ^l are nodes and e_1, …, e_l are edges such that $e_i = \{\tau^{i-1}, \tau^i\}$ for $i \in I_l$. A path is simple if all its nodes and edges are different.

Theorem 4.9. *Let $\mathcal{T}$ be a triangulation of P. Starting at the vertex x^-, the algorithm generates a finite sequence of adjacent J-complete simplices for varying $J \in \mathcal{I}$ which leads to the vertex x^+.*

Proof. By Lemma 4.6, $\{x^-\}$ is an I-complete 0-simplex in $F(I)$ for some unique set $I \in \mathcal{I}$ with $|I| = n - 1$. Since $\{x^-\}$ lies in the boundary of $F(I)$, there exists a unique 1-simplex σ in $F(I)$ having $\{x^-\}$ as its facet. By Lemma 4.6, either σ is an $\bar{I}$-complete simplex in $F(\bar{I})$, where $\bar{I} = I \setminus \{i\}$ for some unique $i \in I$, or σ has exactly one other I-complete facet $\bar{\tau}$. Hence there exists a unique adjacent complete simplex to $\{x^-\}$. That is, $deg(\{x^-\}) = 1$. Similarly, by using Lemmas 4.6 and 4.8, we can prove $deg(\{x^+\}) = 1$.

In all other cases, we prove that if τ is an I-complete $(n - |I| - 1)$-simplex in $F(I)$ for some $I \in \mathcal{I}$, τ has exactly two adjacent complete simplices. There are two possibilities: either τ lies in the interior of $F(I)$ or τ lies in the boundary of $F(I)$. If τ lies in the interior of $F(I)$, then τ is a facet of exactly two $(n - |I|)$-simplices in $F(I)$. It follows from Lemma 4.8 that τ is adjacent to exactly two complete simplices. If τ lies in the boundary of $F(I)$, then there exists exactly one $(n - |I|)$-simplex σ in $F(I)$ having τ as its facet. By Lemma 4.8 either σ is an $\bar{I}$-complete $(n - |\bar{I}| - 1)$-simplex in $F(\bar{I})$ for some unique $\bar{I} \in \mathcal{I}$ with $|\bar{I}| = |I| - 1$ and has no other I-complete facets, or σ has exactly one other I-complete facet. This yields one adjacent complete simplex to τ. On the other hand, since τ lies in the boundary of $F(I)$, τ lies in $F(\tilde{I})$ for some unique set $\tilde{I} \in \mathcal{I}$ with $|\tilde{I}| = |I| + 1$. By Lemma 4.8 either τ is J-complete for some unique set $J \in \mathcal{I}$ with $|J| = |I|$ and $J \neq I$, or τ has exactly one $\tilde{I}$-complete facet. In the former case, τ lies in $F(\tilde{I})$ and hence there exists exactly one simplex $\bar{\sigma}$ in $F(\tilde{I})$ having τ as its facet. It follows again from Lemma 4.8 that there exists exactly one other complete simplex adjacent to τ. This concludes that τ has exactly two adjacent complete simplices. In other words, we have $deg(\tau) = 2$.

As shown above, the degree of each node in the graph $G = (\mathcal{V}, \mathcal{A})$ is at most two. Exactly two nodes have degree equal to one. Since the number of simplices in P is finite, the number of nodes in G must be finite, too. Since $deg(\{x^-\}) = 1$, it is easy to see that there exists a simple finite path starting from $\{x^-\}$. The end node of this path must be a node τ of degree one and different from x^-. The only possibility is that τ is equal to $\{x^+\}$. ∎

In what follows, if a column of $A_{\tau,I}$ corresponds to a vertex x^i, we call it a vertex column; if it corresponds to a constraint vector a^j, we call it an index column. Now we summarize the steps of the algorithm.

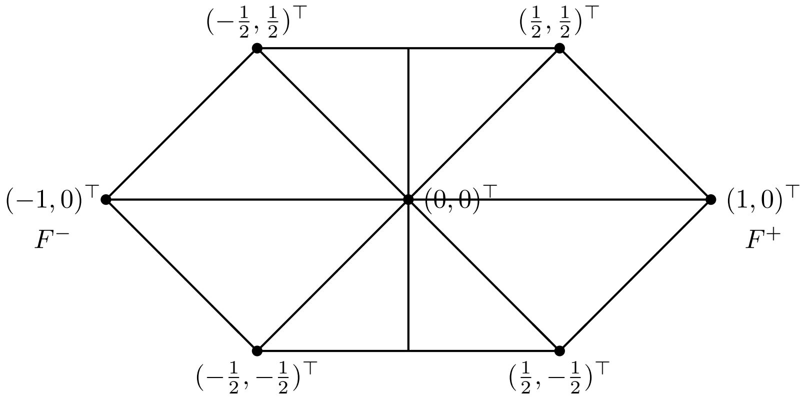

$(-\frac{1}{2},\frac{1}{2})^\top$
$(\frac{1}{2},\frac{1}{2})^\top$
$(-1,0)^\top$
F^-
$(0,0)^\top$
$(1,0)^\top$
F^+
$(-\frac{1}{2},-\frac{1}{2})^\top$
$(\frac{1}{2},-\frac{1}{2})^\top$

FIG. 4.1. *The triangulation of P in Example* 3.1 *underlying the algorithm.*

ALGORITHM.

Step 0 : Take any triangulation of the polytope P. Let $j := 0,\ t := 1,\ x^1 := x^-,\ \tau^j := \tau(x^1)$, and let x^{t+1} be the unique vertex of the simplex in $F(I)$ having τ^j as a facet opposite to it, where I is a set with $n-1$ elements as determined in Lemma 4.3. Go to Step 1.

Step 1 : Let σ be equal to the convex hull of x^{t+1} and τ^j. Pivot $(1, -f(x^{t+1})^\top)^\top$ lexicographically into matrix $A_{\tau^j,I}$. As described in Lemma 4.6, a unique column k from the first n columns of $A_{\tau^j,I}$ will be replaced. If the column k is an index column, then go to Step 3. Otherwise, go to Step 2.

Step 2 : Set $j := j+1$ and let τ^j be the facet of σ opposite the vertex x^k. If $\tau^j = \langle x^+\rangle$, then the algorithm terminates. If τ^j lies in $F(\bar{I})$, where $\bar{I} = I \cup \{l\}$ for some $l \in I_m \setminus I$, then go to Step 4. Otherwise there is exactly another simplex $\widetilde{\sigma}$ in $F(I)$ having τ^j as its facet. Go to Step 1 with x^{t+1} as the unique vertex in $\widetilde{\sigma}$ opposite to the facet τ^j.

Step 3 : Let $i \in I$ be the index corresponding to the column k. Set $I := I\setminus\{i\}$. There is a unique simplex $\widetilde{\sigma}$ in $F(I)$ having σ as a facet. Set $j := j+1,\ t := t+1$, and go to Step 1 with x^{t+1} as the unique vertex in $\widetilde{\sigma}$ opposite to σ, and $\tau^j := \sigma$.

Step 4 : Set $\sigma := \tau^j$. Pivot $(0, a^{l\top})^\top$ lexicographically into matrix $A_{\tau^j,I}$. By Lemma 4.8 there is a unique column k of the first n columns of $A_{\tau^j,I}$ which has to be replaced. If the column k is an index column, then go to Step 3. Otherwise go to Step 2 with $t := t-1$ and $I := \bar{I}$.

It is worth mentioning that the algorithm can also start with the simplex $\langle x^+\rangle$ and terminates with the simplex $\langle x^-\rangle$. Following the above description, it is fairly easy to implement the algorithm on a computer.

From Theorem 4.9 and the system of equations $(*)$ we see that every simplex from $\langle x^-\rangle$ to $\langle x^+\rangle$ contains a parametrized stationary point of the piecewise linear approximation f of φ with respect to $\mathcal{T}$. By taking the straight line segments between the parametrized stationary points of any two adjacent simplices, we obtain a piecewise linear path of parametrized stationary points of f connecting the points x^- and x^+.

COROLLARY 4.10. *Let $\mathcal{T}$ be any triangulation of P. Then with respect to the vector c there is a piecewise linear path of parametrized stationary points of the piecewise linear approximation f of φ with respect to $\mathcal{T}$ and this path connects x^-*

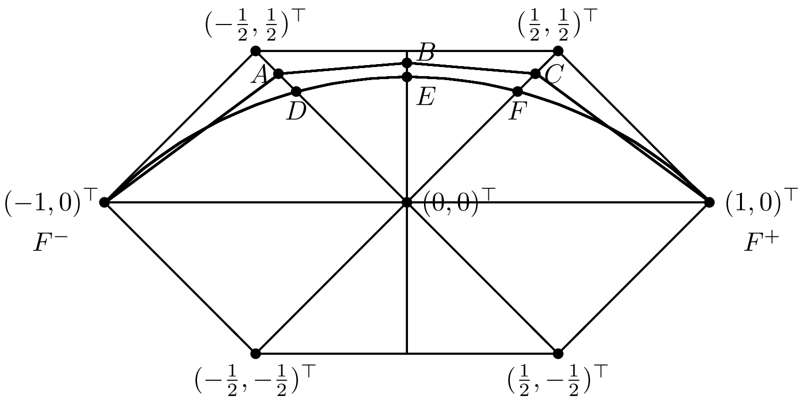

$(-\frac{1}{2},\frac{1}{2})^\top$
$(\frac{1}{2},\frac{1}{2})^\top$
B
A
C
E
D
F
$(-1,0)^\top$
F^-
$(0,0)^\top$
$(1,0)^\top$
F^+
$(-\frac{1}{2},-\frac{1}{2})^\top$
$(\frac{1}{2},-\frac{1}{2})^\top$

FIG. 4.2. *The piecewise linear path generated by the algorithm in Example* 3.1.

and x^+*.*

Now we will illustrate the algorithm with Example 3.1. To implement the algorithm, we take the triangulation depicted in Figure 4.1. Note that this triangulation is very coarse. The piecewise linear path generated by the algorithm is given by the sequence of line segments $[(-1,0)^\top,(-a,a)^\top]$, $[(-a,a)^\top,(0,b)^\top]$, $[(0,b)^\top,(a,a)^\top]$, and $[(a,a)^\top,(1,0)^\top]$, where $a=(5-\sqrt{7})^{-1}$ and $b=(5-2\sqrt{2})^{-1}$; see Figure 4.2. As we see from Figure 4.2 the path of approximate zero points generated by the algorithm and the set of the actual zero points of f are quite close. The finer the triangulation is the more accurate the approximation will be. In this figure $A=(-a,a)^\top$, $B=(0,b)^\top$, $C=(a,a)^\top$, $D=(-d,d)^\top$, $E=(0,e)^\top$, and $F=(d,d)^\top$ with $d=\sqrt{3}/2-1/2$ and $e=\sqrt{2}-1$.

5. Proofs for the existence theorems. In this section we still assume that x^- and x^+ are unique. First it will be argued in Theorems 5.1 and 5.2 that the points lying on the path given in Corollary 4.10 indeed all correspond to approximate parametrized stationary or zero points of the mapping φ. To show this, a sequence of triangulations $\mathcal{T}^r$ with mesh size converging to zero is taken. This yields, according to Corollary 4.10, for every $r\in\mathbb{N}$, a continuous piecewise linear function $\pi^r:[0,1]\to P$ with image set $\pi^r([0,1])$ connecting x^- and x^+. It will be shown that if q^r is an arbitrary point in $\pi^r([0,1])$ and the sequence $(q^r)_{r\in\mathbb{N}}$ converges to q, then q is a parametrized stationary point of $\varphi(q)$ with respect to c. Under the conditions of Theorems 3.2, 3.3, and 3.4, the piecewise linear approximation can be chosen in such a way that q is a zero point of φ. Furthermore, it will be shown in Theorem 5.4 by a limiting argument that there exists a connected set of zero points of φ, containing both x^- and x^+, that is being approximated.

THEOREM 5.1. *Let $\varphi:P\Longrightarrow\mathbb{R}^n$ be a correspondence satisfying Assumption* 2.1 *and let $c\in\mathbb{R}^n\setminus\{0^n\}$. For $r\in\mathbb{N}$, let $\mathcal{T}^r$ be a triangulation of P with mesh size smaller than $\frac{1}{r}$, let $f^r:P\to\mathbb{R}^n$ be a piecewise linear approximation of φ with respect to $\mathcal{T}^r$, and let $\pi^r:[0,1]\to P$ be the corresponding continuous function with image set connecting x^- and x^+. Let $(q^r)_{r\in\mathbb{N}}$ be an arbitrary convergent sequence of points in P with limit q^* where $q^r\in\pi^r([0,1])$. Then q^* is a parametrized stationary point of φ*

with respect to c.

Proof. Let $(\lambda_1^r, \dots, \lambda_{n+1}^r, x^{1^r}, \dots, x^{n+1^r}, s^{1^r}, \dots, s^{n+1^r})_{r\in\mathbb{N}}$ be a sequence of points in $\mathbb{R}_+^{n+1} \times \prod_{i=1}^{n+1} P \times \prod_{i=1}^{n+1} \mathbb{R}^n$ satisfying that $\sum_{j=1}^{n+1} \lambda_j^r = 1$, $\sigma^r(x^{1^r}, \dots, x^{n+1^r})$ is a simplex of $\mathcal{T}^r$, $q^r = \sum_{j=1}^{n+1} \lambda_j^r x^{j^r} \in \pi^r([0,1])$, and $s^{j^r} = f^r(x^{j^r})$. Notice that it may happen that $\lambda_j^r = 0$ for some $j \in I_{n+1}$. By definition, $f^r(q^r) = \sum_{j=1}^{n+1} \lambda_j^r s^{j^r}$. Define $s^r = f^r(q^r)$; then $s^r = \beta^r c + \sum_{j\in I^r} \mu_j^r a^j$ for some $\beta^r \in \mathbb{R}$ for some $\mu_j^r \geq 0\ \forall j \in I^r$, and for some $I^r \in \mathcal{I}$ satisfying that q^r lies in $F(I^r)$. Since $\cup_{q\in P}\varphi(q)$ is bounded, the sequence given above remains in a compact set, and without loss of generality it can be assumed to converge to an element $(\lambda_1^*, \dots, \lambda_{n+1}^*, x^{*1}, \dots, x^{*n+1}, s^{*1}, \dots, s^{*n+1})$. Define $s^* = \sum_{j=1}^{n+1} \lambda_j^* s^{*j}$. Clearly, it holds that $s^r \to s^*$. Since for every $r \in \mathbb{N}$ the mesh size of $\mathcal{T}^r$ is smaller than $\frac{1}{r}$, it holds for every $j \in I_{n+1}$ that $x^{*j} = q^*$. Using that φ is upper semicontinuous this implies that for every $j \in I_{n+1}$, $s^{*j} \in \varphi(q^*)$. Moreover, $\beta^r \to \beta^*$ for some number β^*, without loss of generality $I^r = I^* \forall r$ for some $I^* \in \mathcal{I}$, and $\mu_j^r \to \mu_j^* \forall j \in I^*$ for some nonnegative μ_j^*. Since φ is convex valued, $\sum_{j=1}^{n+1} \lambda_j^* = 1$ and $\lambda_j^* \geq 0\ \forall j \in I_{n+1}$, it holds that $s^* \in \varphi(q^*)$. Moreover, $q^* \in F(I^*)$ and $s^* = \beta^* c + \sum_{j\in I^*} \mu_j^* a^j \in A(I^*)$. Hence, according to Lemma 2.3, q^* is a parametrized stationary point of φ with respect to c. □

In order to give a constructive proof of Theorems 3.2, 3.3, and 3.4, the piecewise linear approximation f of φ with respect to a triangulation $\mathcal{T}$ should be chosen as follows. We call such a piecewise linear approximation a proper one. In case of Theorem 3.2 any piecewise linear approximation of φ with respect to $\mathcal{T}$ can be chosen. Next consider Theorem 3.3. If a point x in the (relative) interior of a face $F(I)$ is a vertex of a simplex of the triangulation, this implies that at least one element in $\varphi(x)$ lies in the set $A_0^*(I)$, and this element is assigned to the piecewise linear approximation at x. In the case of Theorem 3.4 an element of the set $A^*(I)$ in $\varphi(x)$ is assigned to a vertex x of a simplex if x lies in (the interior of) $F(I)$.

THEOREM 5.2. *Let $\varphi : P \Longrightarrow \mathbb{R}^n$ be a correspondence satisfying the conditions in one of Theorems* 3.2, 3.3, *or* 3.4. *For $r \in \mathbb{N}$, let $\mathcal{T}^r$ be a triangulation of P with mesh size smaller than $\frac{1}{r}$ and let $f^r : P \to \mathbb{R}^n$ be a proper piecewise linear approximation of φ with respect to $\mathcal{T}^r$. Let $(q^r)_{r\in\mathbb{N}}$ be an arbitrary convergent sequence of points in P with limit q^* such that for any $r \in \mathbb{N}$ it holds that $q^r \in \pi^r([0,1])$. Then q^* is a zero point of φ.*

Proof. First consider Theorem 3.2. Following the proof of Theorem 5.1 the limit point $s^* \in \varphi(q^*)$ is an element of $A(I^*)$, whereas q^* is an element of $F(I^*)$. The latter property implies that s^* is not an element of $A(I^*)$, unless $s^* = 0^n$. Hence, $s^* = 0^n$. Next consider Theorem 3.3. Consider again the convergent sequence of simplices σ^r in $F(I^*)$ mentioned in the proof of Theorem 5.1. Then the vertex x^{j^r} of σ^r lies in some face $F(I^{j^r})$ of P with $I^* \subset I^{j^r}$. Hence, we have that $A_0^*(I^{j^r}) \subset A_0^*(I^*)$, and so $s^r \in A_0^*(I^*) \forall r$ because of the properness of f^r. Consequently, $s^* \in A_0^*(I^*)$ and therefore $s^* \in A_0^*(I^*) \cap A(I^*) \subset \{0^n\}$, i.e., $s^* = 0^n$. In the case of Theorem 3.4, following a similar argument as in the previous case we obtain that $s^* \in A^*(I^*) \cap A(I^*)$. Since the latter intersection consists of only the zero vector, we obtain again that $s^* = 0^n$. □

From Theorem 5.2 the next result follows immediately.

COROLLARY 5.3. *Let $\varphi : P \Longrightarrow \mathbb{R}^n$ be a correspondence satisfying the conditions in one of Theorems* 3.2, 3.3, *or* 3.4. *For $r \in \mathbb{N}$, let $\mathcal{T}^r$ be a triangulation of P with mesh size smaller than $\frac{1}{r}$ and let $f^r : P \to \mathbb{R}^n$ be a proper piecewise linear approximation of φ with respect to $\mathcal{T}^r$. Then for every $\varepsilon > 0$ there exists an $R \in \mathbb{N}$*

such that for every $r \geq R$ it holds that $q^r \in \pi^r([0,1])$ implies $\|f^r(q^r)\|_\infty < \varepsilon$.

Proof. Suppose that a sequence $(q^r, f^r(q^r))_{r\in\mathbb{N}}$ exists with $q^r \in \pi^r([0,1])$ and $\|f^r(q^r)\|_\infty \geq \varepsilon$ for every $r \in \mathbb{N}$. Since P and $\cup_{q\in P}\varphi(q)$ are compact, there exists a converging subsequence $(q^{r^s}, f^{r^s}(q^{r^s}))_{s\in\mathbb{N}}$, with limit say (q^*, s^*), where $\|s^*\|_\infty \geq \varepsilon > 0$. As in the proof of Theorem 5.2 it can be shown that $s^* = 0^n$, yielding a contradiction. □

Using Theorem 4.9 and Theorem 5.2 it will be shown that there exists a connected set C in P such that $x^- \in C$, $x^+ \in C$, and $0^n \in \varphi(q)\ \forall q \in C$. Hence, there is a continuum of zero points of φ being approximated by the algorithm of section 4. For a nonempty, compact set $S \subset \mathbb{R}^n$, define the continuous function $d_S : \mathbb{R}^n \to \mathbb{R}$ by $d_S(x) = \min\{\|x-y\|_\infty \mid y \in S\}$.

THEOREM 5.4. *Let $\varphi : P \Longrightarrow \mathbb{R}^n$ be a correspondence satisfying the conditions in one of Theorems* 3.2, 3.3, *or* 3.4. *Then there exists a connected set C of points in P such that $x^- \in C$, $x^+ \in C$, and $0^n \in \varphi(q)\ \forall q \in C$.*

Proof. Define $Q = \{q \in P \mid 0^n \in \varphi(q)\}$. From the conditions of the theorems it immediately follows that $x^- \in Q$, $x^+ \in Q$, and Q is compact. Suppose the theorem is false. Then x^+ is not an element of the component of Q containing x^-. By Munkres [21, p. 235] it holds for every compact set X in some Euclidean space and for every element $x \in X$ that the component of X containing x equals the intersection of all sets containing x which are both open and closed in X. Hence, there exists a set Q^0, which is open and closed in Q, such that $x^- \in Q^0$ and $x^+ \notin Q^0$. Define $Q^1 = Q \setminus Q^0$. Then Q^1 is open and closed in Q, $x^- \notin Q^1$, and $x^+ \in Q^1$. Since Q is compact, it follows that Q^0 and Q^1 are disjoint, compact sets. Hence, there exists $\varepsilon > 0$ such that $\min\{\|q^0 - q^1\|_\infty \mid q^0 \in Q^0,\ q^1 \in Q^1\} \geq \varepsilon$. For every $r \in \mathbb{N}$, let $\mathcal{T}^r$ be a triangulation of P with mesh size smaller than $\frac{1}{r}$, let $f^r : P \to \mathbb{R}^n$ be a proper piecewise linear approximation of φ with respect to $\mathcal{T}^r$, and let $\pi^r : [0,1] \to P$ be the corresponding continuous function with image set connecting x^- and x^+. Define $g^r : [0,1] \to \mathbb{R}$ by

$$g^r(t) = d_{Q^0}(\pi^r(t)) - d_{Q^1}(\pi^r(t))\ \forall t \in [0,1].$$

Since g^r is continuous, $g^r(0) \leq -\varepsilon$, and $g^r(1) \geq \varepsilon$, there exists a point $t^r \in [0,1]$ such that $g^r(t^r) = 0$. Hence, $d_{Q^0}(\pi^r(t^r)) = d_{Q^1}(\pi^r(t^r)) = d_Q(\pi^r(t^r)) \geq \frac{1}{2}\varepsilon$. Without loss of generality, it can be assumed that $(\pi^r(t^r))_{r\in\mathbb{N}}$ converges to a point $q^* \in P$. Hence,

$$d_Q(q^*) = d_Q\left(\lim_{r\to\infty} \pi^r(t^r)\right) = \lim_{r\to\infty} d_Q(\pi^r(t^r)) \geq \frac{1}{2}\varepsilon > 0.$$

However, by Theorem 5.2, $d_Q(q^*) = 0$, yielding a contradiction. □

Similarly, one can easily show that in the case of Theorem 3.1 there exists a connected set C of parametrized stationary points in P with respect to c such that $x^- \in C$ and $x^+ \in C$.

6. The general case. The algorithm proposed in the previous section can be adapted for computing a continuum of parametrized stationary points or zero points of φ on P in case the faces F^- and F^+ are not vertices of P. First we take any point v in the interior of F^- and a triangulation of P such that the face F^- itself is being triangulated according to the V-triangulation of Talman and Yamamoto [24]. For $J \in \mathcal{I}$ such that $I^- \subset J$, let $VF(J) = \{x \in F^- | x = \lambda v + (1-\lambda)y, 0 \leq \lambda \leq 1, y \in F(J)\}$. The V-triangulation subdivides any such set $VF(J)$ into $(n - |J| + 1)$-simplices.

Then we apply the algorithm of Talman and Yamamoto [24] to find a parametrized stationary point x^- in F^- of the piecewise linear approximation f of the correspondence φ on P with respect to c. To initiate the algorithm, we solve $\max x^\top f(v)$

subject to $x \in F^-$, which yields, by using Theorem 2.6 of Fujishige and Yang [10] for its dual, a uniquely determined vertex $F(I_0)$ of F^- as a solution, with $|I_0| = n$ and $I^- \subset I_0$. Then starting with the 0-simplex $\{v\}$ and I equal to I_0 the algorithm generates a sequence of adjacent simplices in $VF(I)$ for varying $I \in \mathcal{I}$ such that $I^- \subset I$, and the common facets $\tau(x^1, \dots, x^t)$ satisfy that the system of equations

$$\sum_{j=1}^{t} \lambda_j \begin{pmatrix} 1 \\ -f(x^j) \end{pmatrix} + \sum_{i \in I} \mu_i \begin{pmatrix} 0 \\ a^i \end{pmatrix} = \begin{pmatrix} 1 \\ 0^n \end{pmatrix}$$

has a solution (λ, μ) satisfying $\lambda_j \geq 0, j \in I_t, \mu_i \geq 0$ for $i \in I \backslash I^-$. Notice that $I^- \subset I$ and that we allow I to be equal to I^-.

The algorithm of Talman and Yamamoto stops as soon as a $(t-1)$-simplex $\tau^-(x^1, \dots, x^t)$ in $F(I)$ for some I containing I^- is generated for which the system has a solution (λ, μ). Then $x = \sum_{j=1}^{t} \lambda_j x^j$ is a parametrized stationary point in $F(I^-)$ with respect to c of the piecewise linear approximation f of φ. Next the vector $(c^\top, 0)$ is pivoted semilexicographically into the system, making any $\mu_i, i \in I$, nonnegative. Since $-c = \sum_{i \in I^-} \lambda_i a^i$ for unique $\lambda_i > 0$, one of the μ_i's, say μ_{i_0}, for some $i_0 \in I^-$, will leave the basis. Now the algorithm continues in $F(I \backslash \{i_0\})$ with the unique t-simplex σ in $F(I \backslash \{i_0\})$ having τ^- as a facet and a semilexicographic pivot step is made with $(1, -f^\top(x^{t+1}))^\top$ where x^{t+1} is the vertex of σ opposite to τ, and so on.

In this way the algorithm generates for varying $I \in \mathcal{I}$ by semilexicographic pivoting a unique sequence of adjacent simplices in $F(I)$ with common I-complete facets until a facet τ being a simplex in $F(H)$ for some H containing I^- or a complete facet τ^+ being a simplex in $F(J)$ for some J containing I^+ is generated. In the former case the algorithm continues in the subset $VF(H)$ of $F(I^-)$ as above until again an I-complete simplex in $F(I)$ for some I containing I^- is found, and so on. In the latter case the point $x^+ = \sum_{j=1}^{t} \lambda_j x^j$ at the solution (λ, μ, β) lies in τ^+ and is a parametrized stationary point in F^+ of f with respect to c.

Letting x^- be the last point being generated in F^-, the algorithm generates a piecewise linear path of parametrized stationary points of f with respect to the vector c. This path connects the point x^- in the face F^- with a point x^+ in the face F^+. Taking a sequence of triangulations of P with mesh tending to zero, in the limit a connected set of parametrized stationary points of φ is obtained with respect to c connecting the faces F^- and F^+. In case the correspondence φ satisfies the conditions of Theorems 3.2, 3.3, or 3.4 and the piecewise linear approximations are chosen in an appropriate way, there exists a connected set of zero points of φ connecting F^- and F^+.

Notice that if a sequence of triangulations with mesh tending to zero is taken, for any triangulation in this sequence the points x^- in F^- and x^+ in F^+ being connected through the piecewise linear path generated by the algorithm may differ. In the limit these points converge on a subsequence to two different zero points of φ, one lying in F^- and the other lying in F^+.

7. Examples. In this section we will derive three existing existence theorems from Theorems 3.2, 3.3, and 3.4.

The first example is derived from general equilibrium theory with price rigidities. Let $U^n = \{x \in \mathbb{R}^n \mid 0 \leq x_i \leq 1 \; \forall i \in I_n\}$ be the n-dimensional unit cube. Let 1^n denote the n-vector of ones and for $i \in I_n$ let $e(i)$ denote the ith unit vector in $\mathbb{R}^n$. Under standard assumptions on economic primitives (see, e.g., Herings [12] for details), it can be derived that the total excess demand function of an economy with completely fixed prices $p \in \mathbb{R}^n_{++}$ is a continuous function $f : U^n \to \mathbb{R}^n$ with the following properties:

(A) for every $x \in U^n$, for every $j \in I_n$, $x_j = 0$ implies $f_j(x) \geq 0$, and $x_j = 1$ implies $f_j(x) \leq 0$;

(B) for every $x \in U^n$, $p^\top f(x) = 0$.

THEOREM 7.1. *Let $f : U^n \to \mathbb{R}^n$ be the excess demand function of an economy with completely fixed prices $p \in \mathbb{R}^n_{++}$. Then there exists a connected set C of zero points of f such that $0^n \in C$ and $1^n \in C$.*

Proof. We can rewrite the set U^n as

$$U^n = \{x \in \mathbb{R}^n \mid a^{i\top} x \leq b_i \ \ \forall\, i \in I_{2n}\},$$

where $a^i = e(i)$, $b_i = 1$, $a^{n+i} = -e(i)$, and $b_{n+i} = 0 \forall i \in I_n$. We can partition any $I \in \mathcal{I}$ into two disjoint subsets I^1 and I^2 with $I^1 \subset I_n$ and $I^2 \subset I_{2n} \setminus I_n$. Notice that $i \in I^1$ implies $i + n \notin I^2$, and $i \in I^2$ implies $i - n \notin I^1$.

Let $c = p$. Clearly, $x^- = 0^n$ and $x^+ = 1^n$. We check the condition of Theorem 3.2. Suppose $0^n \neq f(x) \in A(I)$ for some $I \in \mathcal{I}$, so $f(x) = \sum_{i \in I^1} \mu_i e(i) - \sum_{i \in I^2} \mu_i e(i - n) + \beta p$, where $\mu_i \geq 0\ \forall i \in I$, and $\beta \in \mathbb{R}$, and some μ_i or β nonzero. Conditions (A) and (B) imply that $f(0^n) = 0^n$ and $f(1^n) = 0^n$, so $I \neq I_n$ and $I \neq I_{2n} \setminus I_n$. If $I = \emptyset$, then $0 = p^\top f(x) = p^\top \beta p \neq 0$, a contradiction. Take any nonempty set $I \in \mathcal{I}$ not equal to I_n or $I_{2n} \setminus I_n$.

If $I^1 \neq \emptyset$ and $I^2 \neq \emptyset$, then $p_i(\beta + \mu_i) \leq 0$, for $i \in I^1$, and $p_{i-n}(\beta - \mu_i) \geq 0$, for $i \in I^2$, so $\beta = 0$ and $\mu_i = 0 \forall i \in I$, a contradiction.

If $I^1 = \emptyset$, then $I^2 \neq \emptyset$, so $\beta > 0$, but then $f(x) > 0^n$, a contradiction to $0 = p^\top f(x)$.

If $I^2 = \emptyset$, then $I^1 \neq \emptyset$, so $\beta < 0$, but then $f(x) < 0^n$, a contradiction to $0 = p^\top f(x)$. Consequently, Theorem 3.2 holds. □

The proof of Theorem 7.1 shows that our results apply to economies with completely fixed prices $p \in \mathbb{R}^n_{++}$. Since the results apply to all such economies, that is, for all possible specifications of utility functions and initial endowments, it also shows that our results satisfy a certain kind of robustness. The above theorem can also be derived from Theorems 3.3 and 3.4.

The following result is obtained by Herings, Talman, and Yang [15, Theorem 4.3, p. 690] which generalizes Theorem 7.1. It is therefore also related to economies with price rigidities, and it applies even under somewhat weaker assumptions on the economic primitives. We will show that this theorem is a special case of Theorem 3.3.

THEOREM 7.2. *Let $\varphi : U^n \Longrightarrow \mathbb{R}^n$ be any correspondence satisfying Assumption 2.1. Moreover, it holds that*

(A) *for every $x \in U^n$, there exists $f \in \varphi(x)$ such that, for every $j \in I_n$, $x_j = 0$ implies $f_j \geq 0$, and $x_j = 1$ implies $f_j \leq 0$;*

(B) *for every $x \in U^n$, for every $f \in \varphi(x)$, there exists some $p \in \mathbb{R}^n_{++}$ such that $p^\top f = 0$.*

Then there exists a connected set C of zero points of φ such that $0^n \in C$ and $1^n \in C$.

Proof. Rewrite the set U^n as in the proof of Theorem 7.1. Let $c = 1^n$. Clearly, $x^- = 0^n$ and $x^+ = 1^n$. Moreover, it is easy to verify that condition (ii) of Theorem 3.3 is satisfied by condition (A). We have to check condition (i). For $I = \emptyset$, condition (i) is trivially satisfied. Now take any nonempty set I from $\mathcal{I}$. We partition I into two disjoint subsets I^1 and I^2 as in the proof of Theorem 7.1. Suppose there is some $x \in F(I)$ and some $f \in \varphi(x)$ such that $f \in (A_0^*(I) \cap A(I)) \setminus \{0^n\}$. This implies that there exist $\mu_i \geq 0 \; \forall \; i \in I$ and $\beta \in \mathbb{R}$ such that

$$\begin{array}{rcl} f & = & \sum_{i \in I} \mu_i a^i + \beta c, \\ f^\top a^i & \leq & 0 \quad \forall \; i \in I, \\ f & \neq & 0^n. \end{array}$$

Equivalently,

$$\begin{array}{rcl} f & = & \sum_{i \in I^1} \mu_i e(i) - \sum_{i \in I^2} \mu_i e(i-n) + \beta c, \\ f_i & \leq & 0 \quad \forall \; i \in I^1, \\ f_i & \geq & 0 \quad \forall \; n + i \in I^2, \\ f & \neq & 0^n. \end{array}$$

This implies that

$$\begin{array}{rcl} \beta & \geq & \max_{i \in I^2} \mu_i, \\ \beta & \leq & -\max_{i \in I^1} \mu_i. \end{array} \tag{7.1}$$

In case $I^2 = \emptyset$, we have $f < 0^n$. This contradicts condition (B). In case $I^1 = \emptyset$, we have $f > 0^n$, again contradicting condition (B). In cases $I^1 \neq \emptyset$ and $I^2 \neq \emptyset$, without loss of generality there exist $i \in I^1$ and $j \in I^2$ such that $\mu_i > 0$ and $\mu_j > 0$. This would mean both $\beta > 0$ and $\beta < 0$ from (7.1) which is impossible. Hence condition (i) is satisfied. □

Now we use Theorem 3.4 to show that there is a continuum of constrained equilibria in a pure exchange economy with general price rigidities; see, e.g., Schalk and Talman [23] for details. Price vectors in such an economy with n commodities are restricted to an n-dimensional simple polytope

$$P = \{p \in \mathbb{R}^n_+ \mid t^- \leq p^\top c \leq t^+, \; {a^i}^\top p \leq b_i, \; i \in I_m\}$$

for some strictly positive vector c with length 1, $0 < t^- < t^+$, ${a^i}^\top c = 0 \; \forall i \in I_m$. We also assume that there are no redundant constraints and that P is a subset of $\mathbb{R}^n_{++}$. We define

$$Q = \{q \in \mathbb{R}^n \mid t^- \leq q^\top c \leq t^+, \; a^{i\top} q \leq b_i + \varepsilon \; \forall i \in I_m\}$$

for some $\varepsilon > 0$. For ε small enough, $F(I)$, $I \subset I_m$, is a face of Q if and only if $\{x \in P \mid a^i x = b_i, \; i \in I\}$ is a face of P. For any q in Q, let $p(q)$ be the orthogonal projection of q on P and let $I(q)$ be such that $p(q)$ is in the interior of the face $F(I(q))$ of P. Then there exist unique nonnegative numbers $\mu^i(q), i \in I(q)$, such that $q = p(q) + \sum_{i \in I(q)} \mu_i(q) a^i$.

At $q \in Q$ define the price vector by $p(q)$ and a continuous rationing scheme $(r^i(q), d_i(q))$, $i \in I_m$, such that $r^i(q) = a^i$, $d_i(q) = 0$ if $\mu_i(q) = 1$ and $d_i(q) = M$ if $\mu_i(q) = 0$ or i not in $I(q)$, for sufficiently large $M > 0$. This rationing scheme determines the constraints on the net-supply of the consumers. Given a utility function u^h and initial endowment w^h, consumer $h \in H$ maximizes his utility $u^h(x)$ over his budget constraint given by $p(q)^\top x \leq p(q)^\top w^h$ and rationing constraints $r^i(q)^\top (x - w^h) \leq d_i(q), i \in I_m$. The solution set $x^h(q)$ yields the constraint excess demand set $z^h(q) = x^h(q) - \{w^h\}$ of consumer $h \in H$ at q. Adding up these sets over all consumers in H gives the total constraint excess demand correspondence $\zeta : Q \to \mathbb{R}^n$. Under certain standard economic conditions, the correspondence ζ satisfies Assumption 2.1 and $p(q)^\top z = 0$ for any $z \in \zeta(q)$, $q \in Q$.

A constrained equilibrium is obtained if $q \in P$ is such that $0^n \in \zeta(q)$. At such an equilibrium, $p(q)^\top a^i = b_i$ if the ith rationing scheme is binding.

Let F^- be the face of Q on which $c^\top x$ is minimized on Q and let F^+ be the face of Q on which $c^\top x$ is maximized on Q. Notice that $F^- = \{q \in Q \mid q^\top c = t^-\}$ and $F^+ = \{q \in Q \mid q^\top c = t^+\}$.

To see that there is a connected set of constrained equilibria linking F^- and F^+, define the mapping φ on Q by $\varphi(q) = \{y | y = z - (c^\top z)c, z \in \zeta(q)\}, q \in Q$. We will show that $\varphi(q) \subset A^*(I)$ when q lies in the interior of the face $F(I)$ of Q. Clearly, $c^\top y = 0$ for any $y \in \varphi(q)$. Moreover, for any $y = z - (c^\top z)c \in \varphi(q)$ and $i \in I(q) \cap I_m$ it holds that ${a^i}^\top y = {a^i}^\top z - ({a^i}^\top c)c^\top z = {r^i(q)}^\top z \leq 0$. Therefore, any $y \in \varphi(q)$ is an element of $A^*(I)$, and hence $\varphi(q) \subset A^*(I)$ if $q \in F(I)$. According to Theorem 3.4 there exists a connected set C in Q intersecting both F^- and F^+ such that every point $q \in C$ is a zero point of φ, i.e., $0^n \in \varphi(q)$. For such a q it holds that there is $z \in \zeta(q)$ satisfying $z - (c^\top z)c = 0^n$. Because $p(q)^\top z = 0$, we obtain that $(c^\top z)(p(q)^\top c) = 0$. Hence, $c^\top z = 0$, since $p(q)^\top c > 0$. This implies that $z = 0^n$ and therefore $0^n \in \zeta(q)$, inducing a constrained equilibrium. Consequently there exists a connected set of constrained equilibria linking the two faces F^- and F^+.

Finally, we show that the fundamental fixed point theorems of Browder [4] and Mas-Colell [20] can also be derived from Theorem 3.4. Browder proved the continuous function case and Mas-Colell extended the result to the upper semicontinuous correspondence case.

THEOREM 7.3. *Let P be an n-dimensional polytope and let $\varphi : P \times [0,1] \Longrightarrow P$ be any correspondence satisfying Assumption* 2.1. *Then the set*

$$D = \{(x,t) \in P \times [0,1] \mid x \in \varphi(x,t)\}$$

contains a connected set C such that

$$C \cap (\mathrm{P} \times \{0\}) \neq \emptyset \quad \textit{and} \quad C \cap (\mathrm{P} \times \{1\}) \neq \emptyset.$$

Proof. We can rewrite the set $P \times [0,1]$ as

$$\begin{aligned} W = \{(x,t) \in \mathbb{R}^{n+1} \quad | \quad & ({a^i}^\top, 0)(x^\top, t)^\top \leq b_i \quad \forall i \in I_m, \\ & (0, \ldots, 0, -1)(x^\top, t)^\top \leq 0, \\ & (0, \ldots, 0, 1)(x^\top, t)^\top \leq 1\}. \end{aligned}$$

Let $c = (0, \ldots, 0, 1)^\top \in \mathbb{R}^{n+1}$. Obviously, W is simple and no constraint is redundant. Moreover, $F^+ = \{(x,t) \in W \mid t = 1\}$ and $F^- = \{(x,t) \in W \mid t = 0\}$.

Construct the correspondence $\psi : W \Longrightarrow \mathbb{R}^{n+1}$ as

$$\psi(x,t) = (\varphi(x,t) - \{x\}) \times \{0\}.$$

We will show that for any $I \in \mathcal{I}$ and any $(x,t) \in F(I)$, we have $\psi(x,t) \subset A^*(I)$. For $I = \emptyset$, $\psi(x,t) \subset A^*(\emptyset)$ since $c^\top z = 0$ for any $z \in \psi(x,t)$. Now take any nonempty set I from $\mathcal{I}$. We have to consider the following two cases:

(1) In case $I \subset I_m$, take any (x,t) in the face $F(I)$ of W and any z in $\psi(x,t)$. We have

$${a^i}^\top x = b_i \ \forall i \in I; \ z = ((k-x)^\top, 0)^\top$$

for some $k \in P$. Take any $y \in A(I)$. That is,

$$y = \sum_{i \in I} \lambda_i ({a^i}^\top, 0)^\top + \beta c$$

for some $\lambda_i \geq 0$ and $\beta \in \mathbb{R}$. Thus, we have

$$z^\top y = \sum_{i \in I} \lambda_i (k - x)^\top a^i.$$

Since $k \in P$, we have $a^{i\top} k \leq b_i$ for $i \in I_m$. Since $a^{i\top} x = b_i$ for $i \in I$, we have $z^\top y \leq 0$. This means that $\psi(x, t)$ is a subset of $A^*(I)$.

(2) If $I \not\subset I_m$, either $m + 1$ or $m + 2$ is contained in I. For example, suppose that $m + 2$ is in I, i.e., $t = 1$. The case $m + 1 \in I$ follows the same argument. Take any (x, t) in the face $F(I)$ of W and any z in $\psi(x, t)$. We have

$$a^{i\top} x = b_i \ \forall i \in I \setminus \{m + 2\}; \ t = 1; \ z = ((k - x)^\top, 0)^\top$$

for some $k \in P$. Take any $y \in A(I)$. That is,

$$y = \sum_{i \in I \setminus \{m+2\}} \lambda_i (a^{i\top}, 0)^\top + \beta c$$

for some $\lambda_i \geq 0$ and $\beta \in \mathbb{R}$. Thus, we have

$$z^\top y = \sum_{i \in I \setminus \{m+2\}} \lambda_i (k - x)^\top a^i.$$

Since $k \in P$, we have $a^{i\top} k \leq b_i$ for $i \in I_m$. Since $a^{i\top} x = b_i$ for $i \in I \setminus \{m + 2\}$, we have $z^\top y \leq 0$. This means that $\psi(x, t)$ is again a subset of $A^*(I)$. By Theorem 3.4 there exists a connected set C in W such that

$$0^{n+1} \in \psi(x, t) \quad \forall (x, t) \in C; \ F^+ \cap C \neq \emptyset; \ F^- \cap C \neq \emptyset.$$

Clearly, $x \in \varphi(x, t)$ for each $(x, t) \in C$. □

We remark that the above theorem can also be derived from Theorem 3.2 or Theorem 3.3.

Acknowledgment. Zaifu Yang would like to thank the CentER for Economic Research of Tilburg University for its hospitality.

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction,* Springer, Berlin, 1990.

[2] J. P. Aubin, *Optima and Equilibria*, Springer, Berlin, 1998.

[3] L. E. J. Brouwer, *Über Abbildung von Mannigfaltigkeiten*, Math. Ann., 71 (1912), pp. 97–115.

[4] F. E. Browder, *On continuity of fixed points under deformations of continuous mappings*, Summa Brasiliensis Mathematicae, 4 (1960), pp. 183–191.

[5] D. J. Brown, P. M. DeMarzo, and B. C. Eaves, *Computing zeroes of sections of vector bundles using homotopies and relocalization*, Math. Oper. Res., 21 (1996), pp. 26–43.

[6] P. M. DeMarzo and B. C. Eaves, *Computing equilibria of GEI by relocalization on a Grassmann manifold*, J. Math. Econom., 26 (1996), pp. 479–497.

[7] B. C. Eaves, *On the basic theory of complementarity*, Math. Programming, 1 (1971), pp. 68–75.

[8] B. C. Eaves, *Homotopies for computation of fixed points*, Math. Programming, 3 (1972), pp. 1–22.

[9] J. Freidenfelds, *A set intersection theorem and applications*, Math. Programming, 7 (1974), pp. 199–211.

[10] S. Fujishige and Z. Yang, *A lexicographic algebraic theorem and its applications*, Linear Algebra Appl., 279 (1998), pp. 75–91.

[11] P. Hartman and G. Stampacchia, *On some nonlinear elliptic differential functional equations*, Acta Math., 115 (1966), pp. 271–310.

[12] P. J. J. Herings, *Equilibrium existence results for economies with price rigidities*, Econom. Theory, 7 (1996), pp. 63–80.

[13] P. J. J. Herings, *On the existence of a continuum of constrained equilibria*, J. Math. Econom., 30 (1998), pp. 257–273.

[14] P. J. J. Herings and A. J. J. Talman, *Intersection theorems with a continuum of intersection points*, J. Optim. Theory Appl., 96 (1998), pp. 311–335.

[15] P. J. J. Herings, A. J. J. Talman, and Z. Yang, *The computation of a continuum of constrained equilibria*, Math. Oper. Res., 21 (1996), pp. 675–696.

[16] S. Kakutani, *A generalization of Brouwer's fixed point theorem*, Duke Math. J., 8 (1941), pp. 457–459.

[17] B. Knaster, C. Kuratowski, and C. Mazurkiewicz, *Ein Beweis des Fixpunktsatzes für n-dimensionale Simplexe*, Fund. Math., 14 (1929), pp. 132–137.

[18] G. van der Laan and A. J. J. Talman, *A restart algorithm for computing fixed points without an extra dimension*, Math. Programming, 17 (1979), pp. 74–84.

[19] G. van der Laan, A. J. J. Talman, and Z. Yang, *Existence and approximation of robust solutions of variational inequality problems over polytopes*, SIAM J. Control Optim., 37 (1998), pp. 333–352.

[20] A. Mas-Colell, *A note on a theorem of F. Browder*, Math. Programming, 6 (1974), pp. 229–233.

[21] J. Munkres, *Topology*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[22] H. Scarf, *The approximation of fixed points of a continuous mapping*, SIAM J. Appl. Math., 15 (1967), pp. 1328–1343.

[23] S. Schalk and A. J. J. Talman, *Rationing Equilibria under General Price Rigidities*, mimeo, Tilburg University, The Netherlands, 2000.

[24] A. J. J. Talman and Y. Yamamoto, *A simplicial algorithm for stationary point problems on polytopes*, Math. Oper. Res., 14 (1989), pp. 383–399.

[25] M. J. Todd, *The Computation of Fixed Points and Applications*, Lecture Notes in Econom. and Math. Systems 124, Springer, Berlin, 1976.

[26] Y. Yamamoto, *A path-following procedure to find a proper equilibrium of finite games*, Internat. J. Game Theory, 22 (1993), pp. 249–259.

[27] Z. Yang, *A simplicial algorithm for computing robust stationary points of a continuous function on the unit simplex*, SIAM J. Control Optim., 34 (1996), pp. 491–506.

[28] Z. Yang, *Computing Equilibria and Fixed Points*, Kluwer Academic Publishers, Boston, 1999.

SIAM J. Control Optim.
Vol. 39, No. 6, pp. 1874–1928

INPUT-OUTPUT-TO-STATE STABILITY*

MIKHAIL KRICHMAN[†], EDUARDO D. SONTAG[‡], AND YUAN WANG[§]

Abstract. This work explores Lyapunov characterizations of the input-output-to-state stability (IOSS) property for nonlinear systems. The notion of IOSS is a natural generalization of the standard zero-detectability property used in the linear case. The main contribution of this work is to establish a complete equivalence between the IOSS property and the existence of a certain type of smooth Lyapunov function. As corollaries, one shows the existence of "norm-estimators," and obtains characterizations of nonlinear detectability in terms of relative stability and of finite-energy estimates.

Key words. nonlinear stability, Lyapunov function techniques, input-output-to-state stability, input-to-state stability, observers, detectability

AMS subject classifications. 93B07, 93D05, 93D20, 93D09, 34D20

PII. S0363012999365352

1. Introduction.

This paper concerns itself with the following question, for dynamical systems: *is it possible to estimate, on the basis of external information provided by past input and output signals, the magnitude of the internal state $x(t)$ at time t?* The rest of this introduction will explain, in very informal and intuitive terms, the motivation for this question, closely related to the "zero-detectability" problem, sketching the issues that arise and the main results. Precise definitions are provided in the next section.

State estimation is central to control theory. It arises in signal processing applications (Kalman filters), as well as in stabilization based on partial information (observers). By and large, the theory of state estimation is well understood for linear systems, but it is still poorly developed for more general classes of systems, such as finite dimensional deterministic systems, with which this paper is concerned. An outstanding open question is the derivation of useful necessary and sufficient conditions for the existence of observers, i.e., "algorithms" (dynamical systems) which converge to an estimate $\hat{x}(t)$ of the state $x(t)$ of the system of interest, using the information provided by $\{u(s), s \leq t\}$, the set of past input values, and by $\{y(s), s \leq t\}$, the set of past output measurements. In the context of stabilization to an equilibrium, let us say to the zero state $x = 0$ if we are working in a Euclidean space, a weaker type of estimate is sometimes enough: it may suffice to have a norm-estimate, that is to say, an upper bound $\hat{x}(t)$ on the *magnitude* (norm) $|x(t)|$ of the state $x(t)$. Indeed, it is often the case (cf. [32] and Assumption UEC (73) in [18]) that norm-estimates suffice for control applications. To be more precise, one wishes that $\hat{x}(t)$ eventually becomes

*Received by the editors November 28, 1999; accepted for publication (in revised form) October 11, 2000; published electronically April 18, 2001.

http://www.siam.org/journals/sicon/39-6/36535.html

[†]Department of Mechanical and Aerospace Engineering, University of California, Los Angeles, 420 Westwood Plaza, 38-137 Engineering IV, Los Angeles, CA 90095 (krichman@seas.ucla.edu). The research of this author was supported in part by U.S. Air Force grant F49620-98-1-0242.

[‡]Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (sontag@control.rutgers.edu). The research of this author was supported in part by U.S. Air Force grant F49620-98-1-0421.

[§]Department of Mathematical Sciences, Florida Atlantic University, Boca Raton, FL 33431 (ywang@math.fau.edu). The research of this author was supported in part by NSF grant DMS-9457826.

an upper bound on $|x(t)|$ as $t \to \infty$. We are thus interested in *norm-estimators* which, when driven by the input/output data generated by the system, produce such an upper bound $\hat{x}(t)$; cf. Figure 1.1.

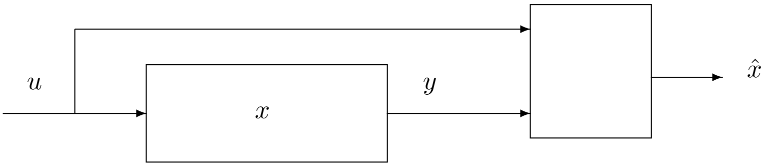


FIG. 1.1. *Norm-estimator.*

In order to understand the issues that arise, let us start by considering the very special case when the external data (inputs u and outputs y) vanish identically. The obvious estimate (assuming, as we will, that everything is normalized so that the zero state is an equilibrium for the unforced system, and the output is zero when $x = 0$) is $\hat{x}(t) \equiv 0$. However, the only way that this estimate fulfills the goal of upper bounding the norm of the true state as $t \to \infty$ is if $x(t) \to 0$. In other words, one obvious necessary property for the possibility of norm-estimation is that the origin must be a globally asymptotically stable state with respect to the "subsystem" consisting of those states for which the input $u \equiv 0$ produces the output $y \equiv 0$. One says in this case that the original system is *zero-detectable.* For *linear systems*, zero-detectability is equivalent to detectability, that is to say, the property that if any two trajectories produce the same output, then they approach each other. Zero-detectability is a central property in the general theory of nonlinear stabilization on the basis of output measurements; see, for instance, among many other references, [34, 17, 50, 10, 16]. Our work can be seen as a contribution toward the better characterization and understanding of this fundamental concept.

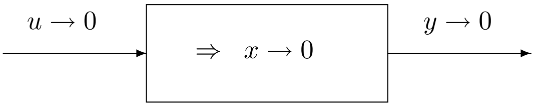


FIG. 1.2. *State converges to zero if external data does.*

However, zero-detectability by itself is far from being sufficient for our purposes, since it fails to be "well-posed" enough. One easily sees that, at the least, one should ask that, when inputs and outputs are small, states should also be small, and if inputs and outputs converge to zero as $t \to \infty$, states do too; cf. Figure 1.2. Moreover, when formally defining the notion of norm-estimator and the natural necessary and sufficient conditions for its existence, other requirements appear: the existence of asymptotic bounds on states, as a function of bounds on input/output data, and the need to describe the "overshoot" (transient behavior) of the state.

One way to approach the formal definition, so as to incorporate all the above characteristics in a simple manner, is to look at the analogous questions for the stability problem, which, for linear systems, is known to be technically dual to detectability. This leads one to the area which deals precisely with this circle of ideas: *input-to-state stability* (ISS).

ISS was introduced in [35] and has proved to be a very useful paradigm in the study of nonlinear stability; see, for instance, the textbooks [16, 20, 22, 23], and the papers [7, 14, 15, 19, 28, 11, 32, 33, 44, 42, 49, 48], as well as its variants such as integral ISS (cf. [2, 4, 24, 38]) and input/output stability (cf. [35, 46, 47]). The notion of ISS takes into account the effect of initial states in a manner fully compatible with Lyapunov stability, and incorporates naturally the idea of "nonlinear gain" functions; the reader may wish to consult [37] and [41] for expositions, as well as [44] for the proofs of several of the main characterizations. Roughly speaking, a system is ISS provided that, no matter what is the initial state, if the inputs are small, then the state must eventually be small. Dualizing this definition one arrives at the notion of detectability which is the main subject of study of this paper: *input-output-to-state stability* (IOSS). (The terminology "IOSS" is not to be confused with the totally different concept called input/output stability (IOS)—cf. [35, 46, 47]—which refers instead to stability of outputs, rather than to detectability.)

A system $\dot{x} = f(x, u)$ with measurement ("output") map $y = h(x)$ is IOSS if there are some functions $\beta \in \mathcal{KL}$ and $\gamma_1, \gamma_2 \in \mathcal{K}_\infty$ such that the estimate

$$|x(t)| \leq \max\left\{\beta(|x(0)|, t), \gamma_1\left(\left\|u|_{[0,t]}\right\|\right), \gamma_2\left(\left\|y|_{[0,t]}\right\|\right)\right\}$$

holds for any initial state $x(0)$ and any input $u(\cdot)$, where $x(\cdot)$ is the ensuing trajectory and $y(t) = h(x(t))$ the respective output function. (States $x(t)$, input values $u(t)$, and output values $y(t)$ lie in appropriate Euclidean spaces. We use $|\cdot|$ to denote Euclidean norm and $\|\cdot\|$ for supremum norm. Precise definitions and technical assumptions are discussed later.) The terminology IOSS is self-explanatory: formally, there is "stability from the input/output data to the state." The term was introduced in the paper [45], but the same notion had appeared before: it represents a natural combination of the notions of "strong" observability (cf. [35]) and ISS, and was called simply "detectability" in [36] (where it is phrased in input/output, as opposed to state space, terms and applied to questions of parameterization of controllers) and was called "strong unboundedness observability" in [19] (more precisely, this last notion also allows an additive nonnegative constant in the right-hand side of the estimate). In [45], two of the authors described relationships between the existence of full state observers and the IOSS property, or more precisely, a property which we called "incremental IOSS." The use of ISS-like formalism for studying observers, and hence implicitly the IOSS property, has also appeared several times in other authors' work, such as the papers [31, 26].

One of the main results of this paper is that a system is IOSS if and only if it admits a norm-estimator (in a sense also to be made precise). This result is in turn a consequence of a necessary and sufficient characterization of the IOSS property in terms of smooth dissipation functions, namely, there is a proper (radially unbounded) and positive definite smooth function V of states (a "storage function" in the language of dissipative systems introduced by Willems [52] and further developed by Hill and Moylan [12, 13] and others) such that a *dissipation inequality*

$$\frac{d}{dt}V(x(t)) \;\leq\; -\sigma_1(|x(t)|) + \sigma_2(|y(t)|) + \sigma_3(|u(t)|) \tag{1.1}$$

holds along all trajectories, with the functions σ_i of class $\mathcal{K}_\infty$. This provides an "infinitesimal" description of IOSS, and a norm-observer is easily built from V. Such a characterization in dissipation terms was conjectured in [45], and we provide here a complete solution to the problem. (The paper [45] also explains how the existence of V links the IOSS property to "passivity" of systems.)

It is worth pointing out that several authors have independently suggested that one should *define* "detectability" in dissipation terms. For example, in [27, eq. 15], one finds detectability defined by the requirement that there should exist a differentiable storage function V satisfying our dissipation inequality but with the special choice $\sigma_2(r) := r^2$ (there were no inputs in the class of systems considered there). A variation of this is to weaken the dissipation inequality, to merely require

$$x \neq 0 \;\Rightarrow\; \frac{d}{dt}V(x(t)) < \sigma_2(|y(t)|)$$

(again, with no inputs), as done, for instance, in the definition of detectability given in [30]. Observe that this represents a slight weakening of our property, in so far as there is no "margin" of stability $-\sigma_1(|x(t)|)$. One of our contributions is to show that such alternative definitions (when posed in the right generality) are in fact equivalent to IOSS.

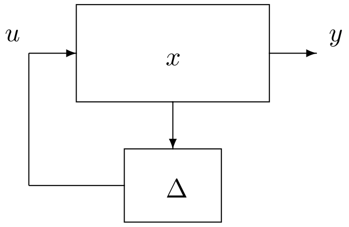


FIG. 1.3. *Robust detectability.*

A key preliminary step in the construction of V, just as it was for the analogous result for the ISS property obtained in [42], is the characterization of the IOSS property in robustness terms, by means of a "small gain" argument. The IOSS property is shown to be equivalent to the existence of a "robustness margin" $\rho \in \mathcal{K}_\infty$. This means that every system obtained by closing the loop with a feedback law Δ (even dynamic and/or time-varying) for which $|\Delta(t)| \leq \rho(|x(t)|)$ for all t (cf. Figure 1.3) is OSS (i.e., is IOSS as a system with no inputs). In order to formulate precisely this notion of robust detectability, we need to consider auxiliary "systems with disturbances." Since such systems must be introduced anyhow, we decided to present all our results (and definitions, even of IOSS) for systems with disturbances, in the process gaining extra generality in our results.

The core of the paper is, thus, the construction of V for "robustly detectable" (more precisely, "robust IOSS") systems $\dot{x} = g(x, d)$ which are obtained by substituting $u = d\rho(|x|)$ in the original system, and letting $d = d(\cdot)$ be an arbitrary measurable function taking values in a unit ball. The function V must satisfy a differential inequality of the form $\dot{V}(x(t)) \leq -\sigma_1(|x(t)|) + \sigma_2(|y(t)|)$ along all trajectories, that is to say, the following partial differential inequality:

$$\nabla V(x) \cdot g(x, d) \;\leq\; -\sigma_1(|x|) + \sigma_2(|y|)$$

for some functions σ_1 and σ_2 of class $\mathcal{K}_\infty$. But one last reduction consists of turning this problem into one of building Lyapunov functions for "relatively asymptotically stable" systems. Indeed, one observes that the main property needed for V is that it should *decrease along trajectories as long as $y(t)$ is sufficiently smaller than $x(t)$*. This leads us to the notion of "global asymptotic stability modulo outputs" and its Lyapunov-theoretic characterization.

The construction of V relies upon the solution of an appropriate optimal control problem, for which V is the value function. This problem is obtained by "fuzzifying" the dynamics near the set where $y \ll x$, so as to obtain a problem whose value function is continuous. Several elementary facts about relaxed controls are used in deriving the conclusions. The last major ingredient is the use of techniques from nonsmooth analysis, and in particular inf-convolutions, in order to obtain a Lipschitz, and from there by a standard regularization argument, a smooth, function V, starting from the continuous V that was obtained from the optimal control problem.

Finally, we will also discuss a version of detectability which relies upon "energy" estimates instead of uniform estimates. Such versions of detectability are fairly standard in control theory; see, for instance, [10], which defined "L^2-detectability" by a requirement that the state trajectory should be in L^2 if the observations are. The corresponding "integral to integral" notion uses a very interesting concept introduced in [29], that of "unboundedness observability" (UO), which amounts to a "relative (modulo outputs) forward completeness" property. It is shown that, for systems with no controls, the integral variant of OSS is equivalent to the conjunction of OSS and UO.

It is worth remarking that the main result in this paper amounts to providing necessary and sufficient conditions for the existence of a smooth (and proper and positive definite) solution V to a partial differential inequality which is equivalent to asking that (1.1) holds along all trajectories, namely,

$$\max_{u\in\mathbf{R}^m} \left\{\nabla V(x)\cdot f(x,u) + \sigma_1(|x|) - \sigma_2(|h(x)|) - \sigma_3(|u|)\right\} \;\leq\; 0\,. \tag{1.2}$$

It is a consequence of our results that if there is an (even just) lower semicontinuous such solution (when "solution" is interpreted in a weak sense, for example, in terms of viscosity or proximal subdifferentials), then there is also a smooth solution (usually, however, with different comparison functions σ_i's). This is because the existence of a weak solution is already equivalent to IOSS, as shown in [21]. It is a routine observation that the above partial differential inequality can be posed in an equivalent way as a *Hamilton–Jacobi inequality (HJI)*, in the special case of quadratic input "cost" $\sigma_3(r) = r^2$ and for systems $\dot{x} = f(x,u)$ which are affine in controls, i.e., systems of the form

$$\dot{x} \;=\; g_0(x) \;+\; \sum_{i=1}^{m} u_i\, g_i(x) \tag{1.3}$$

(we are denoting by u_i the ith component of u). Indeed, one need only replace the expression in (1.2) by its maximum value obtained at $u_i = (1/2)\nabla V(x)\cdot g_i(x)$, $i = 1,\dots,m$, thereby obtaining the following HJI:

$$\nabla V(x)\cdot g_0(x) + \frac{1}{4}\sum_{i=1}^{m}(\nabla V(x)\cdot g_i(x))^2 + \sigma_1(|x|) - \sigma_2(|h(x)|) \;\leq\; 0\,. \tag{1.4}$$

2. Definitions and statements of the main results.

2.1. Systems of interest.

We study a system whose dynamics depend on two types of inputs, which we respectively call *controls* and *disturbances*:

$$\dot{x}(t) = f(x(t),\mathbf{u}(t),\mathbf{w}(t)), \quad y(t) = h(x(t)). \tag{2.1}$$

Here, states evolve in $\mathbf{X} = \mathbf{R}^n$, controls are measurable, essentially bounded functions $\mathbf{u}$ on $\mathcal{I} = \mathbf{R}_{\geq 0}$ with values in $\mathbb{U} := \mathbf{R}^{m_u}$, and disturbances are measurable functions

$\mathbf{w} : \mathcal{I} \to \Gamma$ with values in Γ, which is a compact, convex subset of $\mathbf{R}^{m_w}$. We will denote the set of all such functions by $\mathcal{M}_\Gamma$. In those cases when a different interval $\mathcal{I} \subset \mathbf{R}_{\geq 0}$ of definition for a control $\mathbf{u}$ is specified, we always apply the definitions to the extension of $\mathbf{u}$ to $\mathbf{R}_{\geq 0}$, using $\mathbf{u} \equiv 0$ on $\mathbf{R}_{\geq 0} \setminus \mathcal{I}$. The function $f : \mathbf{X} \times \mathbb{U} \times \Gamma \to \mathbf{X}$ is locally Lipschitz in (x, u) uniformly on w, jointly continuous in x, u, and w, and such that $f(0, 0, w) = 0$ for any $w \in \Gamma$; and $h : \mathbf{X} \to \mathcal{Y} := \mathbf{R}^p$ is smooth (C^1) and vanishes at 0.

A function $\alpha : \mathbf{R}_{\geq 0} \to \mathbf{R}_{\geq 0}$ is *of class* $\mathcal{K}$ if it is continuous, positive definite, and strictly increasing, and is *of class* $\mathcal{K}_\infty$ if it is also unbounded. A function $\beta : \mathbf{R}_{\geq 0} \times \mathbf{R}_{\geq 0} \to \mathbf{R}_{\geq 0}$ is said to be *of class* $\mathcal{KL}$ if for each fixed $t \geq 0$, $\beta(\cdot, t)$ is of class $\mathcal{K}$, and for each fixed $s \geq 0$, $\beta(s, t)$ decreases to 0 as $t \to \infty$. Let $z(\cdot)$ be a measurable function.

The L_∞ (essential supremum) norm of the restriction of z to the interval $[t_1, t_2]$ is denoted by $\left\| z|_{[t_1,t_2]} \right\|$.

Given a state $\xi \in \mathbf{X}$, for each pair $(\mathbf{u}, \mathbf{w})$ denote by $x(t, \xi, \mathbf{u}, \mathbf{w})$ the unique maximal solution of the system (2.1), which is defined on some maximal interval $[0, t_{\max}(\xi, \mathbf{u}, \mathbf{w}))$. We will use the notation $y(t, \xi, \mathbf{u}, \mathbf{w}) := h(x(t, \xi, \mathbf{u}, \mathbf{w}))$, and, when unimportant or clear from the context, we will write $t_{\max}$ instead of $t_{\max}(\xi, \mathbf{u}, \mathbf{w})$, $x(t)$ instead of $x(t, \xi, \mathbf{u}, \mathbf{w})$, and $y(t)$ instead of $y(t, \xi, \mathbf{u}, \mathbf{w})$.

2.2. Notions of "uniform detectability" and dissipation functions.

DEFINITION 2.1. *A system of type* (2.1) *is said to be* uniformly input-output-to-state stable (UIOSS) *if there exist functions $\beta \in \mathcal{KL}$ and $\gamma_1, \gamma_2 \in \mathcal{K}$ such that the estimate*

$$|x(t, \xi, \mathbf{u}, \mathbf{w})| \leq \max\left\{\beta(|\xi|, t), \gamma_1\left(\left\|\mathbf{u}|_{[0,t]}\right\|\right), \gamma_2\left(\left\|y|_{[0,t]}\right\|\right)\right\} \tag{2.2}$$

holds for any initial state $\xi \in \mathbf{X}$, control $\mathbf{u}$, disturbance $\mathbf{w}$, and $t \in [0, t_{\max}(\xi, \mathbf{u}, \mathbf{w}))$.

DEFINITION 2.2. *A smooth (C^∞) function $V : \mathbf{X} \to \mathbf{R}_{\geq 0}$ is a UIOSS-Lyapunov function for system* (2.1) *if*

- *there exist $\mathcal{K}_\infty$-functions α_1, α_2 such that*

$$\alpha_1(|\xi|) \leq V(\xi) \leq \alpha_2(|\xi|) \tag{2.3}$$

  holds for all ξ in $\mathbf{X}$, and
- *there exists a $\mathcal{K}_\infty$-function α and $\mathcal{K}$-functions σ_1, σ_2 such that*

$$\nabla V(\xi) \cdot f(\xi, u, w) \leq -\alpha(|\xi|) + \sigma_1(|u|) + \sigma_2(|h(\xi)|) \tag{2.4}$$

  for all ξ in $\mathbf{X}$, for all control values $u \in \mathbb{U}$, and for all disturbance values $w \in \Gamma$. □

Property (2.3) amounts to positive definiteness and properness of V; requiring the existence of an upper bound α_2 is redundant, as it follows from the fact that V is continuous and satisfies $V(0) = 0$. However, it is convenient to specify this bound explicitly, as it will be used in various estimates. Condition (2.4) is a *dissipation inequality* in the sense of [52].

Remark 2.1. A smooth function $V : \mathbf{X} \to \mathbf{R}_{\geq 0}$, satisfying (2.3) on $\mathbf{X}$ with some α_1, α_2 of class $\mathcal{K}_\infty$, is a UIOSS-Lyapunov function for a system (2.1) if and only if there exist functions α_3 of class $\mathcal{K}_\infty$, and γ and χ_1 of class $\mathcal{K}$ such that

$$\nabla V(\xi) \cdot f(\xi, u, w) \leq -\alpha_3(|\xi|) + \gamma(|h(\xi)|) \tag{2.5}$$

for any $\xi \in \mathbf{X}$, $w \in \Gamma$, and $u \in \mathbb{U}$ such that $|\xi| \geq \chi_1(|u|)$.

Indeed, clearly (2.4) implies (2.5) with $\alpha_3(\cdot) := \alpha(\cdot)/2$, $\gamma \equiv \sigma_2$, and $\chi_1 := \alpha^{-1} \circ (2\sigma_1)$. To prove the other implication, assume now that (2.5) holds with some $\alpha_3 \in \mathcal{K}_\infty$ and γ, $\chi_1 \in \mathcal{K}$. Define $\sigma_1(\cdot) = \max\{0, \hat{\sigma_1}(\cdot)\}$, where

$$\hat{\sigma_1}(r) := \max\left\{\nabla V(\xi) \cdot f(\xi, u, w) + \alpha_3(\chi_1(|u|)) : |u| \leq r,\ |\xi| \leq \chi_1(r), w \in \Gamma\right\}.$$

Then σ_1 is continuous, $\sigma_1(0) = 0$, and one can assume that σ_1 is a $\mathcal{K}_\infty$-function (majorize it by one if it is not). We claim that (2.4) holds with $\alpha \equiv \alpha_3$ and $\sigma_2 \equiv \gamma$. Indeed, if $|\xi| \geq \chi_1(|u|)$, then (2.5) holds, from which (2.4) trivially follows. If $|\xi| < \chi_1(|u|)$, then, by definition of σ_1,

$$\sigma_1(|u|) \geq \nabla V(\xi) \cdot f(\xi, u, w) + \alpha_3(\chi_1(|u|))$$

for every w, which, in turn, implies (2.4).

A few particular cases of the UIOSS property have been studied in the literature. If the system (2.1) in consideration has no outputs and no disturbances, UIOSS reduces to the well-known ISS property, whose Lyapunov characterization was obtained in [42]. In case (2.1) is autonomous, UIOSS becomes OSS. This property was introduced in [43] where Lyapunov-type necessary and sufficient conditions were obtained. Finally, for systems with no disturbances, UIOSS is just IOSS. This property was introduced in [45], where it was conjectured that any IOSS control system admits a smooth IOSS-Lyapunov function. This conjecture will be proven here in a more general setting, for systems forced by both controls and disturbances. A few interesting applications of this Lyapunov characterization were also discussed in [45], one of them to be defined next.

2.3. Norm-estimators.

DEFINITION 2.3. *A* state-norm-estimator *(or* state-norm-observer*) for a system Σ of type* (2.1) *is a pair $(\Sigma_{n.o}, k(\cdot,\cdot))$, where $k : \mathbf{R}^\ell \times \mathcal{Y} \to \mathbf{R}$, and $\Sigma_{n.o}$ is a system*

$$\dot{p} = g(p, u, y) \tag{2.6}$$

evolving in $\mathbf{R}^\ell$ and driven by the controls and outputs of Σ, such that the following conditions are satisfied.

- *There exist $\mathcal{K}$-functions $\hat{\gamma}_1$ and $\hat{\gamma}_2$ and a $\mathcal{KL}$-function $\hat{\beta}$ such that for any initial state $\zeta \in \mathbf{R}^\ell$, all inputs $\mathbf{u}$ and $\mathbf{y}$, and any t in the interval of definition of the solution $p(\cdot, \zeta, \mathbf{u}, \mathbf{y})$, the following inequality holds:*

  $$|k(p(t, \zeta, \mathbf{u}, \mathbf{y}), \mathbf{y}(t))| \leq \hat{\beta}(|\zeta|, t) + \hat{\gamma}_1\left(\left\|\mathbf{u}|_{[0,t]}\right\|\right) + \hat{\gamma}_2\left(\left\|\mathbf{y}|_{[0,t]}\right\|\right) \tag{2.7}$$

  (in other words, the system (2.6) *is IOS with respect to the inputs u and y and output k).*
- *There are functions $\rho \in \mathcal{K}$ and $\beta \in \mathcal{KL}$ so that, for any pair of initial states ξ and ζ of systems* (2.1) *and* (2.6)*, respectively, any control $\mathbf{u} : \mathbf{R}_{\geq 0} \to \mathbb{U}$ and any disturbance $\mathbf{w} \in \mathcal{M}_\Gamma$, we have*

  $$|x(t, \xi, \mathbf{u}, \mathbf{w})| \leq \beta(|\xi| + |\zeta|, t) + \rho(|k(p(t, \zeta, \mathbf{u}, \mathbf{y}_{\xi,\mathbf{u},\mathbf{w}}), \mathbf{y}_{\xi,\mathbf{u},\mathbf{w}}(t))|) \tag{2.8}$$

  for all $t \in [0, t_{\max}(\xi, \mathbf{u}, \mathbf{w}))$. (Here $\mathbf{y}_{\xi,\mathbf{u},\mathbf{w}}$ denotes the output trajectory of Σ, that is, $\mathbf{y}_{\xi,\mathbf{u},\mathbf{w}}(t) = y(t, \xi, \mathbf{u}, \mathbf{w})$.)

2.4. Statement of the main result. The main theorem to be proved in this paper, summarizing the equivalent characterizations of UIOSS, will be as follows.

THEOREM 2.4. *Let* Σ *be a system of type* (2.1). *Then the following are equivalent:*

1. Σ *is UIOSS.*
2. Σ *admits a UIOSS-Lyapunov function.*
3. *There is a state-norm-estimator for* Σ.

The main contribution is in showing that 1 implies 2; the remaining implications are much easier.

2.5. Example: Linear systems. A particular class of systems (2.1) is as follows. A *linear*, time-invariant system Σ_{lin} with outputs is one for which f and h are linear, that is,

$$\begin{aligned} \dot{x} &= \mathbf{A}x + \mathbf{B}u, \\ y &= \mathbf{C}x, \end{aligned} \tag{2.9}$$

where $\mathbf{A} \in \mathbf{R}^{n\times n}$, $\mathbf{B} \in \mathbf{R}^{n\times m}$, and $\mathbf{C} \in \mathbf{R}^{p\times n}$. (We assume that $m_w = 0$.)

Recall the following definition.

DEFINITION 2.5. *A linear, time-invariant system* (2.9) *is* detectable*, or* asymptotically observable*, if the implication*

$$\mathbf{C}x(t) \equiv 0 \implies x(t) \to 0 \tag{2.10}$$

holds for any trajectory $x(t)$ *of* (2.9)*, corresponding to the zero control* $\mathbf{u} \equiv 0$.

The following result is a totally routine linear systems theory fact, but we include the proof as a motivation for the nonlinear material to follow.

PROPOSITION 2.6. *If a linear system* (2.9) *is detectable, then it is IOSS.*

Proof. It is a well-known fact (see, for example, [39]) that if a system (2.9) is detectable, then there exists a matrix $\mathbf{L} \in \mathbf{R}^{n\times p}$, such that the matrix $\mathbf{A} + \mathbf{LC}$ is Hurwitz, and, furthermore, the system

$$\dot{z}(t) = \mathbf{A}z(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{L}(\mathbf{C}z(t) - y(t)), \tag{2.11}$$

referred to as an *observer* and driven by the controls and outputs of (2.9), has the property that if $x(t)$ and $z(t)$ are any solutions of (2.9) and (2.11), respectively, then $|x(t) - z(t)| \to 0$, and, in particular, if $x(0) = z(0)$, then $x(t) = z(t)$ for all nonnegative t. Fix an initial state ξ and a control $\mathbf{u}$. Then the solution $x(t, \xi, \mathbf{u})$ of (2.9) is also the solution of (2.11) with $z(0) = \xi$, so that

$$x(t, \xi, \mathbf{u}) = e^{t(\mathbf{A}+\mathbf{LC})}\xi + \int_0^t e^{s(\mathbf{A}+\mathbf{LC})}[\mathbf{B}\mathbf{u}(t-s) - \mathbf{L}y(t-s)]ds.$$

Choose two positive numbers δ' and δ so that $\Re\lambda \leq -\delta' < -\delta$ for every eigenvalue λ of $\mathbf{A} + \mathbf{LC}$. Then there exists a polynomial $P(\cdot)$ and, consequently, a constant K such that

$$\begin{aligned} |x(t, \xi, \mathbf{u})| &\leq P(t)e^{-\delta' t}|\xi| + \int_0^t P(s)e^{-\delta' s}\left[\|\mathbf{B}\|\,|\mathbf{u}(t-s)| + \|\mathbf{L}\|\,|y(t-s)|\right]ds \\ &\leq Ke^{-\delta t}|\xi| + K\frac{\|\mathbf{B}\|}{\delta}\left\|\mathbf{u}|_{[0,t]}\right\| + K\frac{\|\mathbf{L}\|}{\delta}\left\|y|_{[0,t]}\right\|. \end{aligned} \tag{2.12}$$

Thus, the IOSS estimate (2.2) holds for (2.9) with the linear gains $\beta(r, t) := Ke^{-\delta t}r$ and $\gamma_1(r) = \gamma_2(r) := K\frac{\|\mathbf{B}\|}{\delta}r$. □

To see this, first note that the solution $z(\cdot,\zeta,\mathbf{u},\mathbf{y})$ of (2.11) satisfies the estimate

$$|z(t,\zeta,\mathbf{u},\mathbf{y})| \leq Ke^{-\delta t}\,|\xi| + K\frac{\|\mathbf{B}\|}{\delta}\left\|\mathbf{u}|_{[0,t]}\right\| + K\frac{\|\mathbf{L}\|}{\delta}\left\|y|_{[0,t]}\right\|,$$

where K,δ are the same as in (2.12). Also, with $e := x - z$, we have

$$\dot{e}(t) = (A + LC)e(t).$$

From this we see that, along any trajectory $(x(t), z(t))$ of (2.9) and (2.11),

$$|x(t)| \leq K\,|\xi - \zeta|\,e^{-\delta t} + |z(t)|$$

for all $t \geq 0$, where K,δ are again as in (2.12).

To find an IOSS-Lyapunov function for system (2.9), take any symmetric matrix $\mathbf{P} \in \mathbf{R}^{n\times n}$ such that

$$\mathbf{P}(\mathbf{A}+\mathbf{LC}) + (\mathbf{A}+\mathbf{LC})'\mathbf{P} = -I$$

(such a matrix $\mathbf{P}$ exists, because $\mathbf{A}+\mathbf{LC}$ is Hurwitz). Define

$$V(x) := x'\mathbf{P}x. \tag{2.13}$$

Notice that, since $\mathbf{P}(\mathbf{A}+\mathbf{LC}) = ((\mathbf{A}+\mathbf{LC})'\mathbf{P})'$, we have $x'\mathbf{P}(\mathbf{A}+\mathbf{LC})x = x'(\mathbf{A}+\mathbf{LC})'\mathbf{P}x$. Therefore

$$\begin{aligned}
\nabla V(x)\cdot f(x,u) &= 2x'\mathbf{P}\,(\mathbf{A}x + \mathbf{B}u)\\
&= 2x'\mathbf{P}\,((\mathbf{A}+\mathbf{LC})x + \mathbf{B}u - \mathbf{L}y)\\
&= 2x'\mathbf{P}(\mathbf{A}+\mathbf{LC})x + 2x'\mathbf{PB}u - 2x'\mathbf{PL}y\\
&= x'(\mathbf{P}(\mathbf{A}+\mathbf{LC}) + (\mathbf{A}+\mathbf{LC})'\mathbf{P})x + 2x'\mathbf{PB}u - 2x'\mathbf{PL}y\\
&\leq -|x|^2 + 2\,\|\mathbf{P}\|\,\|\mathbf{B}\|\,|u|\,|x| + 2\,\|\mathbf{P}\|\,\|\mathbf{L}\|\,|y|\,|x|\\
&\leq -|x|^2 + |x|^2/4 + 4\,\|\mathbf{P}\|^2\,\|\mathbf{B}\|^2\,|u|^2 + |x|^2/4 + 4\,\|\mathbf{P}\|^2\,\|\mathbf{L}\|^2\,|y|^2\\
&\leq -|x|^2/2 + 4\,\|\mathbf{P}\|^2\,\|\mathbf{B}\|^2\,|u|^2 + 4\,\|\mathbf{P}\|^2\,\|\mathbf{L}\|^2\,|y|^2.
\end{aligned}$$

So, the UIOSS dissipation inequality holds for V with gains defined by $\alpha(r) = r/2$, $\sigma_1(r) = 4\,\|\mathbf{P}\|^2\,\|\mathbf{B}\|^2\,r^2$, $\sigma_2(r) = 4\,\|\mathbf{P}\|^2\,\|\mathbf{L}\|^2\,r^2$.

2.6. Systems without controls. Let Ω be a compact metric space (which is always assumed to be of the form $[-1,1]^m$ unless specified otherwise). Consider systems of the type

$$\dot{x} = f(x(t),\mathbf{d}(t)), \quad y(t) = h(x(t)), \tag{2.14}$$

where $f : \mathbf{X}\times\Omega$ is locally Lipschitz in x uniformly on d and jointly continuous in x and d, and $f(0,d) = 0$ for any $d \in \Omega$. The inputs are measurable functions $\mathbf{d} : \mathcal{I} \to \Omega$, and we use the term *disturbances* to refer to such Ω-valued inputs. We will use $\mathcal{M}_\Omega$ to denote the collection of all such functions.

This system can be seen as a particular case of (2.1) that eschews controls and is driven only by disturbances. However, it will play an important role in our studies; therefore for convenience we will define the corresponding stability property and dissipation inequality for this system separately from the main Definition 2.1.

Definition 2.7. *A system* (2.14) *is UOSS (uniformly output-to-state stable) if there exist some $\beta \in \mathcal{KL}$ and $\gamma_2 \in \mathcal{K}$ such that*

$$|x(t,\xi,\mathbf{d})| \leq \max\left\{\beta(|\xi|,t), \gamma_2\left(\left\|y|_{[0,t]}\right\|\right)\right\} \tag{2.15}$$

for any disturbance $\mathbf{d}$*, initial state* $\xi \in \mathbf{X}$*, and* $t \in [0, t_{\max})$.

Definition 2.8. *A UOSS-Lyapunov function for system* (2.14) *is a smooth function* $V : \mathbf{X} \to \mathbf{R}_{\geq 0}$ *satisfying* (2.3) *and*

$$\nabla V(x) \cdot f(x,d) \leq -\alpha_3(|x|) + \gamma(|h(x)|) \quad \forall\, x \in \mathbf{X},\ \forall\, d \in \Omega, \tag{2.16}$$

with some class $\mathcal{K}_\infty$ functions α_i and a $\mathcal{K}$ function γ. For systems with no disturbances we simply say that V is an OSS-Lyapunov function.

2.6.1. "Modulo outputs" relative stability.

Recall the classical notion of uniform global asymptotic stability for systems of type (2.14), ensuring that every solution of the system tends to the equilibrium and never goes too far from it. Suppose now that it does not matter how the system behaves when the information provided by the output is adequate, that is, the norm of the output dominates the norm of the current state. On the other hand, we want the system to decay nicely when the output does not help in determining how large the state is. This motivates the following "modulo output" definition of stability.

Definition 2.9. *A system of type* (2.14) *satisfies the GASMO (global asymptotic stability modulo output) property if there exist a function ρ of class $\mathcal{K}_\infty$ and a function λ of class $\mathcal{KL}$ such that, for all $\xi \in \mathbf{X}$, $\mathbf{d} \in \mathcal{M}_\Omega$, and any $T < t_{\max}(\xi,\mathbf{d})$, if*

$$|x(t,\xi,\mathbf{d})| \geq \rho(|h(x(t,\xi,\mathbf{d}))|) \quad \forall\, 0 \leq t \leq T,$$

then the estimate

$$|x(t,\xi,\mathbf{d})| \leq \lambda(|\xi|,t) \quad \forall\, 0 \leq t \leq T \tag{2.17}$$

holds.

Remark 2.2. If a system in consideration has no outputs, then the GASMO property becomes global asymptotic stability (GAS).

The following proposition provides an "ε-δ" characterization of the GASMO property.

Proposition 2.10. *A system of type* (2.14) *satisfies the GASMO property if and only if there exists a $\mathcal{K}_\infty$-function ρ so that the following two properties hold.*

1. *For any $\varepsilon > 0$ and any $r > 0$, there exists some $T_{r,\varepsilon}$ such that for any $|\xi| \leq r$, any $\mathbf{d}$, and any $T \in [0, t_{\max}(\xi,\mathbf{d}))$ such that $T \geq T_{r,\varepsilon}$, if*

$$|x(t,\xi,\mathbf{d})| \ \geq\ \rho(|y(t,\ \xi,\mathbf{d})|)\ \forall\, 0 \leq t \leq T\,,$$

then

$$|x(t,\ \xi,\mathbf{d})|\ <\ \varepsilon\ \ \forall\, t \in [T_{r,\varepsilon}, T]\,.$$

2. *There exists a $\mathcal{K}$-function ϑ such that for any $\xi \in \mathbf{X}$, any disturbance $\mathbf{d}$, and any $T < t_{\max}(\xi,\mathbf{d})$ such that*

$$|x(t,\xi,\mathbf{d})|\ \geq\ \rho(|y(t,\xi,\mathbf{d})|)\ \forall\, 0 \leq t \leq T\,,$$

the following "bounded overshoot" estimate holds:

$$|x(T,\xi,\mathbf{d})|\ \leq\ \vartheta(|\xi|)\,.$$

The necessity part is obvious. To prove the sufficiency, we need the following lemma, proved in section 3 of [25], although not explicitly stated in this form.

LEMMA 2.11. *Let* $\Phi(r,t) : (\mathbf{R}_{\geq 0})^2 \to \mathbf{R}_{\geq 0}$ *be a map such that*

1. *for all* $\varepsilon > 0$ *and for all* $R > 0$ *there exists* T *such that* $\Phi(r,t) < \varepsilon$ *for all* $0 \leq r \leq R$ *and for all* $t \geq T$,
2. *for all* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that if* $r \leq \delta$, *then* $\Phi(r,t) < \varepsilon$ *for all* $t > 0$.

Then Φ *can be majorized by a* $\mathcal{KL}$*-function.*

Proof of sufficiency for Proposition 2.10. Consider the function

$$\Phi(r,t) := \sup\left\{|x(t,\xi,\mathbf{d})| : \ |\xi| \leq r,\, \mathbf{d} \in \mathcal{M}_\Omega,\ |x(s,\xi,\mathbf{d})| \geq \rho(|y(s,\xi,\mathbf{d})|)\ \forall\, s \in [0,t]\right\}.$$

Then the conditions 1 and 2 of Lemma 2.11 follow from the assumptions 1 and 2 of the proposition, so that one can majorize Φ by a $\mathcal{KL}$-function λ. □

2.6.2. Integral variants. The UOSS property gives uniform estimates on states as a function of uniform bounds on outputs. There is a "finite energy output implies finite energy state" version as well.

DEFINITION 2.12. *A system of type* (2.14) *is* integral-to-integral uniformly output-to-state stable (iiUOSS) *if there exist functions* γ, κ *of class* $\mathcal{K}$, *and* $\chi \in \mathcal{K}_\infty$ *such that*

$$\int_0^t \chi(|x(s,\xi,\mathbf{d})|)\, ds \ \leq\ \kappa(|\xi|) + \int_0^t \gamma(|h(x(s,\xi,\mathbf{d}))|)\, ds \tag{2.18}$$

for any initial state ξ, *any disturbance* $\mathbf{d} \in \mathcal{M}_\Omega$, *and any time* $t \in [0, t_{\max}(\xi,\mathbf{d}))$.

Without loss of generality, γ and κ can be assumed to be of class $\mathcal{K}_\infty$.

DEFINITION 2.13. *A system* (2.1) *is called* forward complete *if for every initial condition* ξ, *every input signal* $\mathbf{u}$, *and every disturbance* $\mathbf{d}$ *defined on* $[0,+\infty)$, *the corresponding trajectory* $x(t,\xi,\mathbf{u},\mathbf{d})$ *is defined for all* $t \geq 0$, *i.e.,* $t_{\max}(\xi,\mathbf{u},\mathbf{d}) = +\infty$.

The following property, which is strictly weaker than forward completeness, was introduced in [29].

DEFINITION 2.14. *A system* (2.14) *has the* unboundedness observability *property (UO) if*

$$\limsup_{t \nearrow t_{\max}(\xi,\mathbf{d})} |y(t,\xi,\mathbf{d})| = +\infty \tag{2.19}$$

holds for each initial state ξ *and disturbance* $\mathbf{d}$ *with* $t_{\max}(\xi,\mathbf{d}) < \infty$.

The following useful characterization of UO was provided in [3].

PROPOSITION 2.15. *A system* (2.14) *has the UO property if and only if there exist class* $\mathcal{K}$ *functions* ρ_1, χ_1, χ_2, *and a constant* c, *such that the following implication holds:*

$$\begin{aligned} &|h(x(t,\xi,\mathbf{d}))| \leq \rho_1(|x(t,\xi,\mathbf{d})|) \quad \forall\, t \in [0,T] \\ &\Rightarrow |x(t,\xi,\mathbf{d})| \leq \chi_1(t) + \chi_2(|\xi|) + c \quad \forall t \in [0,T] \end{aligned} \tag{2.20}$$

for all $\xi \in \mathbf{X}$, $\mathbf{d} \in \mathcal{M}_\Omega$, *and all* $T \in [0, t_{\max}(\xi,\mathbf{d}))$.

This proposition provides a uniform bound on all the states that can be reached by a UO system in given time from a given bounded set via a trajectory *not dominated by the output.* Notice that *for systems with disturbances* (2.14), *the UOSS property implies the UO property.*

2.6.3. Statement of the main result for the case of no controls.

THEOREM 2.16. *Let Σ be a system of type* (2.14). *Then the following are equivalent:*

1. Σ *is UOSS.*
2. Σ *is GASMO.*
3. Σ *is iiUOSS and UO.*
4. Σ *admits a UOSS-Lyapunov function.*

2.7. Organization of the paper. Implications 2 ⇒ 3 ⇒ 1 of Theorem 2.4 are proven in section 5. The part of Theorem 2.4 most difficult to prove is the implication 1 ⇒ 2. The main technical result needed for this proof is implication 2 ⇒ 4 of Theorem 2.16. This is proven in section 4. The construction of a UIOSS-Lyapunov function for an original system (2.1) is reduced, via a small gain argument, to the construction of a UOSS-Lyapunov function for a special system (2.14) related to the original system (2.1). This reduction is done in section 3.1, and section 3.2 completes the construction of UIOSS-Lyapunov functions.

Finally, implications 3 ⇒ 2, 1 ⇒ 2, and 4 ⇒ 3 of Theorem 2.16 are proven in section 3.3, and 4 ⇒ 1 follows from Theorem 2.4.

3. Reduction to the case of no controls.

In this part we show how to reduce our main result to the particular case of systems with no controls. Let $\mathbb{U}_1$ denote a closed unit ball $\{u \in \mathbb{U} : |u| \leq 1\}$ in $\mathbb{U}$.

3.1. Robust output-to-state stability.

DEFINITION 3.1. *System* (2.1) *is said to be* robustly output-to-state stable (ROSS) *if there exists a locally Lipschitz $\mathcal{K}_\infty$-function φ, called a* stability margin, *such that the system*

$$\dot{x}(t) = g(x(t), \mathbf{d}(t)) := f(x(t), \mathbf{d}_u(t)\varphi(|x(t)|), \mathbf{w}(t)) \tag{3.1}$$

with disturbances $d := [d_u, w] \in \mathbb{U}_1 \times \Gamma$ and outputs $y = h(x)$ is UOSS.

Notice that the set $\mathbb{U}_1 \times \Gamma$ is a compact, convex subset of $\mathbf{R}^{m_u+m_w}$. We will denote it by Ω in this section. Observe also that the dynamics g of system (3.1) are locally Lipschitz in x uniformly in d, and also $g(0, d) = 0$ for all $d \in \Omega$.

LEMMA 3.2. *If a system* (2.1) *is UIOSS, then it is ROSS.*

The proof will follow from a few preliminary lemmas.

Let $\beta \in \mathcal{KL}$ and γ_1, $\gamma_2 \in \mathcal{K}_\infty$ be as in (2.2). Let $\vartheta(r) = \beta(r, 0)$. Without loss of generality, we may assume that ϑ is $\mathcal{K}_\infty$ and $\vartheta(r) \geq r$ (so that $\vartheta^{-1}(r) \leq r$).

Define $\varphi(r)$ to be a locally Lipschitz $\mathcal{K}_\infty$-function, which minorizes ${\gamma_1}^{-1}(\frac{1}{4}\vartheta^{-1}(r))$ and can be extended as a Lipschitz function to a neighborhood of $[0, \infty)$. To prove the lemma we will show that φ is a stability margin for (2.1).

PROPOSITION 3.3. *Fix a $\xi \in \mathbf{X}$, a control $\mathbf{u}$, and a disturbance $\mathbf{w}$, and let $x(\cdot) := x(\cdot, \xi, \mathbf{u}, \mathbf{w})$ be the corresponding solution of the system* (2.1). *Let $T \in [0, t_{\max}(\xi, \mathbf{u}, \mathbf{w}))$. Then if $|\mathbf{u}(t)| \leq \varphi(|x(t)|)$ for almost all $t \in [0, T]$, the estimate*

$$|x(t)| \leq \max\left\{\beta(|\xi|, t),\ \gamma_2(\|y|_{[0,t]}\|),\ \frac{|\xi|}{4}\right\} \tag{3.2}$$

holds for all $t \in [0, T]$.

Proof.

Claim 1. Suppose $T < t_{\max}(\xi, \mathbf{u}, \mathbf{w})$. If

$$|\mathbf{u}(t)| \leq \varphi(|x(t)|) \text{ for almost all } t \in [0, T), \tag{3.3}$$

then, for all $t \in [0, T)$,

$$|x(t)| \le \max\left\{\vartheta(|\xi|), 2\gamma_2(\left\|y|_{[0,t]}\right\|)\right\}. \tag{3.4}$$

Proof of the claim. Suppose first that $\xi = 0$. In this case $x(t) \equiv 0$. Indeed, define $\mathbf{d}_u(\cdot)$ on $[0, t_{\max}(\xi, \mathbf{u}, \mathbf{w}))$ by

$$\mathbf{d}_u(t) = \begin{cases} 0 & \text{if } x(t) = 0, \\ \mathbf{u}(t)/\varphi(|x(t)|) & \text{if } x(t) \ne 0. \end{cases}$$

Then (3.3) implies that $\mathbf{d}_u \in \mathcal{M}_{\mathbb{U}_1}$ and that $x(\cdot)$ is the solution of (3.1) with $\xi = 0$, $\mathbf{w}$ as we picked, and $\mathbf{d}_u(\cdot)$ as we defined. Noticing that the constant function equal to 0 is also a solution of this system, with the same initial state and the same disturbance, we conclude by uniqueness of solutions that $x(t) = 0$ for all nonnegative t, so that (3.4) trivially holds.

Suppose now that $\xi \ne 0$. Fix ε, such that $1 < \varepsilon < 2$. We will first show that if $|\mathbf{u}(t)| \le \varphi(|x(t)|)$ for almost all $t < T$, then the estimate

$$|x(t)| \le \max\left\{\varepsilon\vartheta(|\xi|), 2\gamma_2(\left\|y|_{[0,t]}\right\|)\right\} \tag{3.5}$$

holds for all $t \in [0, T)$. Indeed, notice that (3.5) is true as a strict inequality at $t = 0$ because $|\xi| \le \vartheta(|\xi|) < \varepsilon\vartheta(|\xi|)$. If (3.5) fails at some $t \in [0, T)$, then there exists a

$$t_0 = \min\left\{t < T \ : \ x(t) = \max\left\{\varepsilon\vartheta(|\xi|), 2\gamma_2(\left\|y|_{[0,t]}\right\|)\right\}\right\}.$$

Note that $t_0 > 0$ because at $t = 0$ we have a strict inequality in (3.5). So, (3.5) holds for all $t \in [0, t_0)$ and $x(t_0) = \max\left\{\varepsilon\vartheta(|\xi|), 2\gamma_2(\left\|y|_{[0,t_0]}\right\|)\right\}$. Therefore for almost all $t \in [0, t_0)$ we have

$$\begin{aligned}
&\gamma_1(\left\|\mathbf{u}|_{[0,t]}\right\|) \le \gamma_1(\left\|\varphi(|x(\cdot)|)|_{[0,t]}\right\|) \\
&\le \max\left\{\frac{1}{4}\vartheta^{-1}(\varepsilon\vartheta(|\xi|)), \frac{1}{4}\vartheta^{-1}(2\gamma_2(\left\|y|_{[0,t]}\right\|))\right\} \\
&\le \max\left\{\frac{1}{4}\varepsilon\vartheta(|\xi|), \frac{1}{4}2\gamma_2(\left\|y|_{[0,t]}\right\|)\right\} \\
&\le \max\left\{\vartheta(|\xi|), \gamma_2(\left\|y|_{[0,t]}\right\|)\right\}.
\end{aligned}$$

Then, since our system is UIOSS and $x(\cdot)$ is continuous, for all $t \in [0, t_0]$ we have

$$\begin{aligned}
|x(t)| &\le \max\left\{\vartheta(|\xi|), \gamma_1(\left\|\mathbf{u}|_{[0,t]}\right\|), \gamma_2(\left\|y|_{[0,t]}\right\|)\right\} \\
&= \max\left\{\vartheta(|\xi|), \gamma_2(\left\|y|_{[0,t]}\right\|)\right\}.
\end{aligned}$$

On the other hand, $|x(t_0)| = \max\left\{\varepsilon\vartheta(|\xi|), 2\gamma_2(\left\|y|_{[0,t_0]}\right\|)\right\}$ by definition of t_0. The contradiction proves the estimate (3.5). Letting ε tend to 1, we conclude that estimate (3.4) holds for all $t \in [0, T)$, completing the proof of Claim 1.

Hence, under the assumption of Claim 1, we have

$$\begin{aligned}
\gamma_1\left(\left\|\mathbf{u}|_{[0,t]}\right\|\right) &\le \gamma_1\left(\left\|\varphi(|x(\cdot)|)|_{[0,t]}\right\|\right) \\
&\le \max\left\{\frac{|\xi|}{4}, \frac{1}{4}\vartheta^{-1}\left(2\gamma_2\left(\left\|y|_{[0,t]}\right\|\right)\right)\right\}
\end{aligned}$$

for all t in $[0, t_{\max})$. So,

$$|x(t)| \le \max\left\{\beta(|\xi|, t),\ \gamma_2\left(\left\|y|_{[0,t]}\right\|\right),\ \frac{|\xi|}{4},\ \frac{1}{4}\vartheta^{-1}\left(2\gamma_2\left(\left\|y|_{[0,t]}\right\|\right)\right)\right\}$$

for all $t \in [0, t_{\max})$. Noticing that

$$\frac{1}{4}\vartheta^{-1}\left(2\gamma_2\left(\|y|_{[0,t]}\|\right)\right) \leq \gamma_2\left(\|y|_{[0,t]}\|\right)$$

(because $\vartheta^{-1}(r) \leq r$), we arrive at (3.2). □

LEMMA 3.4. *Given any $\mathcal{KL}$-function $\hat{\beta}$, there exists a $\mathcal{KL}$-function β and a $\mathcal{K}_\infty$-function ν such that for any $\tau > 0$, any continuous function $\mu : [0,\tau] \to \mathbf{R}_{\geq 0}$, and any nonnegative constant C, the following implication holds:*

$$\forall t_1, t_2,\ \ 0 \leq t_1 < t_2 \leq \tau,\quad \mu(t_2) \leq \max\left\{\hat{\beta}(\mu(t_1), t_2 - t_1),\ \frac{\mu(t_1)}{2},\ C\right\} \tag{3.6}$$

implies

$$\mu(\tau) \leq \max\left\{\beta(\mu(0),\tau),\ \ \nu(C)\right\}. \tag{3.7}$$

Proof. By Proposition 7 in [38], there exist μ_1 and $\mu_2 \in \mathcal{K}_\infty$ such that

$$\hat{\beta}(r,t) \leq \mu_1(\mu_2(r)e^{-t}),$$

so, by majorizing $\hat{\beta}$ as above if necessary, we can assume without loss of generality that $\hat{\beta}$ is continuous in its second variable, and $\hat{\beta}(r,0) \geq r$ for all r. For any $r > 0$ define T_r to be the first time when $\hat{\beta}(r, T_r) = r/2$. By replacing $\hat{\beta}$ with the $\mathcal{KL}$-function $\widetilde{\beta}$, defined by

$$\widetilde{\beta}(r,t) := \max\left\{\hat{\beta}(r,t),\ \hat{\beta}(r,0)e^{-t}\right\},$$

we can assume without loss of generality that the series $\sum_{i=0}^{\infty} T_{\frac{r}{2^i}}$ diverges for every $r > 0$.

Define a function $\phi : \mathbf{R}_{\geq 0} \times \mathbf{R}_{\geq 0} \to \mathbf{R}_{\geq 0}$ as follows:

$$\phi(r,t) = \begin{cases} \hat{\beta}(r,t) & \text{for } t \in [0, T_r), \\ \hat{\beta}\left(\frac{r}{2^k}, t - \sum_{i=0}^{k-1} T_{\frac{r}{2^i}}\right) & \text{for } \ t \in \left[\sum_{i=0}^{k-1} T_{\frac{r}{2^i}}, \sum_{i=0}^{k} T_{\frac{r}{2^i}}\right),\ \ k = 1,2,3\dots. \end{cases}$$

Notice that the following two conditions hold for ϕ.

(1) For every R, $\varepsilon > 0$, there exists $\widetilde{t} > 0$ such that $\phi(r,t) \leq \varepsilon$ for all $r < R$ and $t > \widetilde{t}$.

Indeed, fix positive R and ε and find $k \in \mathbb{Z}$ such that $\hat{\beta}(R/2^k, 0) < \varepsilon$. Next, by continuity of $\hat{\beta}$ and by compactness of $[0,R]$ we can find a $\widetilde{t}$, such that $\sum_{i=0}^{k-1} T_{\frac{r}{2^i}} < \widetilde{t}$ for all positive $r < R$. Then, if $r < R$ and $t > \widetilde{t}$, then

$$\phi(r,t) = \hat{\beta}\left(\frac{r}{2^k}, t - \sum_{i=0}^{k-1} T_{\frac{r}{2^i}}\right) < \hat{\beta}\left(\frac{r}{2^k}, 0\right) < \varepsilon.$$

(2) For all $\varepsilon > 0$ there exists $\delta > 0$ such that if $r \leq \delta$, then $\phi(r,t) < \varepsilon$ for all $t \geq 0$.

Indeed, for all r and t, $\phi(r,t) \leq \hat{\beta}_0(r) := \hat{\beta}(r,0)$. For any positive ε, take $\delta = \delta(\varepsilon) := \hat{\beta}_0^{-1}(\varepsilon)$. Then, for all $t \geq 0$ we have

$$\phi(r,t) \leq \hat{\beta}\left(\hat{\beta}_0^{-1}(\varepsilon), 0\right) \leq \varepsilon.$$

Therefore, by Lemma 2.11, ϕ can be majorized by a $\mathcal{KL}$-function β.

Let $\nu(r) = \hat{\beta}(2r, 0)$.

Now pick any μ, C, and τ satisfying (3.6). Define $T = \min\{t : \mu(t) \leq 2C\}$ and $T = \tau$ if $\mu(t) > 2C$ for all $t \geq 0$.

For any t_1 and t_2 in $[0, \tau]$ such that $0 \leq t_1 \leq t_2 \leq T$, we have $\mu(t_1) > 2C$, so that $\mu(t_1)/2 > C$, hence

$$\mu(t_2) \leq \max\left\{\hat{\beta}(\mu(t_1), t_2 - t_1), \mu(t_1)/2\right\}. \tag{3.8}$$

Suppose now that $\tau = T$. If $0 \leq \tau < T_{\mu(0)}$, then (3.8) with $t_1 = 0$, $t_2 = \tau$ yields

$$\mu(\tau) \leq \max\left\{\hat{\beta}(\mu(0), \tau), \frac{\mu(0)}{2}\right\} = \hat{\beta}(\mu(0), \tau),$$

where the equality follows from the definition of $T_{\mu(0)}$. Likewise, if

$$\tau \in \left[\sum_{i=0}^{k-1} T_{\frac{\mu(0)}{2^i}}, \sum_{i=0}^{k} T_{\frac{\mu(0)}{2^i}}\right),$$

then

$$\begin{aligned}\mu(\tau) &\leq \max\left\{\hat{\beta}\left(2^{-k}\mu(0), \tau - \sum_{i=0}^{k-1} T_{\frac{\mu(0)}{2^i}}\right), 2^{-(k+1)}\mu(0)\right\} \\ &= \hat{\beta}\left(2^{-k}\mu(0), \tau - \sum_{i=0}^{k-1} T_{\frac{\mu(0)}{2^i}}\right),\end{aligned}$$

where the inequality follows from (3.8) and the equality is implied by the definition of $T_{\frac{\mu(0)}{2^k}}$. Therefore we have

$$\mu(\tau) \leq \phi(\mu(0), \tau). \tag{3.9}$$

In case $\tau > T$, inequality (3.6) implies

$$\begin{aligned}\mu(\tau) &\leq \max\left\{\hat{\beta}(2C, \tau - T), \mu(T)/2, C\right\} \\ &= \max\left\{\hat{\beta}(2C, \tau - T), C, C\right\} \leq \hat{\beta}(2C, 0) = \nu(C).\end{aligned} \tag{3.10}$$

Combining (3.9) and (3.10) we obtain

$$\mu(\tau) \leq \max\{\phi(\mu(0), \tau), \nu(C)\} \leq \max\{\beta(\mu(0), \tau), \nu(C)\}. \qquad \square$$

Proof of Lemma 3.2. We need to show that the system (3.1), corresponding to our system (2.1) with the stability margin φ we have defined, is UOSS. Apply Lemma 3.4 to the $\mathcal{KL}$-function $\hat{\beta} := \beta$ to find appropriate functions $\beta_1 \in \mathcal{KL}$ and $\nu \in \mathcal{K}$. Assume given any initial state ξ and disturbance $\mathbf{d} = [\mathbf{d}_u, \mathbf{w}]$, and let $x(t) := x(t, \xi, \mathbf{d}_u, \mathbf{w})$ be the corresponding solution. Fix any positive $t < t_{\max}(\xi, \mathbf{d})$, and define $\mathbf{u}$ by

$$\mathbf{u}(s) := \begin{cases} \varphi(|x(s)|)\mathbf{d}_u(s), & s \leq t \\ 0, & s > t. \end{cases}$$

Then, for all $s \le t$ we have $x(s) = x(s, \xi, \mathbf{u}, \mathbf{w})$, where the latter is the solution of the original system (2.1) with control $\mathbf{u}$ and disturbance $\mathbf{w}$. Let $C = \gamma_2(\left\|y|_{[0,t]}\right\|)$. Then, for any t_1 and t_2 in $[0, t]$ we have $C \ge \gamma_2(\left\|y|_{[t_1,t_2]}\right\|)$. So, since $|\mathbf{u}(s)| \le \varphi(|x(s)|)$ for all $s \in [0, t]$, Proposition 3.3 will imply that

$$|x(t_2)| \le \max\left\{\beta(|x(t_1)|, t_2 - t_1),\ \frac{|x(t_1)|}{2},\ C\right\}.$$

By the choice of β_1 and ν we then have

$$|x(t)| \le \max\left\{\beta_1(|\xi|, t), \nu(\gamma_2(\left\|y|_{[0,t]}\right\|))\right\},$$

proving the UOSS property for system (3.1) corresponding to the original UIOSS system. Thus, φ is indeed a stability margin for the original system, and the proof of Lemma 3.2 is now complete. □

3.2. A UIOSS system admits a UIOSS-Lyapunov function. We show now how the main implication of Theorem 2.4 follows from Theorem 2.16.

LEMMA 3.5 (see Lemma 2.13 in [42]). *Suppose a system* Σ *of type* (2.1) *is ROSS. Let* V *be a UOSS-Lyapunov function for the system* (3.1) *associated with* Σ*. Then* V *is a UIOSS-Lyapunov function for* Σ*.*

Proof. Let φ be a stability margin for Σ. Since V is a UOSS-Lyapunov function for (3.1), inequalities (2.3) and (2.16) hold with some α_1, α_2, α_3, and γ. Pick a state $\xi \in \mathbf{X}$ and disturbance value $w \in \Gamma$. For any control value $u \in \mathbb{U}$ with $|u| \le \varphi(|\xi|)$ we can find a $d_u \in \mathbb{U}_1$ such that $u = d_u \varphi(|\xi|)$, so that by the dissipation inequality (2.16) for V (applied with $d := [d_u, w]$) we have

$$\nabla V(\xi) \cdot f(\xi, u, w) = \nabla V(\xi) \cdot g(\xi, d) \le -\alpha_3(|\xi|) + \gamma(|h(\xi)|),$$

proving (2.5) for V. So, the condition as in Remark 2.1 is satisfied for V with $\chi_1 = \varphi^{-1}$, and α_i and γ as before. Thus, V is a UIOSS-Lyapunov function for Σ. □

By Theorem 2.16, the system (3.1) admits a UOSS-Lyapunov function V. Hence the following corollary follows.

COROLLARY 3.6. *If a system* (2.1) *is ROSS, then it admits a UIOSS-Lyapunov function.*

By Lemma 3.2, every UIOSS system is also ROSS, hence the implication 1 ⇒ 2 of Theorem 2.4 follows.

3.3. UOSS and iiUOSS imply the GASMO property.

LEMMA 3.7. *A UOSS system of type* (2.14) *satisfies the GASMO property.*

Proof. Assume that system (2.14) is UOSS. Without loss of generality, we may assume that γ_2 in (2.15) is of class $\mathcal{K}_\infty$.

Let $\vartheta(s) = \beta(s, 0)$. Recall that we have assumed that $\vartheta(s) > s$ for all $s > 0$. Now let ρ be any $\mathcal{K}_\infty$-function satisfying the inequality $\rho(s) > \vartheta(4\gamma_2(s))$ for all $s > 0$.

Claim. For any $\xi \in \mathbf{X}$, any $\mathbf{d} \in \mathcal{M}_\Omega$, and any $\tau \in [0, t_{\max}(\xi, \mathbf{d}))$, if

$$|x(t, \xi, \mathbf{d})| \ge \rho(|y(t, \xi, \mathbf{d})|)\ \forall\ 0 \le t \le \tau,$$

then

$$\gamma_2(|y(t, \xi, \mathbf{d})|) \le |\xi| / 2\ \forall\ 0 \le t \le \tau,$$

and hence

$$|x(t, \xi, \mathbf{d})| \le \beta(|\xi|, 0) = \vartheta(|\xi|)\ \forall\ 0 \le t \le \tau.$$

In particular, if $|x(t,\ \xi,\mathbf{d})| \geq \rho(|y(t,\xi,\mathbf{d})|)$ for all $t \in [0, t_{\max}(\xi,\mathbf{d}))$, then $t_{\max}(\xi,\mathbf{d}) = \infty$.

Proof of the claim. If $\xi = 0$, the result is clear. Pick any $\xi \neq 0$, $\mathbf{d} \in \mathcal{M}_\Omega$ and assume that $|x(t,\xi,\mathbf{d})| \geq \rho(|y(t,\xi,\mathbf{d})|)$ for all $0 \leq t \leq \tau$ for some $\tau \in (0,\ t_{\max}(\xi,\mathbf{d}))$. Then, at $t = 0$,

$$\gamma_2(|y(0,\xi,\mathbf{d})|) \leq \gamma_2(\rho^{-1}(|\xi|)) \leq \gamma_2(\rho^{-1}(\vartheta(|\xi|))) < |\xi|/4.$$

Hence, $\gamma_2(|y(t,\xi,\mathbf{d})|) < |\xi|/4$ for all $t \in [0,\delta)$ for some $\delta > 0$. Let

$$t_1 = \inf\left\{t > 0 : \gamma_2(|y(t,\xi,\mathbf{d})|) \geq |\xi|/2\right\}.$$

Then $t_1 > 0$. Assume now that $t_1 \leq \tau$. Then

$$\gamma_2(|y(t_1,\xi,\mathbf{d})|) = |\xi|/2 \quad\text{and}\quad \gamma_2(|y(t,\xi,\mathbf{d})|) < |\xi|/2$$

for each $t \in [0,\ t_1)$, and hence for such t, $|x(t,\xi,d)| \leq \vartheta(|\xi|)$. Then, for each $0 \leq t \leq t_1$,

$$\gamma_2(|y(t,\xi,\mathbf{d})|) \leq \gamma_2(\rho^{-1}(|x(t,\xi,\mathbf{d})|)) \leq \gamma_2(\rho^{-1}(\vartheta(|\xi|))) < |\xi|/4.$$

By continuity, $\gamma_2(|y(t_1,\xi,\mathbf{d})|) \leq |\xi|/4$, contradicting the definition of t_1. This shows that it is impossible to have $t_1 \leq \tau$, and the proof of the claim is complete.

For each $r > 0$ let T_r be any nonnegative number so that $\beta(r,t) < r/2$ for all $t \geq T_r$. Now, given any $r > 0$ and any $\varepsilon > 0$, for each $i = 1, 2, \ldots$, let $r_i := 2^{1-i}r$, and let $k(\varepsilon)$ be any positive integer so that $2^{-k(\varepsilon)}r < \varepsilon$ and define $T_{r,\varepsilon}$ as $T_{r_1} + T_{r_2} + \cdots + T_{r_{k(\varepsilon)}}$.

Pick any trajectory $x(t,\xi,\mathbf{d})$ as in the statement of Proposition 2.10, defined on an interval of the form $[0,T]$, with $T \geq T_{r,\varepsilon}$, with initial condition $|\xi| \leq r$, and disturbance $\mathbf{d} \in \mathcal{M}_\Omega$, satisfying $|x(t,\xi,\mathbf{d})| \geq \rho(|y(t,\xi,\mathbf{d})|)$ for all $t \in [0,T]$. Then, the above claim implies that $\gamma_2(|y(t,\xi,\mathbf{d})|) < |\xi|/2$ for all such t. Therefore, for any $t > T_{r_1} = T_r$,

$$\begin{aligned}|x(t,\xi,\mathbf{d})| &\leq \max\left\{\beta(|\xi|,t),\ |\xi|/2\right\}\\ &\leq \max\left\{\beta(r,t),\ r/2\right\} \leq r/2.\end{aligned}$$

Consider now the restriction of the trajectory to the interval $[T_{r_1}, T]$. This is the same as the trajectory that starts from the state $x(T_{r_1},\xi,\mathbf{d})$, which has norm less than r_1, so by the same argument and the definition of T_{r_2} we have that $|x(t,\xi,d)| \leq r/4$ for all $t \geq T_{r_2}$. Repeating on each interval $[T_{r_i}, T_{r_{i+1}}]$, we conclude that $|x(t,\xi,d)| < \varepsilon$ for all $T_{r,\varepsilon} \leq t \leq T$. □

LEMMA 3.8. *Suppose a system of type* (2.14) *is iiUOSS and UO. Then it satisfies the GASMO property with* $\rho(\cdot) := \max\left\{\chi^{-1}(2\gamma(\cdot)), \rho_1^{-1}(\cdot)\right\}$, *where* χ *and* γ *are as in the definition of iiUOSS and* ρ_1 *is as in Proposition* 2.15.

To prove this lemma, we need the following elementary observation, which is a variant of what is usually referred to as "Barbălat's lemma."

PROPOSITION 3.9. *Let* $\mathcal{X} := \{x_\alpha,\ \alpha \in \mathcal{A}\}$ *be a family of absolutely continuous curves in* $\mathbf{X}$, *each of which is defined on an interval* $\mathcal{I}_\alpha$, *either half-open* ($\mathcal{I}_\alpha = [0,\lambda_\alpha)$) *or closed* ($\mathcal{I}_\alpha = [0,\lambda_\alpha]$). *Suppose the following.*

- $\mathcal{X}$ *is closed with respect to shifts, that is, for all* $\alpha \in \mathcal{A}$ *and* $T \in \mathcal{I}_\alpha$, *there exists an* $\alpha' \in \mathcal{A}$ *such that* $x_{\alpha'} \equiv x_{\alpha T}$, *where* $x_{\alpha T}$ *is defined by* $x_{\alpha T}(t) := x_\alpha(t+T)$, *and* $\lambda_{\alpha'} = \lambda_\alpha - T$.
- *There exists a nonnegative, increasing function* ν_3 *such that*

$$|\dot{x}_\alpha(t)| \leq \nu_3(|x_\alpha(t)|)\ \ \forall\, \alpha \in \mathcal{A} \quad \textit{for almost all } t \in \mathcal{I}_\alpha.$$

- *There exist functions κ and χ of class $\mathcal{K}_\infty$ such that*

$$\kappa(|x_\alpha(0)|) \geq \int_0^t \chi(|x_\alpha(s)|)\,ds \quad \forall\, \alpha \in \mathcal{A},\ t \in \mathcal{I}_\alpha.$$

Then for any two positive numbers r and ε there exists a $T_{r,\varepsilon}$, such that for all $\alpha \in \mathcal{A}$ and $t \in \mathcal{I}_\alpha$ the following holds:

$$t \geq T_{r,\varepsilon} \text{ and } |x_\alpha(0)| \leq r \Rightarrow |x_\alpha(t)| < \varepsilon.$$

Proof.

Claim 1. Given any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon)$ such that if $|x_\alpha(0)| \leq \delta$, then $|x_\alpha(t)| < \varepsilon$ for all $t \in I_\alpha$.

Proof of Claim 1. Fix a positive ε, and set

$$\delta(\varepsilon) = \min\left\{\frac{\varepsilon}{2}, \kappa^{-1}\left(\frac{\varepsilon\chi(\varepsilon/2)}{2\nu_3(\varepsilon)}\right)\right\}.$$

Pick any $\alpha \in \mathcal{A}$ such that $|x_\alpha(0)| \leq \delta$.

Suppose $\left|x_\alpha(\widetilde{t}_2)\right| \geq \varepsilon$ for some $\widetilde{t}_2 \in \mathcal{I}_\alpha$. Then there exist t_1 and t_2 with $t_1 < t_2 \leq \widetilde{t}_2$ such that $|x_\alpha(t_1)| = \varepsilon/2$ and $\varepsilon/2 < |x_\alpha(t)| < \varepsilon$ for all $t \in (t_1, t_2)$. Then

$$\frac{\varepsilon}{2} = |x_\alpha(t_2)| - |x_\alpha(t_1)| \leq |x_\alpha(t_2) - x_\alpha(t_1)| \leq \sup_{t_1 \leq t \leq t_2} |\dot{x}_\alpha(t)|\,(t_2 - t_1) \leq \nu_3(\varepsilon)(t_2 - t_1).$$

So,

$$\kappa(\delta) \geq \kappa(|\xi|) \ \geq \ \int_{t_1}^{t_2} \chi(|x_\alpha(s)|)\,ds > \frac{1}{2}\chi\left(\frac{\varepsilon}{2}\right)(t_2 - t_1) \geq \frac{\varepsilon\chi(\varepsilon/2)}{2\nu_3(\varepsilon)} \geq \kappa(\delta).$$

The obtained contradiction proves the claim.

Claim 2. Given positive numbers r and δ, there exists a time $\tau(r,\delta)$ such that if $|x_\alpha(0)| \leq r$ and $\tau(r,\delta) \in \mathcal{I}_\alpha$, then $\exists t_0 < \tau(r,\delta)$ such that $|x(t_0, \xi, \mathbf{d})| \leq \delta$.

Proof of Claim 2. Take $\tau(r,\delta) = \frac{2\kappa(r)}{\chi(\delta)}$. Then, if $\tau(r,\delta) \in I_\alpha$ and $|x_\alpha(t)| > \delta$ for all $t \in [0, \tau(r,\delta))$, then we have

$$\kappa(r) \geq \kappa(|x_\alpha(0)|) \geq \int_0^{\tau(r,\delta)} \chi(|x_\alpha(s)|)\,ds > \chi(\delta)\frac{\kappa(r)}{\chi(\delta)} = \kappa(r).$$

The obtained contradiction proves the claim.

Fix arbitrary positive r and ε. By Claim 1, find $\delta(\varepsilon)$ such that if $|x_\alpha(0)| < \delta(\varepsilon)$, then $|x_\alpha(t)| \leq \varepsilon$ for all $t \in [0, \lambda_\alpha)$. Define $T_{r,\varepsilon} := \tau(r, \delta(\varepsilon))$, where $\tau(r,\delta(\varepsilon))$ is furnished by Claim 2. If $|x_\alpha(0)| < r$, then, by Claim 2, there is a $t_0 < \tau(r,\delta(\varepsilon))$ with $|x_\alpha(t_0)| < \delta(\varepsilon)$. Consider now a function $x_{\alpha'}(\cdot) := x_\alpha(t_0 + \cdot)$ (it belongs to $\mathcal{X}$ by assumption). Since $|x_{\alpha'}| \leq \delta(\varepsilon)$, Claim 1 ensures that

$$|x_\alpha(t)| = |x_{\alpha'}(t - t_0)| \leq \varepsilon \quad \forall\, t \geq t_0 \geq T_{r,\varepsilon}.$$

This shows that $T_{r,\varepsilon}$ satisfies the conclusion of the proposition. □

We now return to the proof of Lemma 3.8.

Proof. Recall that we have defined $\lambda_{\xi,\mathbf{d}} := \inf\{t \in [0, t_{\max}(\xi,\mathbf{d})) : |x(t,\xi,\mathbf{d})| \leq \rho(|y(t,\xi,\mathbf{d})|)\}$, and let $\lambda_{\xi,\mathbf{d}} = t_{\max}$ if $|x(t,\xi,d)| > \rho(|y(t,\xi,\mathbf{d})|)$ for all $t \in [0, t_{\max})$.

Note that, given ξ and $\mathbf{d}$, for all $t < \lambda_{\xi,\mathbf{d}}$ we have $\chi(|x(t,\xi,\mathbf{d})|) > 2\gamma(|h(x(t,\xi,\mathbf{d}))|)$ so that

$$\kappa(|\xi|) \geq \int_0^t (\chi(|x(s,\xi,\mathbf{d})|) - \gamma(|h(x(s,\xi,\mathbf{d}))|))\, ds > \frac{1}{2}\int_0^t \chi(|x(s,\xi,\mathbf{d})|)\, ds. \tag{3.11}$$

Let ν_3 be a $\mathcal{K}$-function such that $\max_{d\in\Omega}|f(x,d)| \leq \nu_3(|x|)$. Write $x_{\xi,\mathbf{d}}(\cdot) := x(\cdot,\xi,\mathbf{d})$. Notice that the family $\{x_{\xi,\mathbf{d}}(\cdot),\ \xi \in \mathbf{X},\ \mathbf{d} \in \mathcal{M}_\Omega\}$ with $\mathcal{I}_{\xi,\mathbf{d}} := [0, \lambda_{\xi,\mathbf{d}})$ satisfies all the assumptions of Proposition 3.9 (with "κ" $= 2\kappa$). Given any positive r, ε, Proposition 3.9 furnishes $T_{r,\varepsilon}$. This $T_{r,\varepsilon}$ obviously fits the first condition in the characterization of the GASMO property, provided by Proposition 2.10.

To find a function ϑ to ensure that the second part of Proposition 2.10 is satisfied, recall that, by Proposition 2.15, if a system (2.14) has the UO property, then there exist class $\mathcal{K}_\infty$-functions ρ_1, μ_1, μ_2, and a constant $c > 0$ such that the following implication holds for all $\xi \in \mathbf{X}$, all $\mathbf{d} \in \mathcal{M}_\Omega$, and all $T \in [0, t_{\max}(\xi,\mathbf{d}))$:

$$\begin{aligned}|h(x(t,\xi,\mathbf{d}))| \ &\leq \ \rho_1(|x(t,\xi,\mathbf{d})|) \ \ \forall t \in [0,T] \\ &\Rightarrow |x(t,\xi,\mathbf{d})| \leq \mu_1(t) + \mu_2(|\xi|) + c \ \ \forall t \in [0,T].\end{aligned}$$

Therefore, if $|\xi| \leq r$ and $T_{r,r/2}$ is as defined above, then for all $t \in [0, \lambda_{\xi,\mathbf{d}})$ we have

$$|x(t,\xi,\mathbf{d})| \leq \mu_1(T_{r,r/2}) + \mu_2(r) + c \ \text{ if } t < T_{r,r/2}$$

and

$$|x(t,\xi,\mathbf{d})| \leq r/2 \ \text{ if } t \geq T_{r,r/2}.$$

Thus, the following estimate holds for all such t:

$$|x(t,\xi,\mathbf{d})| \leq \widetilde{\vartheta}(|\xi|) := \max\left\{|\xi|/2,\ \mu_1(T_{|\xi|,|\xi|/2}) + \mu_2(|\xi|) + c\right\}.$$

Next, take a sequence $\{\varepsilon_k\}$, $k = 0, 1, 2, \ldots$, strictly decreasing to 0, with $\varepsilon_0 = 1$. For each ε_k, find $\delta_k = \delta(\varepsilon_k)$ as in the proof of Claim 1. Since $\delta_k \leq \varepsilon_k/2$, the sequence $\{\delta_k\}$ converges to 0 as well. Find a function ϑ of class $\mathcal{K}$, such that

(1) $\vartheta(\delta_{k+1}) > \varepsilon_k \ \ \forall k > 0$,
(this will ensure that $|x(t,\xi,\mathbf{d})| \leq \vartheta(|\xi|)$ for all ξ with $|\xi| < \delta_0$, for all $t \in [0, \lambda_{\xi,\mathbf{d}}]$)
(2) $\vartheta(s) \geq \widetilde{\vartheta}(s) \ \ \forall s > \delta_0$.

Then ϑ satisfies the second condition in the Proposition 2.10. This completes the proof. □

Remark 3.1. The unboundedness observability assumption is crucial in proving the last lemma. The following example illustrates a disturbance-free integral-to-integral output-to-state stable (iiOSS) system which fails to be OSS (and, equivalently, fails to be GASMO).

Let $1_A(\cdot)$ denote the indicator function of a set A, and let ϕ_ε be a C^∞-bump function with support in $(-\varepsilon, \varepsilon)$:

$$\phi_\varepsilon(\xi) := \begin{cases} e^{-\frac{|\xi|^2}{\varepsilon^2 - |\xi|^2}}, & |\xi| < \varepsilon, \\ 0, & |\xi| \geq \varepsilon. \end{cases} \tag{3.12}$$

Fix an arbitrary positive $\varepsilon < 0.25$ and consider a one dimensional autonomous system

$$\Sigma: \qquad \dot{x} = f(x), \quad y = h(x),$$

where

$$\begin{aligned} f(x) = x^3 & \left[1_{(-\infty,-1]}(x)(1-\phi_\varepsilon(x+1)) + 1_{[1,+\infty)}(x)(1-\phi_\varepsilon(x-1))\right] \\ & -x\left[1_{(-1,1)}(x)(1-\phi_\varepsilon(x+1))(1-\phi_\varepsilon(x-1))\right], \end{aligned}$$

and h is a smooth function such that $h(x) = x$ for all x in $[-2, 2]$, and $h(x) = 0$ if $|x| \geq 3$.

Claim 1. The system Σ is iiOSS.

Proof. Note that Σ has a stable equilibrium at $x = 0$ and two unstable ones at 1 and -1. If $|x| < 1$, then $\text{sign}(x) = -\text{sign}(f(x))$, so, if $|\xi| \leq 1$, then $|x(t,\xi)| \leq 1$ for any nonnegative t. Therefore, if $\xi \in [-1, 1]$, then for all $t \geq 0$ we have

$$\int_0^t |x(s,\xi)|\,ds = \int_0^t |h(x(s,\xi))|\,ds, \tag{3.13}$$

so, estimate (2.18) trivially follows for all $\xi \in [-1, 1]$ and all $t \in t_{\max}(\xi)$ with $\gamma = Id$ and any $\kappa \in \mathcal{K}$.

If $|\xi| \geq 1 + \varepsilon$, then $f(x) = x^3$, so that

$$x(t,\xi) = \frac{\text{sign}(\xi)}{\sqrt{\xi^{-2} - 2t}}.$$

Thus, in this case the solution $x(t,\xi)$ is defined for all nonnegative $t < t_{\max}(\xi) = \xi^{-2}/2$ and

$$\int_0^t |x(s,\xi)|\,ds \leq \int_0^{t_{\max}(\xi)} |x(s,\xi)|\,ds = \frac{1}{\xi} \leq \frac{1}{1+\varepsilon}.$$

Let κ be any $\mathcal{K}$-function such that $\kappa(1) \geq (1+\varepsilon)^{-1}$. Suppose $1 < |\xi| < 1 + \varepsilon$. Let $\hat{t}$ be the time when $\left|x(\hat{t},\xi)\right| = 1 + \varepsilon$. Then $t_{\max}(\xi) = \hat{t} + (1+\varepsilon)^{-2}/2$. Also, $x(s,\xi) = h(x(s,\xi))$ for all $s \in [0,\hat{t}]$, so, in particular, equality (3.13) holds for all $t < \hat{t}$, which, again, trivially implies (2.18) with $\gamma = Id$ and any $\kappa \in \mathcal{K}$.

If $t > \hat{t}$, then

$$\begin{aligned} \int_0^t |x(s,\xi)|\,ds &= \int_0^{\hat{t}} |x(s,\xi)|\,ds + \int_{\hat{t}}^t |x(s,\xi)|\,ds \\ &\leq \int_0^{\hat{t}} |x(s,\xi)|\,ds + \int_{\hat{t}}^{t_{\max}(\xi)} |x(s,\xi)|\,ds \\ &= \int_0^{\hat{t}} |h(x(s,\xi))|\,ds + \int_0^{t_{\max}(1+\varepsilon)} |x(s,1+\varepsilon)|\,ds \\ &\leq \int_0^{\hat{t}} |h(x(s,\xi))|\,ds + \kappa(\xi) \\ &\leq \int_0^{t} |h(x(s,\xi))|\,ds + \kappa(\xi). \end{aligned}$$

This shows that Σ is iiOSS, as estimate (2.18) holds for Σ with the κ that we constructed and $\gamma = Id$. □

Claim 2. System Σ is not OSS.

Proof. Indeed, pick any initial state ξ of large enough magnitude so that $h(\xi) = 0$. Then $h(x(t,\xi)) = 0$ for all $t < t_{\max}(\xi) = \xi^{-2}/2$. If Σ were OSS, then there would exist some $\mathcal{KL}$-function β such that $|x(t,\xi)| \le \beta(|\xi|, t)$, but $|x(t,\xi)|$ tends to ∞ as $t \to t_{\max}(\xi)$, whereas $\beta(|\xi|, t) \le \beta(|\xi|, 0)$. This contradiction proves the claim. □

We now prove implication 4 ⇒ 3 of Theorem 2.16.

Proof. Suppose a system Σ of type (2.14) is UOSS. We have already remarked that Σ is UO, so we must show that it is iiUOSS. By assumption there exists a smooth function V satisfying (2.16) and (2.3) with some α_1, α_2, α_3, and γ. Pick any ξ, $\mathbf{d}$, and $t \in [0, t_{\max}(\xi, \mathbf{d}))$. Integrating inequality (2.16) along the trajectory $x(\cdot, \xi, \mathbf{d})$ over $[0, t]$ we get

$$\begin{aligned}\int_0^t \alpha_3(|x(t,\xi,\mathbf{d})|)dt &\le V(x(0,\xi,\mathbf{d})) - V(x(t,\xi,\mathbf{d})) + \int_0^t \gamma(|h(x(t,\xi,\mathbf{d}))|)\, dt \\ &\le \alpha_2(|\xi|) + \int_0^t \gamma(|h(x(t,\xi,\mathbf{d}))|)\, dt,\end{aligned}$$

proving inequality (2.18) for system Σ, with $\chi = \alpha_3$ and $\kappa = \alpha_2$. □

With Lemma 3.7 in mind we conclude that the only step missing in establishing the Lyapunov characterization for UOSS is proving the implication 2.16 ⇒ 2.16 in Theorem 2.16.

3.4. A remark on the GASMO ⇒ UOSS implication.

Remark 3.2. Once the Lyapunov characterization for UOSS is proven, the implication 2 ⇒ 1 of Theorem 2.16 will automatically follow by appealing to the converse Lyapunov theorem (2 ⇒ 4 ⇒ 1). However, it is worth mentioning that GASMO ⇒ UOSS implication can easily be proven directly without this intermediate step.

Indeed, fix $\xi \in \mathbf{X}$ and $\mathbf{d} \in \mathcal{M}_\Omega$. Take any $t \in [0, t_{\max}(\xi, \mathbf{d}))$. If $t \le \lambda_{\xi,\mathbf{d}}$, then the GASMO property provides the estimate

$$|x(t,\xi,\mathbf{d})| \le \lambda(|\xi|, t). \tag{3.14}$$

Suppose now that $t > \lambda_{\xi,\mathbf{d}}$. Obviously, at least one of the two following conditions must be satisfied:

(1) $|x(t,\xi,\mathbf{d})| > 2\rho(|h(x(t,\xi,\mathbf{d}))|)$,

(2) $|x(t,\xi,\mathbf{d})| \le 3\rho(|h(x(t,\xi,\mathbf{d}))|)$.

If condition (2) applies, then we have a bound

$$|x(t,\xi,\mathbf{d})| \le 3\rho(|h(x(t,\xi,\mathbf{d}))|) \le 3\rho(\left\|y|_{[0,t]}\right\|). \tag{3.15}$$

In case condition (1) holds, let

$$\widetilde{t} := \max\left\{s,\ \lambda_{\xi,\mathbf{d}} < s < t \ : \ |x(s,\xi,\mathbf{d})| = 2\rho(|h(x(s,\xi,\mathbf{d}))|)\right\}.$$

Then, again by the GASMO property applied with initial state $x(\widetilde{t}, \xi, \mathbf{d})$, we have

$$\begin{aligned}|x(t,\xi,\mathbf{d})| &\le \lambda(\left|x(\widetilde{t},\xi,\mathbf{d})\right|, t - \widetilde{t}) \\ &\le \lambda(2\rho(\left|y(\widetilde{t},\xi,\mathbf{d})\right|), 0) \le \lambda(2\rho(\left\|y|_{[0,t]}\right\|), 0).\end{aligned} \tag{3.16}$$

Combining estimates (3.14), (3.15), and (3.16), we conclude that inequality (2.15) holds for Σ with $\beta(\cdot) := \lambda(\cdot)$ and $\gamma_2(\cdot) := \max\{\lambda(2\rho(\cdot), 0),\ 3\rho(\cdot)\}$.

4. The case of no controls.

4.1. Setup. Suppose a system Σ of type (2.14) satisfies the GASMO property with some $\mathcal{K}_\infty$ function ρ. By majorizing ρ with another $\mathcal{K}_\infty$-function if necessary, we will assume that ρ is smooth when restricted to $s > 0$ and also $\rho(s) > s$ for all positive s. We let

$$\mathcal{D} := \{\xi \in \mathbf{X}\colon\ |\xi| \le \rho(|h(\xi)|)\}\ ,$$

$$\mathcal{E} := \mathbf{X} \backslash \mathcal{D}\,,$$

and

$$\mathcal{E}_1 := \{\xi \in \mathbf{X} : |\xi| > 2\rho(|h(\xi)|)\}\ .$$

For each $\mathbf{d} \in \mathcal{M}_\Omega$ and $\xi \in \mathcal{E}$, define

$$\lambda_{\xi,\mathbf{d}} = \inf\{t \in [0, t_{\max}) : x(t, \xi, \mathbf{d}) \in \mathcal{D}\}\,, \tag{4.1}$$

with the convention $\lambda_{\xi,\mathbf{d}} = t_{\max}(\xi, \mathbf{d})$ if the trajectory never enters $\mathcal{D}$. If $\mathcal{D} = \mathbf{X}$, then any proper, smooth, and positive definite function $V : \mathbf{X} \to \mathbf{R}$ is a UOSS-Lyapunov function for (2.14). Indeed, because it is proper and finite, V obviously satisfies (2.3) for some α_1 and α_2. Since V is smooth, $|\nabla V(\xi)|$ is bounded above by a nondecreasing continuous function $\nu(|\xi|)$ and

$$\frac{d}{dt}V(x(t)) \;=\; \nabla V(x(t)) \cdot f(x(t), \mathbf{d}(t)) \;\le\; \nu(|x(t)|)\nu_3(|x(t)|),$$

where $\nu_3(|\cdot|)$ is a $\mathcal{K}$-function majorizing $f(\cdot, d)$ for all $d \in \Omega$. Then, since $|x| \le \rho(|h(x)|)$ for all $x \in \mathbf{X}$, we have

$$\frac{d}{dt}V(x(t)) \;\le\; -\nu(|x(t)|)\nu_3(|x(t)|) \;+\; 2\nu(\rho(|h(x(t))|))\nu_3(\rho(|h(x(t))|))\,.$$

So, V satisfies inequality

$$\nabla V(x) \cdot f(x, d) \le -\alpha_3(|x|) + \gamma(|h(x)|) \quad \forall\, x \in \mathbf{X},\ \forall\, d \in \Omega$$

(with $\alpha_3(\cdot) = \nu(\cdot)\nu_3(\cdot)$ and $\gamma(\cdot) = [2\nu \circ \rho(\cdot)][\nu_3 \circ \rho(\cdot)]$), which is the same as (2.4) for systems of type (2.14).

Suppose now that $\mathcal{D} \ne \mathbf{X}$. Recall that we have defined, for each $\xi \notin \mathcal{D}$ and $d \in \mathcal{M}_\Omega$, $\lambda_{\xi,\mathbf{d}} = \inf\{t \in [0, t_{\max}) : x(t, \xi, \mathbf{d}) \in \mathcal{D}\}$, with the convention $\lambda_{\xi,\mathbf{d}} = t_{\max}(\xi, \mathbf{d})$ if the trajectory never enters $\mathcal{D}$.

The GASMO property then implies

$$|x(t, \xi, \mathbf{d})| \le \lambda(|\xi|\,, t) \qquad \forall\, \xi \in \mathcal{E},\ \forall\, \mathbf{d} \in \mathcal{M}_\Omega,\ \forall\, t \in [0, \lambda_{\xi,\mathbf{d}}) \tag{4.2}$$

for some $\lambda \in \mathcal{KL}$.

Note that, because of property (4.2), the system cannot have any equilibrium in $\mathcal{E}$, that is,

$$f(\xi, d) \ne 0$$

for every $\xi \in \mathcal{E}$ and every $d \in \Omega$. Moreover, replacing $\rho(s)$ by $c\rho(s)$ for some $c > 1$ if necessary, one may also assume that $f(\xi, d) \ne 0$ for all $\xi \in \partial\mathcal{D} \setminus \{0\}$ and all $d \in \Omega$.

We introduce an auxiliary system $\hat{\Sigma}$ which slows down the motions of the original one:

$$\dot{z} \;=\; \widehat{f}(z,d) \;=\; \frac{1}{1+|f(z,d)|^2+\kappa(z)}f(z,d), \tag{4.3}$$

where κ is any smooth function $\mathbf{X} \to [0,\infty)$ with the property that

$$\kappa(\xi) \;\geq\; 2\max_{d\in\Omega} |\nabla(\rho\circ|h|)(\xi)\cdot f(\xi,d)| \tag{4.4}$$

whenever $|h(\xi)| \geq 1$. (Recall that ρ was assumed, without loss of generality, to be smooth for positive arguments.) For each disturbance $\hat{\mathbf{d}}$ (defined on $\mathbf{R}_{\geq 0}$) denote by

$$z(s,\xi,\hat{\mathbf{d}})$$

the value at time s of the solution of the equation $\dot{z} = \widehat{f}(z,\hat{\mathbf{d}})$ with initial state ξ. Observe that, as $\widehat{f}$ is bounded, this solution exists for all nonnegative s.

Claim 1. For each ξ and each $\mathbf{d}$,

$$x(t,\xi,\mathbf{d}) = z(\sigma_{\xi,\mathbf{d}}(t),\xi,\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1}) \qquad \forall\, t\in[0,t_{\max}(\xi,\mathbf{d})), \tag{4.5}$$

where $\sigma_{\xi,\mathbf{d}} : [0,t_{\max}(\xi,\mathbf{d})) \to \mathbf{R}_{\geq 0}$ is defined by

$$\sigma_{\xi,\mathbf{d}}(t) = \int_0^t \Big[1+|f(x(s,\xi,\mathbf{d}),\mathbf{d}(s))|^2+\kappa(x(s,\xi,\mathbf{d}))\Big]\,ds.$$

Moreover, $\sigma_{\xi,\mathbf{d}}(t)\to\infty$ as $t\to t_{\max}(\xi,\mathbf{d})$, so, we can define $\sigma_{\xi,\mathbf{d}}(t_{\max}(\xi,\mathbf{d})) := +\infty$ for convenience.

Proof of Claim 1. Indeed, writing $s=\sigma_{\xi,\mathbf{d}}(t)$ and computing the derivative of $x({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d})$ with respect to s, one has

$$\begin{aligned} f(x(t,\xi,\mathbf{d}),\mathbf{d}(t)) = \frac{d}{dt}x(t,\xi,\mathbf{d}) &= \frac{d}{dt}x({\sigma_{\xi,\mathbf{d}}}^{-1}\circ\sigma_{\xi,\mathbf{d}}(t),\xi,\mathbf{d})\\ &= \frac{d}{ds}x\left({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d}\right)\cdot\frac{d}{dt}\sigma_{\xi,\mathbf{d}}(t)\\ &= \frac{d}{ds}x({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d})\Big[1+|f(x(t,\xi,\mathbf{d}),\mathbf{d}(t))|^2+\kappa(x(t,\xi,\mathbf{d}))\Big]. \end{aligned}$$

Therefore

$$\begin{aligned} \frac{d}{ds}x({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d}) &= \frac{f(x(t,\xi,\mathbf{d}),\mathbf{d}(t))}{1+|f(x(t,\xi,\mathbf{d}),\mathbf{d}(t))|^2+\kappa(x(t,\xi,\mathbf{d}))}\\ &= \frac{f\left(x({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d}),\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1}(s)\right)}{1+|f\left(x({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d}),\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1}(s)\right)|^2+\kappa\left(x\left({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d}\right)\right)}\\ &= \widehat{f}\left(x\left({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d}\right),\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1}(s)\right) \qquad \forall\, 0\leq s<\sigma_{\xi,\mathbf{d}}(t_{\max}(\xi,\mathbf{d})). \end{aligned}$$

Thus, the functions $z(s,\xi,\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1})$ and $x({\sigma_{\xi,\mathbf{d}}}^{-1}(s),\xi,\mathbf{d})$ satisfy the same differential equation (4.3) with initial state ξ on $[0,\ \sigma_{\xi,\mathbf{d}}(t_{\max}(\xi,\mathbf{d})))$, therefore they coincide on $[0,\ \sigma_{\xi,\mathbf{d}}(t_{\max}(\xi,\mathbf{d})))$.

To show the limit property of $\sigma_{\xi,\mathbf{d}}$, suppose

$$\lim_{t\to t_{\max}(\xi,\mathbf{d})} \sigma_{\xi,\mathbf{d}}(t) = b < \infty$$

(note that the limit exists because $\sigma_{\xi,\mathbf{d}}(\cdot)$ is increasing). Let

$$K = \{z(s,\xi,\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1}) : 0 \le s < b\}.$$

Then K is bounded, so $\bar{K}$ is compact, and by (4.5), $x(t,\xi,\mathbf{d}) \in K \subseteq \bar{K}$ for all $t \in [0, t_{\max}(\xi,\mathbf{d}))$, contradicting the maximality of $t_{\max}(\xi,\mathbf{d})$. □

For each initial state ξ and each disturbance function $\mathbf{d}$, define

$$\theta_{\xi,\mathbf{d}} = \inf\{t \ge 0 : \; z(t,\xi,\mathbf{d}) \in \mathcal{D}\}\,, \tag{4.6}$$

where $\theta_{\xi,\mathbf{d}} = \infty$ if $z(t,\xi,\mathbf{d}) \notin \mathcal{D}$ for all $t \ge 0$. Note that $\theta_{\xi,\mathbf{d}} > 0$ for all $\mathbf{d} \in \mathcal{M}_\Omega$ and all $\xi \in \mathcal{E}$ because $\mathcal{E}$ is open. Observe also that if $\mathbf{d}_1 = \mathbf{d}\circ\sigma_{\xi,\mathbf{d}}$,

$$\theta_{\xi,\mathbf{d}} = \sigma_{\xi,\mathbf{d}_1}(\lambda_{\xi,\mathbf{d}_1}).$$

Claim 2. System $\hat{\Sigma}$ satisfies the GASMO property.

Proof. According to (4.2) and (4.5), we have, for every $\xi \in \mathcal{E}$ and each $\mathbf{d}$,

$$\begin{aligned}|z(t,\xi,\mathbf{d})| &= \left|x({\sigma_{\xi,\mathbf{d}_1}}^{-1}(t),\xi,\mathbf{d}_1)\right| \\ &\le \lambda(|\xi|\,, {\sigma_{\xi,\mathbf{d}_1}}^{-1}(t)) \le \vartheta(|\xi|),\end{aligned}$$

for all $t \in [0,\theta_{\xi,\mathbf{d}})$, where $\vartheta(s) = \lambda(s,0)$, and $\mathbf{d}_1$ is such that $\mathbf{d} = \mathbf{d}_1 \circ {\sigma_{\xi,\mathbf{d}_1}}^{-1}$. Let

$$M_r = 1 + \max_{d\in\Omega, |\xi|\le\vartheta(r)} |f(\xi,d)|^2 + \max_{|\xi|\le\vartheta(r)} \kappa(\xi).$$

Then, for any $\xi \in \mathcal{E}$ with $|\xi| \le r$, it holds that

$$\sigma_{\xi,\mathbf{d}_1}(t) = \int_0^t \Big[1 + |f(x(s,\xi,\mathbf{d}_1),\mathbf{d}_1(s))|^2 + \kappa(x(s,\xi,\mathbf{d}_1))\Big]\, ds \;\le\; M_r t$$

for all $t \in [0,\lambda_{\xi,\mathbf{d}_1})$, and hence, ${\sigma_{\xi,\mathbf{d}_1}}^{-1}(t) \ge \frac{t}{M_r}$ for all $|\xi| \le r$, $t \in [0,\theta_{\xi,\mathbf{d}})$. Consequently, we have

$$|z(t,\xi,\mathbf{d})| \le \hat{\lambda}(|\xi|\,,t) \qquad \forall\, t \in [0,\theta_{\xi,\mathbf{d}}), \tag{4.7}$$

where $\hat{\lambda}(s,t) = \lambda(s, \frac{t}{M_s})$ is clearly of class $\mathcal{KL}$. Therefore, this shows that system $\hat{\Sigma}$ is GASMO. □

From now on, we let the function λ of class $\mathcal{KL}$ be as in Definition 2.9 for the system $\hat{\Sigma}$, that is, the following estimate holds for system (4.3):

$$|z(t,\xi,\mathbf{d})| \le \lambda(|\xi|\,,t) \qquad \forall\, t \in [0,\theta_{\xi,\mathbf{d}}), \tag{4.8}$$

for all $\xi \in \mathcal{E}$, and all $\mathbf{d} \in \mathcal{M}_\Omega$.

According to Proposition 7 in [38], there exist $\mathcal{K}_\infty$-functions μ_1 and μ_2 such that

$$\lambda(r,t) \le \mu_1(\mu_2(r)e^{-t}) \qquad \forall\, r,\; t \ge 0. \tag{4.9}$$

Define

$$\Xi(s) := \mu_1^{-1}(s). \tag{4.10}$$

The proof will now develop as follows. We first construct a continuous Lyapunov-like function V_0, defined on the set $\mathcal{E}_1$. Next V_0 is approximated by a Lipschitz continuous function (by the methods of nonsmooth analysis). The resulting function is then approximated by a smooth function V_1. Finally we extend V_1 to the rest of the state space, obtaining a Lyapunov-like function, smooth away from the origin, which is then approximated by a smooth Lyapunov function.

4.2. Definitions and basic facts on relaxed controls. Recall that our disturbances $\mathbf{d}$ are measurable functions $\mathbf{R}_{\geq 0} \to \Omega$, a compact, convex subset of $\mathbf{R}^m$.

Let $\mathcal{P}(\Omega)$ be the set of all Radon probability measures on Ω. Bishop's theorem furnishes a weak norm on $\mathcal{P}(\Omega)$, whose corresponding metric topology coincides with the weak star topology on $\mathcal{P}(\Omega)$ (see [51, pp. 40 and 267]).

For any $T > 0$, we define $\mathcal{S}_T$ to be the set of all measurable functions from $[0,T]$ to $\mathcal{P}(\Omega)$, and $\mathcal{S}$ to be the set of all measurable functions from $\mathbf{R}_{\geq 0}$ to $\mathcal{P}(\Omega)$. We topologize $\mathcal{S}_T$ by weak convergence: $\{\nu_k(\cdot)\} \to \nu(\cdot)$ in $\mathcal{S}_T$ if and only if

$$\int_0^T \int_\Omega g(t,\omega) d[\nu_k(t)](\omega)\, dt \to \int_0^T \int_\Omega g(t,\omega) d[\nu(t)](\omega)\, dt$$

for all functions $g : [0,T] \times \Omega \to \mathbf{R}$ which are continuous in ω, measurable in t, and such that

$$\max\{|g(t,\omega)|\,,\ \omega \in \Omega\}$$

is integrable on $[0,T]$. We say that $\{\nu_k\} \to \nu$ weakly in $\mathcal{S}$ if, for every $T > 0$, the sequence $\left\{\nu_k|_{[0,T]}\right\}$ of restrictions of ν_k to $[0,T]$ converges to $\nu|_{[0,T]}$ in $\mathcal{S}_T$.

Notice that, since every element of Ω can be identified with the δ-measure, concentrated in it, Ω can be embedded into $\mathcal{P}(\Omega)$, $\mathcal{M}_\Omega$ into $\mathcal{S}$, and $\mathcal{M}_\Omega^T$ into $\mathcal{S}_T$ in the obvious way, where $\mathcal{M}_\Omega^T$ is the set of functions in $\mathcal{M}_\Omega$ restricted to $[0,T]$.

For each $\nu \in \mathcal{P}(\Omega)$, we denote

$$f(x,\nu) \;=\; \int_\Omega f(x,r)\, d\nu(r)\,. \tag{4.11}$$

Notice that for any relaxed control $\nu(\cdot)$, the function $f(\cdot,\nu(\cdot)) : (x,t) \to f(x,\nu(t))$ is Lipschitz in x and measurable in t. Moreover, as for all $x \in \mathbf{X}$ and all $\nu \in \mathcal{P}(\Omega)$, we have a bound

$$|f(x,\nu)| \leq \max_{d\in\Omega} |f(x,d)|\,,$$

the solution of the "system" $\dot{x}(t) = f(x(t),\nu(t))$ exists for any initial condition ξ and relaxed control ν on some maximal interval $[0, t_{\max}(\xi,\nu))$. Write $x(\cdot,\xi,\nu)$ to denote this solution. Just as in the case with ordinary controls, we will define

$$\lambda_{\xi,\nu} := \inf\{t \leq t_{\max}\ :\ x(s,\xi,\nu) \in \mathcal{D}\}\,. \tag{4.12}$$

The basic three facts we will be using in the next section are as follows.

Fact 1. For any $T > 0$, the space $\mathcal{S}_T$ is sequentially compact (see [51, Thm IV.2.1, p. 272]). Consequently, $\mathcal{S}$ is sequentially compact by a diagonalization argument.

Fact 2. For any $T > 0$, the set $\mathcal{M}_\Omega^T$ of ordinary controls on $[0,T]$ is dense in $\mathcal{S}_T$ (see [5, p. 691], also [51]). Consequently, $\mathcal{M}_\Omega$ is dense in $\mathcal{S}$. The topology of $\mathcal{S}$

induces a topology on the subspace $\mathcal{M}_\Omega$. This topology is stronger than the topology of L^p for any positive p: in fact, $\{\mathbf{d}_k(t)\} \to \mathbf{d}$ would imply

$$\int_0^T g(\mathbf{d}_k(t))dt \to \int_0^T g(\mathbf{d}(t))\,dt$$

for any continuous function $g : \mathbb{U} \to \mathbf{R}$ and any positive T.

Fact 3. For any $T > 0$, the mapping $(t, \xi, \nu(\cdot)) \mapsto z(t, \xi, \nu)$ is continuous on $[0, T] \times \mathbf{X} \times \mathcal{S}_T$ (see [1, Lemma 3.12]).

Some relevant immediate consequences of Facts 1, 2, and 3 are as follows.

Claim 1. The function $(\xi, \nu) \to \lambda_{\xi,\nu}$, mapping initial states and disturbances to the first hitting times as defined in (4.12), is lower semicontinuous on both ν and ξ. (This easily follows from continuous dependence on initial conditions and control, and from the fact that the set $\mathcal{D}$ is closed.)

Claim 2. If system (2.14) is GASMO, then, for any initial value $\xi \in \mathcal{E}$, a relaxed disturbance $\nu \in \mathcal{S}$, and time $T \leq \lambda_{\xi,\nu}$ we have an estimate

$$x(T, \xi, \nu) \leq \lambda(|\xi|, T).$$

Proof. Pick $\xi \in \mathbf{X}$ and $\nu \in \mathcal{S}$. Let $\{\mathbf{d}_k\}$ be a sequence of ordinary disturbances, converging to ν in $\mathcal{S}$. Then, for large enough k we have $t_{\max}(\xi, \mathbf{d}_k) > T$, and $x(\cdot, \xi, \mathbf{d}_k)$ converge to $x(\cdot, \xi, \nu)$ uniformly on $[0, T]$. Also, by Claim 1, $\liminf_{k\to\infty} \lambda_{\xi,\mathbf{d}_k} \geq \lambda_{\xi,\nu}$. Since we assume the system to be GASMO, the estimate (2.17) holds for $\mathbf{d} := \mathbf{d}_k$ for all large enough k and for all $t \leq T$, so the claim follows. □

With the previous claim in mind, we can say that the system (2.14) with $d \in \mathcal{S}$ is GASMO (this is a slight abuse of terminology because, strictly speaking, (2.14) with relaxed disturbances is not really a "system," as that would mean, by definition, that disturbances take values in a finite dimensional space). Consequently, the auxiliary system (4.3) is also GASMO for $\mathbf{d} \in \mathcal{S}$. Thus, we can assume that (4.8) holds for all $\xi \in \mathcal{E}$ and all $\mathbf{d} \in \mathcal{S}$.

4.3. Constructing a continuous Lyapunov-like function on $\mathcal{E}_1$. In section 4.1 we introduced the system $\hat{\Sigma} : \dot{z} = \hat{f}(z, d) := \frac{f(z,d)}{1+|f(z,d)|^2+\kappa(z)}$, which slows down the motions of the original system Σ. Recall also that $\mathcal{D} := \{\xi : |\xi| \leq \rho(|h(\xi)|)\}$, and define the set

$$\mathcal{B} := \{\xi : \rho(|h(\xi)|) \leq |\xi| \leq 1.5\rho(|h(\xi)|)\}.$$

Let $f_0 : \mathbf{X} \to \mathbf{R}$ be defined by

$$f_0(\xi) = \max_{\mathrm{d}\in\Omega} \left|\widehat{f}(\xi, \mathrm{d})\right|.$$

Note that f_0 is locally Lipschitz, and recall that we have assumed with no loss of generality that Σ has no equilibria on the set $\{x \in \mathbf{X} : |x| \geq \rho(|h(x)|)\}$ (otherwise replace $\rho(\cdot)$ by $c\rho(\cdot)$, where $c > 1$). In particular, $f_0(\xi) \neq 0$ for any $\xi \in \partial\mathcal{D} \setminus \{0\}$. Let $\phi : \mathbf{X} \setminus \{0\} \to [0, 1]$ be smooth and

$$\phi(x) = \begin{cases} 1, & x \in \mathcal{D}, \\ 0, & x \in \mathbf{X} \setminus (\mathcal{D} \cup \mathcal{B}). \end{cases}$$

Now introduce another system, on the state space $\mathbf{X} \setminus \{0\}$,

$$\widetilde{\Sigma} : \quad \dot{z} = \widetilde{f}(z, d, v),$$

where disturbances d are as before, and auxiliary controls $\mathbf{v}$ are measurable functions of time, taking values in $[-1,1]^n$ (note that the dimension of the control set for v's is the same as that of $\mathbf{X}$), where, for each $i = 1, 2, \ldots, n$,

$$\widetilde{f}_i(z, d, v) := \widehat{f}_i(z, d) + 2\phi(z) f_0(z) v_i.$$

For each $T > 0$, let $\mathcal{W}_T$ denote the set of the auxiliary controls defined on $[0, T]$ (i.e., measurable functions $\mathbf{v} : [0, T] \to [-1, 1]^n$) equipped with the weak convergence topology; that is, "$\{\mathbf{v}_k\}$ converges weakly to $\mathbf{v}$ in $\mathcal{W}_T$" means

$$\int_0^T \varphi(s)\mathbf{v}_k(s)\, ds \;\to \int_0^T \varphi(s)\mathbf{v}(s)\, ds$$

for all functions φ that are integrable over $[0, T]$. With the weak topology, $\mathcal{W}_T$ is sequentially compact (cf. [39, Proposition 10.1.5]). Consequently, given any sequence $\{\mathbf{v}_k\}$ of controls defined on $[0, \infty)$, there exist some control $\mathbf{v}$ and a subsequence $\{\mathbf{v}_{k_j}\}$ such that $\mathbf{v}_{k_j} \to \mathbf{v}$ weakly on every interval $[0, T]$.

We denote the set of the auxiliary controls defined on $[0, \infty)$ by $\mathcal{W}$. Let $\{\mathbf{v}_k\} \subset \mathcal{W}$ and $\mathbf{v} \in \mathcal{W}$. We say that $\{\mathbf{v}_k\}$ weakly converges to $\mathbf{v}$ if, for every $T > 0$, the sequence of restrictions $\{\mathbf{v}_k|_{[0,T]}\}$ weakly converges to $\mathbf{v}|_{[0,T]}$ in $\mathcal{W}_T$.

Recall that $z(t, \xi, \mathbf{d})$ denotes the solution of $\hat{\Sigma}$. Write $z(t, \xi, \mathbf{d}, \mathbf{v})$ for the value at time t of the solution of $\widetilde{\Sigma}$ with initial state $\xi \neq 0$, disturbance $\mathbf{d} \in \mathcal{M}_\Omega$, and auxiliary control $\mathbf{v} \in \mathcal{W}$.

Observe the following.

- $\widetilde{\Sigma}$ is affine in v.
- since for any $\xi \in \mathbf{X}$ and $d \in [-1, 1]^m$, $|\hat{f}(\xi, d)| \le 1$, we have $|\widetilde{f}(\xi, d, v)| \le 3$ for any ξ, d and v. In particular, this implies that $\widetilde{\Sigma}$ is forward complete.
- Suppose $\xi \notin \mathcal{D} \cup \mathcal{B}$ and pick $\mathbf{d} \in \mathcal{M}_\Omega$ and $\mathbf{v} \in \mathcal{W}$. Then there is some $t_0 > 0$ such that $z(t, \xi, \mathbf{d}) \notin \mathcal{D} \cup \mathcal{B}$ for all $t \in [0, t_0]$, and $z(t, \xi, \mathbf{d}, \mathbf{v}) \equiv z(t, \xi, \mathbf{d})$ on $[0, t_0]$.

To extend the definition of $z(t, \xi, \mathbf{d}, \mathbf{v})$ to the case when $\mathbf{d} \in \mathcal{S}$, we let, for a fixed $\mathbf{d} \in \mathcal{S}$ and $\mathbf{v} \in \mathcal{W}$,

$$g_{\mathbf{d},\mathbf{v}}(z, t) := \hat{f}(z, \mathbf{d}(t)) + 2\phi(z) f_0(z)\mathbf{v}(t)$$

(where $f(z, \mathbf{d}(t))$ is as defined in (4.11) for $\nu := \mathbf{d}(t) \in \mathcal{P}(\Omega)$). Then $g_{\mathbf{d},\mathbf{v}} : \mathbf{X} \setminus \{0\} \times \mathbf{R}_{\ge 0} \to \mathbf{X}$ is locally Lipschitz in its first variable, and $|g_{\mathbf{d},\mathbf{v}}(z, t)| \le 3$ for all $(z, t) \in \mathbf{X} \setminus \{0\} \times \mathbf{R}_{\ge 0}$. Hence, the solution of

$$\begin{aligned} \dot{z}(t) &= g_{\mathbf{d},\mathbf{v}}(z, t), \\ z(0) &= \xi \end{aligned}$$

exists for all $\xi \in \mathbf{X} \setminus \{0\}$ and $t \ge 0$. We will denote it by $z(t, \xi, \mathbf{d}, \mathbf{v})$.

Observe that $\widetilde{f} : \mathbf{X} \setminus \{0\} \times \Omega \times [-1, 1]^n \to \mathbf{R}^n$ is continuous, and $\widetilde{f}(z, d, v)$ is locally Lipschitz in z on $\mathbf{X} \setminus \{0\}$ uniformly on $(d, v) \in \Omega \times [-1, 1]^n$. The system $\widetilde{\Sigma}$ evolves in the state space $\mathbf{X} \setminus \{0\}$. As $|\widetilde{f}| \le 3$ everywhere, trajectories are defined and unique for each initial value $\xi \in \mathbf{X} \setminus \{0\}$ and each pair of inputs $\mathbf{d}, \mathbf{v}$. Moreover, if $z(\cdot)$ is a maximal such trajectory, then either $z(t)$ is defined for all $t \ge 0$, or there is some $T > 0$ such that $\lim_{t \to T} z(t) = 0$. We prove next that this last case cannot happen.

Lemma 4.1. *For every ball U around 0, there is a constant c such that for any $\xi \in U$, $\xi \neq 0$, $\mathbf{d} \in \mathcal{S}$, $\mathbf{v} \in \mathcal{W}$, and $t \geq 0$, we have a lower bound*

$$|z(t, \xi, \mathbf{d}, \mathbf{v})| > \frac{1}{2} |\xi| \, e^{-ct}. \tag{4.13}$$

Proof. Since $\hat{f}$ is locally Lipschitz in x uniformly in d, and $\hat{f}(0, d) = 0$ for all d, we can find a positive constant c, such that $|\hat{f}(z, d)| \leq c|z|/3$ for all $z \in U$ and all $d \in \Omega$. Then also $|f_0(z)| \leq c|z|/3$; therefore

$$\left|\widetilde{f}(z, d, v)\right| \leq \left|\hat{f}(z, d)\right| + 2 \, |f_0(z)| \leq c|z|$$

for all $v \in [-1, 1]^n$, all $d \in \Omega$, and hence, all $d \in \mathcal{P}(\Omega)$. Now fix $\xi \in U \setminus \{0\}$, $\mathbf{d} \in \mathcal{S}$, and $\mathbf{v} \in \mathcal{W}$, and write $z(t) := z(t, \xi, \mathbf{d}, \mathbf{v})$. Note that the inequality (4.13) holds for $t = 0$; therefore it holds for all small enough $t > 0$. Suppose that (4.13) fails at some $t_2 > 0$, so that

$$|z(t_2)| \leq \frac{1}{2} |\xi| \, e^{-ct_2} < |\xi| \, . \tag{4.14}$$

Then there exists a $t_1 < t_2$ such that $|z(t_1)| = |\xi|$ and $|z(t)| \leq |\xi|$ for all $t \in [t_1, t_2]$. Let

$$w(t) := |z(t)|^2 /2.$$

Then, for almost all $t \in [t_1, t_2]$

$$|\dot{w}(t)| = |z(t) \cdot \dot{z}(t)| \leq c \, |z(t)|^2 = 2cw(t).$$

In particular, this implies that $\dot{w}(t) + 2cw(t) \geq 0$. So, for all $t \in [t_1, t_2]$ we have

$$0 \leq e^{2ct}(\dot{w}(t) + 2cw(t)) = \frac{d(e^{2ct}w(t))}{dt},$$

implying that $e^{2ct}w(t) \geq e^{2ct_1}w(t_1)$ for all $t \in [t_1, t_2]$. Thus,

$$\frac{1}{2} |z(t_2)|^2 = w(t_2) \geq e^{2ct_1}w(t_1)e^{-2ct_2} = \frac{1}{2} |z(t_1)|^2 e^{-2c(t_2 - t_1)} \geq \frac{1}{2} |\xi|^2 e^{-2ct_2},$$

so that $|z(t_2)| \geq e^{-ct_2} |\xi|$, contradicting (4.14). □

Corollary 4.2. *For every $r > 0$ and $T > 0$ there is a $\sigma = \sigma(r, T) > 0$ such that for any $\mathbf{d} \in \mathcal{S}$, $\mathbf{v} \in \mathcal{W}$, $|\xi| \geq r$, and $t \leq T$ we have*

$$|z(t, \xi, \mathbf{d}, \mathbf{v})| \geq \sigma.$$

Lemma 4.3. *For each $\mathbf{d} \in \mathcal{S}_T$ and each $\mathbf{v} \in \mathcal{W}_T$, if $\xi_k \to \xi$ in $\mathbf{X} \setminus \{0\}$, $\mathbf{d}_k \to \mathbf{d}$ in $\mathcal{S}_T$, and $\mathbf{v}_k \to \mathbf{v}$ in $\mathcal{W}_T$, $\{z(t, \xi_k, \mathbf{d}_k, \mathbf{v}_k)\}$ converges to $z(t, \xi, \mathbf{d}, \mathbf{v})$ uniformly on $[0, T]$.*

Proof. Assume without loss of generality that $\xi_k \in U := B_{\frac{|\xi|}{2}}(\xi)$. Since $|\widetilde{f} \leq 3|$ and by Corollary 4.2, $\mathcal{R}_T\,(U) \subseteq B_{1.5|\xi|+3T}(0) \setminus B_\sigma(0)$, where $\sigma = \sigma(|\xi|\,/2, T)$ is as in Corollary 4.2. Let M_1 and M_2 be Lipschitz constants for $\hat{f}(\cdot, d)$ (uniformly for

$d \in \Omega$) and $2f_0\,\phi$, respectively, on $B_{1.5|\xi|+3T}(0) \setminus B_\sigma(0)$. Write $z(t) := z(t,\xi,\mathbf{d},\mathbf{v})$, and $z_k(t) := z(t,\xi_k,\mathbf{d}_k,\mathbf{v}_k)$. Now, for all $t \in [0,T]$ we have

$$\begin{align}
|z(t)-z_k(t)| &= \left|\xi_k - \xi + \int_0^t \left(\widetilde{f}(z_k(t),\mathbf{d}_k(t),\mathbf{v}_k(t)) - \widetilde{f}(z(t),\mathbf{d}(t),\mathbf{v}(t))\right) dt\right| \notag\\
&\le |\xi_k - \xi| + \left|\int_0^t \left(\widetilde{f}(z(t),\mathbf{d}_k(t),\mathbf{v}_k(t)) - \widetilde{f}(z(t),\mathbf{d}(t),\mathbf{v}(t))\right) dt\right| \notag\\
&\qquad + \int_0^t \left|\widetilde{f}(z_k(t),\mathbf{d}_k(t),\mathbf{v}_k(t)) - \widetilde{f}(z(t),\mathbf{d}_k(t),\mathbf{v}_k(t))\right| dt \notag\\
&\le |\xi_k - \xi| + \left|\int_0^t \left(\hat{f}(z(t),\mathbf{d}_k(t)) - \hat{f}(z(t),\mathbf{d}(t))\right) dt\right| \tag{4.15}\\
&\qquad +2\left|\int_0^t \left(f_0(z(t))\phi(z(t))\mathbf{v}_k(t) - f_0(z(t))\phi(z(t))\mathbf{v}(t)\right) dt\right| \tag{4.16}\\
&\qquad + \int_0^t (M_1+M_2)\,|z_k(t)-z(t)|\,dt. \notag
\end{align}$$

The integrals in (4.15) and (4.16) tend to 0 because of the convergence of $\{\mathbf{d}_k(\cdot)\}$ to $\mathbf{d}$ in $\mathcal{S}$ and the weak convergence of $\mathbf{v}_k$ to $\mathbf{v}$. So, for any $\varepsilon > 0$ we can find a K such that, for all $k \ge K$,

$$\begin{aligned}
|\xi_k - \xi| &+ \left|\int_0^T \left(\hat{f}(z(t),\mathbf{d}_k(t)) - \hat{f}(z(t),\mathbf{d}(t))\right) dt\right| \\
&+ 2\left|\int_0^t \left(f_0(z(t))\phi(z(t))\mathbf{v}_k(t) - f_0(z(t))\phi(z(t))\mathbf{v}(t)\right) dt\right| \le \varepsilon e^{-(M_1+M_2)T}.
\end{aligned}$$

Then, for all $k \ge K$ and $t \in [0,T]$ we have, by the Gronwall inequality,

$$|z(t)-z_k(t)| \le \varepsilon e^{-(M_1+M_2)T}\, e^{(M_1+M_2)t} \le \varepsilon.$$

As ε was arbitrary, this proves uniform convergence. □

Also, since for any $\xi \in \partial D \setminus \{0\}$, $f_0(\xi) \ne 0$, and $f_0(\xi) \ge |\hat{f}(\xi,d)|$ for any $d \in \Omega$, we have the following controllability property on $\partial\mathcal{D} \setminus \{0\}$.

LEMMA 4.4. *Let $\xi \ne 0$ be on $\partial\mathcal{D}$. Then, for each $\tau > 0$, there exists a neighborhood U of ξ, such that for any $\eta \in U$ and any $\mathbf{d} \in \mathcal{M}_\Omega$, there is some control $\mathbf{v}$, and some $0 \le t_1 \le \tau$, such that $z(t_1,\eta,\mathbf{d},\mathbf{v}) = \xi$ and $z(t,\eta,\mathbf{d},\mathbf{v}) \in U$ for all $0 \le t \le t_1$.*

Proof. Since $\phi(\xi)\,f_0(\xi) \ne 0$ and the function $\phi(\cdot)\,f_0(\cdot)$ is continuous, we can find a ball U_1 centered at ξ and a constant c_1 such that $\phi(z)f_0(z) > c_1$ for every $z \in U_1$. Since $\phi(\xi) = 1$ and ϕ is continuous, one could also find a ball $U_2 \subseteq U_1$ centered at ξ, so that

$$\left|\widehat{f}(z,d)\right| \le 1.5\phi(z)f_0(z) \quad \forall z \in U_2,\ d \in \Omega.$$

Fix $\tau > 0$ and let $B(\xi)$ be the ball of radius $\tau c_1/2$ centered at ξ. Define $U := B(\xi) \cap U_2$. Pick a point $\eta \in U$. Then $|\xi - \eta| < \tau c_1/2$, so,

$$\bar{v}_2 := \frac{2(\xi-\eta)}{\tau c_1}$$

has norm smaller than 1. Consider the "feedback law"

$$k(z,d) = \frac{1}{2}(1.5k_1(z,d) + 0.5\bar{v}_2),$$

where

$$k_1(z,d) := -\frac{\widehat{f}(z,d)}{1.5\phi(z)f_0(z)}.$$

Notice that for all $z \in U$ and $d \in \Omega$ we have $|k_1(z,d)| \leq 1$ and

$$\begin{aligned}\widetilde{f}(z,d,k(z,d)) &= \widehat{f}(z,d) + 2\phi(z)f_0(z)k(z,d)\\ &= \widehat{f}(z,d) + 1.5\phi(z)f_0(z)k_1(z,d) + 0.5\phi(z)f_0(z)\bar{v}_2\\ &= 0.5\phi(z)f_0(z)\bar{v}_2 = \frac{(\xi-\eta)\phi(z)f_0(z)}{\tau c_1}.\end{aligned}$$

Thus, with any initial condition $\eta \in U$ and disturbance $\mathbf{d} \in \mathcal{M}_\Omega$, if the control $v(t) := k(z(t), d(t))$ is applied, then the trajectory of the system $\widetilde{\Sigma}$ will be the line segment, connecting ξ and η, transversed with a velocity greater than $(\xi - \eta)/\tau$. So, there exists a $t_0 \leq \tau$ such that $z(t_0, \eta, \mathbf{d}, \mathbf{v}) = \xi$ and, since U is convex, $z(t, \eta, \mathbf{d}, \mathbf{v}) \in U$ for all $t \geq t_0$. □

For each $\mathbf{d} \in \mathcal{S}$, let

$$\theta_{\mathbf{d}}(\xi, \mathbf{v}) = \inf\{t \geq 0 : z(t, \xi, \mathbf{d}, \mathbf{v}) \in \mathcal{D}\},$$

where as before, $\theta_{\mathbf{d}}(\xi, \mathbf{v}) = \infty$ if the trajectory never reaches $\mathcal{D}$.

LEMMA 4.5. *The map* $(\xi, \mathbf{v}, \mathbf{d}) \mapsto \theta_{\mathbf{d}}(\cdot,\cdot)$ *is lower semicontinuous on* $\mathcal{E} \times \mathcal{W} \times \mathcal{S}$.

Proof. Let $\{\xi_k\} \subset \mathcal{E}$, $\{\mathbf{v}_k\} \subset \mathcal{W}$ and $\{\mathbf{d}_k\} \subset \mathcal{S}$ be such that $\xi_k \to \xi, \mathbf{v}_k \to \mathbf{v}$, and $\mathbf{d}_k \to \mathbf{d}$ for some $\xi \in \mathcal{E}$, $\mathbf{v} \in \mathcal{W}$, and $\mathbf{d} \in \mathcal{S}$. We need to show that

$$\theta_{\mathbf{d}}(\xi, \mathbf{v}) \leq \liminf_{k\to\infty} \theta_{\mathbf{d}_k}(\xi_k, \mathbf{v}_k). \tag{4.17}$$

Let $\theta_k = \theta_{\mathbf{d}_k}(\xi_k, \mathbf{v}_k)$. Without loss of generality, we may assume that

$$\liminf_{k\to\infty} \theta_k = \theta_0 < \infty.$$

Passing to a subsequence if necessary, we assume that $\theta_k \to \theta_0$. Thus, there exists some K such that $\theta_k \leq \theta_0 + 1$ for all $k \geq K$. Since $\{z(t, \xi_k, \mathbf{d}_k, \mathbf{v}_k)\}$ converges to $z(t, \xi, \mathbf{d}, \mathbf{v})$ uniformly on $[0, \theta_0 + 1]$, it follows that

$$z(\theta_0, \xi, \mathbf{d}, \mathbf{v}) = \lim_{k\to\infty} z(\theta_k, \xi_k, \mathbf{d}_k, \mathbf{v}_k).$$

Since $\mathcal{D}$ is closed and $z(\theta_k, \xi_k, \mathbf{d}_k, \mathbf{v}_k) \in \mathcal{D}$ for each k, we know that $z(\theta_0, \xi, \mathbf{d}, \mathbf{v}) \in \mathcal{D}$, and hence, $\theta_{\mathbf{d}}(\xi, \mathbf{v}) \leq \theta_0$. □

Define, for $\xi \in \mathcal{E}, \mathbf{d} \in \mathcal{S}$, and $\mathbf{v} \in \mathcal{W}$,

$$V_{\mathbf{v}}(\xi, \mathbf{d}) := \int_0^{\theta_{\mathbf{d}}(\xi,\mathbf{v})} \Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|)\, dt$$

(where Ξ is as in (4.10)) and, for $\xi \in \mathcal{E}$ and $\mathbf{d} \in \mathcal{S}$,

$$\widetilde{V}_0(\xi, \mathbf{d}) := \inf_{\mathbf{v}\in\mathcal{W}} V_{\mathbf{v}}(\xi, \mathbf{d}).$$

Note that for some $\mathbf{v}$ and $\mathbf{d}$, $V_{\mathbf{v}}(\xi, \mathbf{d})$ may take ∞ as its value, but $\widetilde{V}_0(\xi, \mathbf{d})$ is always finite, since

$$\widetilde{V}_0(\xi, \mathbf{d}) \leq V_O(\xi, \mathbf{d}), \tag{4.18}$$

where $O(\cdot)$ is the control identically equal to 0. Recall that by definition of Ξ we have then, for all $\xi \in \mathcal{E}_1$ and $\mathbf{d} \in \mathcal{S}$,

$$
\begin{aligned}
V_O(\xi, \mathbf{d}) &= \int_0^{\theta_{\mathbf{d}}(\xi, O)} \Xi(|z(t, \xi, \mathbf{d}, O)|)\, dt \\
&= \int_0^{\theta_{\mathbf{d}}(\xi, O)} \Xi(|z(t, \xi, \mathbf{d})|)\, dt \leq \int_0^{\infty} \Xi(\mu_1(\mu_2(|\xi|)e^{-t}))dt \leq \mu_2(|\xi|), \qquad (4.19)
\end{aligned}
$$

where μ_1 and μ_2 are as in (4.9).

LEMMA 4.6. *The function $V_{(\cdot)}(\cdot, \cdot) : \mathcal{E} \times \mathcal{S} \times \mathcal{W} \to \mathbf{R}_{\geq 0}$ is lower semicontinuous.*

Proof. Let $(\xi, \mathbf{d}, \mathbf{v}) \in \mathcal{E} \times \mathcal{S} \times \mathcal{W}$, and let $\{\xi_k\} \to \xi$, $\{\mathbf{d}_k\} \to \mathbf{d}$ and $\{\mathbf{v}_k\} \to \mathbf{v}$, where $\xi_k \in \mathcal{E}$ for all k.

Case 1. $V_{\mathbf{v}}(\xi, \mathbf{d}) < \infty$. In this case, for any $\varepsilon > 0$, there exists some $0 < T < \theta_{\mathbf{d}}(\xi, \mathbf{v})$ such that

$$
V_{\mathbf{v}}(\xi, \mathbf{d}) \ = \ \int_0^{\theta_{\mathbf{d}}(\xi, \mathbf{v})} \Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|)\, dt \ \leq \ \int_0^{T} \Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|)\, dt + \varepsilon.
$$

Without loss of generality we can assume that all ξ_k are within the unit distance from ξ. Recall that the reachable set from the unit ball around ξ is bounded. Since $z(t, \xi_k, \mathbf{d}_k, \mathbf{v}_k)$ converges to $z(t, \xi, \mathbf{d}, \mathbf{v})$ uniformly on $[0, T]$, and $\Xi(\cdot)$ is uniformly continuous on compacts, there exists some $K > 0$ such that

$$
|\Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|) - \Xi(|z(t, \xi_k, \mathbf{d}_k, \mathbf{v}_k)|)| < \frac{\varepsilon}{1+T} \ \ \forall k > K, \ \forall t \in [0, T].
$$

This implies that

$$
\int_0^{T} \Xi(|z(t, \xi_k, \mathbf{d}_k, \mathbf{v}_k)|)\, dt \ \geq \ \int_0^{T} \Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|)\, dt - \varepsilon
$$

for all $k \geq K$. By Lemma 4.5, there exists some $K_1 \geq K$ such that $\theta_{\mathbf{d}_k}(\xi_k, \mathbf{v}_k) > T$ for all $k \geq K_1$. Thus, for all $k \geq K_1$,

$$
\begin{aligned}
V_{\mathbf{v}_k}(\xi_k, \mathbf{d}_k) &\geq \int_0^{T} \Xi(|z(t, \xi_k, \mathbf{d}_k, \mathbf{v}_k)|)\, dt \\
&\geq \int_0^{T} \Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|)\, dt - \varepsilon \ \geq \ V_{\mathbf{v}}(\xi, \mathbf{d}) - 2\varepsilon.
\end{aligned}
$$

As ε was arbitrary, we conclude that

$$
V_{\mathbf{v}}(\xi, \mathbf{d}) \leq \liminf V_{\mathbf{v}_k}(\xi_k, \mathbf{d}_k).
$$

Case 2. $V_{\mathbf{v}}(\xi, \mathbf{d}) = \infty$.

In this case, $\theta_{\mathbf{d}}(\xi, \mathbf{v}) = \infty$. Fix an integer $k \geq 0$. There exists some T_k such that

$$
\int_0^{T_k} \Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|)\, dt \geq k.
$$

Repeating the same argument used above, one sees that

$$
\int_0^{T_k} \Xi(|z(t, \xi_l, \mathbf{d}_l, \mathbf{v}_l)|)\, dt \geq k - 1
$$

for all $l \geq L_0$ for some L_0. By Lemma 4.5, there is some $L_1 \geq L_0$ such that $\theta_{\mathbf{d}_l}(\xi_l, \mathbf{v}_l) \geq T_k$ for all $l \geq L_1$. Consequently, for all $l \geq L_1$,

$$V_{\mathbf{v}_l}(\xi_l, \mathbf{d}_l) \geq \int_0^{T_k} \Xi(|z(t, \xi_l, \mathbf{d}_l, \mathbf{v}_l)|)\, dt \geq k - 1.$$

Since $k > 0$ can be picked arbitrarily, it follows that

$$\liminf V_{\mathbf{v}_l}(\xi_l, \mathbf{d}_l) = \infty.$$

In both cases, we have shown that $\liminf V_{\mathbf{v}_l}(\xi_l, \mathbf{d}_l) \geq V_{\mathbf{v}}(\xi, \mathbf{d})$. The lower semicontinuity property follows readily. □

LEMMA 4.7. *For every $\xi \in \mathcal{E}$, $\mathbf{d} \in \mathcal{S}$, there exists a control $\bar{\mathbf{v}}$ such that*

$$V_{\bar{\mathbf{v}}}(\xi, \mathbf{d}) = \widetilde{V}_0(\xi, \mathbf{d}). \tag{4.20}$$

Proof. Let $\mathbf{v}_k$ be a sequence of controls such that $V_{\mathbf{v}_k}(\xi, \mathbf{d}) \searrow \widetilde{V}_0(\xi, \mathbf{d})$. Without loss of generality we are assuming that all these controls are defined for all positive t (by letting them equal to 0 where they are not defined). Extract from $\{\mathbf{v}_k\}$ a subsequence $\{\mathbf{v}_{k_l}\}$ converging weakly to some limit $\bar{v}$ in $\mathcal{W}$. Without relabeling, we assume that $\mathbf{v}_k \to \bar{\mathbf{v}}$. By Lemma 4.6,

$$V_{\bar{\mathbf{v}}}(\xi, \mathbf{d}) \leq \lim_{k\to\infty} V_{\mathbf{v}_k}(\xi, \mathbf{d}) = \widetilde{V}_0(\xi, \mathbf{d}).$$

Combining this with the fact that $\widetilde{V}_0(\xi, \mathbf{d}) \leq V_{\mathbf{v}}(\xi, \mathbf{d})$ for all $\mathbf{v} \in \mathcal{W}$, one thus proves (4.20). □

COROLLARY 4.8. *For any $\xi \in \mathcal{E}, \mathbf{d} \in \mathcal{S}, \mathbf{v} \in \mathcal{W}$, and $0 \leq T < \theta_{\mathbf{d}}(\xi, \mathbf{v})$, it holds that*

$$\widetilde{V}_0(\xi, \mathbf{d}) \;\leq\; \int_0^T \Xi(|z(s, \xi, \mathbf{d}, \mathbf{v})|)\, ds + \widetilde{V}_0(z(T, \xi, \mathbf{d}, \mathbf{v}), \mathbf{d}_T), \tag{4.21}$$

where $\mathbf{d}_T(t) = \mathbf{d}(t + T)$ for all $t \geq 0$.

Proof. Suppose the assertion is not true. Then there exist $\xi \in \mathcal{E}$, $T > 0$, $\mathbf{v} \in \mathcal{W}$, and $\mathbf{d} \in \mathcal{S}$ such that (4.21) fails. By Lemma 4.7, one can find a control $\mathbf{v}_1$ such that $\widetilde{V}_0(z(T, \xi, \mathbf{d}, \mathbf{v}), \mathbf{d}_T) = V_{\mathbf{v}_1}(z(T, \xi, \mathbf{d}, \mathbf{v}), \mathbf{d}_T)$. Define $\bar{\mathbf{v}}$ to be the concatenation of $\mathbf{v}$ and $\mathbf{v}_1$. Then, letting $\theta := \theta_{\mathbf{d}_T}(z(T, \xi, \mathbf{d}, \mathbf{v}), \mathbf{v}_1)$ and noticing that $\theta_{\mathbf{d}}(\xi, \bar{\mathbf{v}}) = \theta + T$, we get, by our assumption,

$$\begin{aligned}
\widetilde{V}_0(\xi, \mathbf{d}) &> \int_0^T \Xi(|z(s, \xi, \mathbf{d}, \mathbf{v})|)\, ds + \widetilde{V}_0(z(T, \xi, \mathbf{d}, \mathbf{v}), \mathbf{d}_T) \\
&= \int_0^T \Xi(|z(s, \xi, \mathbf{d}, \mathbf{v})|)\, ds + \int_0^\theta \Xi(|z(t, z(T, \xi, \mathbf{d}, \mathbf{v}), \mathbf{d}_T, \mathbf{v}_1)|)\, dt \\
&= \int_0^{\theta+T} \Xi(|z(s, \xi, \mathbf{d}, \bar{\mathbf{v}})|)\, ds,
\end{aligned}$$

which contradicts with the minimality of $\widetilde{V}_0(\xi, \mathbf{d})$. □

LEMMA 4.9. *For each $\xi \in \partial\mathcal{D}$, the following holds:*

$$\lim_{\eta\to\xi} \widetilde{V}_0(\eta, \mathbf{d}) = 0 \tag{4.22}$$

uniformly in $\mathbf{d} \in \mathcal{S}$*, that is, for any* $\varepsilon > 0$*, there is a neighborhood* U *of* ξ *such that* $\widetilde{V}_0(\eta, \mathbf{d}) < \varepsilon$ *for all* $\eta \in U \cap \mathcal{E}$ *and all* $\mathbf{d} \in \mathcal{S}$*.*

Proof. If $\xi = 0$, the result follows from (4.18) and (4.19). Suppose now that $\xi \neq 0$.

Let $\varepsilon > 0$ be given. Let $\tau = \frac{\varepsilon}{\Xi(|\xi|+1)}$. Find a neighborhood U of ξ as in Lemma 4.4. Shrinking U if necessary, we assume that $|\xi - \eta| \leq 1$ for all $\eta \in U$.

Suppose $\eta \in U \cap \mathcal{E}$ and $\mathbf{d} \in \mathcal{M}_\Omega$. By the controllability property, there is some control v such that $z(t_1, \eta, \mathbf{d}, \mathbf{v}) = \xi \in \partial\mathcal{D}$ for some $t_1 \in [0, \tau]$, and that $z(t, \eta, \mathbf{d}, \mathbf{v}) \in U$ for all $t \in [0, t_1]$. Thus,

$$V_{\mathbf{v}}(\eta, \mathbf{d}) \leq \int_0^{t_1} \Xi(|z(s, \eta, \mathbf{d}, \mathbf{v})|)\, ds \leq \tau \cdot \Xi(|\xi| + 1) \leq \varepsilon,$$

from which it follows that $\widetilde{V}_0(\eta, \mathbf{d}) \leq \varepsilon$.

The above shows that $\widetilde{V}_0(\eta, \mathbf{d}) \leq \varepsilon$ for all nonrelaxed $\mathbf{d} \in \mathcal{M}_\Omega$, $\eta \in U \cap \mathcal{E}$. Pick a relaxed $\mathbf{d} \in \mathcal{S}$. Then there exists a sequence $\{\mathbf{d}_k\} \subset \mathcal{M}_\Omega$ such that $\mathbf{d}_k \to \mathbf{d}$ in the topology of relaxed controls. Let $\eta \in U \cap \mathcal{E}$. Let $\mathbf{v}_k$ be such that $\widetilde{V}_0(\eta, \mathbf{d}_k) = V_{\mathbf{v}_k}(\eta, \mathbf{d}_k)$. By weak sequential compactness of $\mathcal{W}$, one may assume, after taking a subsequence, that $\mathbf{v}_k \to \bar{\mathbf{v}}$ for some $\bar{\mathbf{v}}$. By Lemma 4.6,

$$V_{\bar{\mathbf{v}}}(\eta, \mathbf{d}) \leq \liminf_{k \to \infty} V_{\mathbf{v}_k}(\eta, \mathbf{d}_k) \leq \varepsilon,$$

and consequently, $\widetilde{V}_0(\eta, \mathbf{d}) \leq \varepsilon$. This shows that $\widetilde{V}_0(\eta, \mathbf{d}) \leq \varepsilon$ for all $\eta \in U \cap \mathcal{E}$ and all $d \in \mathcal{S}$. □

To prove the continuity of $\widetilde{V}_0$, we also need the following result.

LEMMA 4.10. *Suppose for some* $\xi \in \mathcal{E}$*,* $\mathbf{d} \in \mathcal{S}$*, and* $\mathbf{v} \in \mathcal{W}$*,* $V_{\mathbf{v}}(\xi, \mathbf{d}) < \infty$*. Then there exists some* $\xi_0 \in \partial\mathcal{D}$ *such that*

$$\lim_{t \to \theta_{\mathbf{d}}(\xi, \mathbf{v})} z(t, \xi, \mathbf{d}, \mathbf{v}) = \xi_0. \tag{4.23}$$

Proof. Suppose $V_{\mathbf{v}}(\xi, \nu) < \infty$. This means that

$$\int_0^{\theta_\nu(\xi, \mathbf{v})} \Xi(|z(s, \xi, \nu, \mathbf{v})|)\, ds < \infty. \tag{4.24}$$

If $\theta_{\mathbf{d}}(\xi, \mathbf{v}) < \infty$, then (4.23) follows from the continuity of $z(\cdot, \xi, \nu, \mathbf{v})$ with $\xi_0 = z(\theta_{\mathbf{d}}(\xi, \mathbf{v}))$.

Suppose now that $\theta_{\mathbf{d}}(\xi, \mathbf{v}) = \infty$. Since the integral in (4.24) converges, it follows that

$$\int_t^\infty \Xi(|z(s, \xi, \mathbf{d}, \mathbf{v})|)\, ds \to 0 \quad \text{as} \quad t \to \infty.$$

Consider the family of functions $\{x_t(\cdot),\ t > 0\}$, defined by $x_t(s) := z(t + s, \xi, \mathbf{d}, \mathbf{v})$, $\mathcal{I}_t = [0, \infty)$. By Lemma 4.1, the trajectory $z(s, \xi, \mathbf{d}, \mathbf{v})$ does not reach the origin in finite time, hence, there exists a positive, strictly decreasing function φ such that $\varphi(s) < |z(s, \xi, \mathbf{d}, \mathbf{v})|$ for all $s > 0$. Find a $\mathcal{K}_\infty$-function κ such that

$$\kappa(\varphi(t)) > \int_t^\infty \Xi(|z(s, \xi, \mathbf{d}, \mathbf{v})|)\, ds = \int_0^\infty \Xi(|x_t(s)|)\, ds.$$

Then the family $\{x_t(\cdot),\ t > 0\}$ satisfies all the conditions of Proposition 3.9 (with $\chi := \Xi$). Take $r := |\xi|$. Then, for any $\varepsilon > 0$, $|z(t, \xi, \mathbf{d}, \mathbf{v})| < \varepsilon$ for all $t > T_{r,\varepsilon}$. So, the conclusion of the lemma follows. □

Proposition 4.11. *The function* $\widetilde{V}_0 : \mathcal{E} \times \mathcal{S} \to \mathbf{R}$ *is continuous.*

Proof. Fix $\xi \in \mathcal{E}$, $\mathbf{d} \in \mathcal{S}$. Suppose $\xi_k \to \xi, \mathbf{d}_k \to \mathbf{d}$, where $\xi_k \in \mathcal{E}$. Let $\{k_j\}$ be a subsequence of $\{k\}$ such that

$$\lim_{j\to\infty} \widetilde{V}_0(\xi_{k_j}, \mathbf{d}_{k_j}) = \liminf_{k\to\infty} \widetilde{V}_0(\xi_k, \mathbf{d}_k). \tag{4.25}$$

For each k, let $\mathbf{v}_k$ be such that $\widetilde{V}_0(\xi_k, \mathbf{d}_k) = V_{\mathbf{v}_k}(\xi_k, \mathbf{d}_k)$. Notice that $\lim_{k\to\infty} V_{\mathbf{v}_k}(\xi_k, \mathbf{d}_k)$ exists, because of (4.25).

By sequential compactness of $\mathcal{W}$, there exists a subsequence of $\{\mathbf{v}_{k_j}\}$ converging to some $\bar{\mathbf{v}} \in \mathcal{W}$. Without relabeling, we assume that $\mathbf{v}_{k_j} \to \bar{\mathbf{v}}$. It then follows from Lemma 4.6 that

$$V_{\bar{\mathbf{v}}}(\xi, \mathbf{d}) \leq \lim\, V_{\mathbf{v}_{k_j}}(\xi_{k_j}, \mathbf{d}_{k_j}) = \liminf\, \widetilde{V}_0(\xi_k, \mathbf{d}_k).$$

Consequently, $\widetilde{V}_0(\xi, \mathbf{d}) \leq \liminf\, \widetilde{V}_0(\xi_k, \mathbf{d}_k)$. To complete the proof, we will show that

$$\widetilde{V}_0(\xi, \mathbf{d}) \geq \limsup_{k\to\infty} \widetilde{V}_0(\xi_k, \mathbf{d}_k). \tag{4.26}$$

Let $\mathbf{v}$ be a control such that $\widetilde{V}_0(\xi, \mathbf{d}) = V_{\mathbf{v}}(\xi, \mathbf{d})$. Let $\varepsilon > 0$ be given. By Lemma 4.10, there is some $\xi_0 \in \partial\mathcal{D}$ such that (4.23) holds. By Lemma 4.9, there is a neighborhood U of ξ_0 such that

$$\widetilde{V}_0(\eta, \nu) < \varepsilon/4 \ \ \forall \eta \in U \cap \mathcal{E},\ \forall \nu \in \mathcal{S}. \tag{4.27}$$

Let $0 < T < \theta_{\mathbf{d}}(\xi, \mathbf{v})$ be such that $z(T, \xi, \mathbf{d}, \mathbf{v}) \in U$. Then, since $\{z(t, \xi_k, \mathbf{d}_k, \mathbf{v})\}$ converges to $z(t, \xi, \mathbf{d}, \mathbf{v})$ uniformly on $[0, T]$, it follows that $z(T, \xi_k, \mathbf{d}_k, \mathbf{v}) \in U$ for $k \geq K_1$ for some K_1. By Lemma 4.5, one may assume that $T < \theta_{\mathbf{d}_k}(\xi_k, \mathbf{v})$ for all $k \geq K_1$. Consequently, $\eta = z(T, \xi_k, \mathbf{d}_k, \mathbf{v})$ is also in $\mathcal{E}$, so, applying (4.27) with $\nu = (\mathbf{d}_k)_T$, we have

$$\widetilde{V}_0(z(T, \xi_k, \mathbf{d}_k, \mathbf{v}), (\mathbf{d}_k)_T) < \varepsilon/4 \qquad \forall\, k \geq K_1,$$

where $(\mathbf{d}_k)_T(t) = \mathbf{d}_k(T+t)$. By the uniform convergence property of $\{z(t, \xi_k, \mathbf{d}_k, \mathbf{v})\}$, it follows that there is some compact set $\mathcal{K}$ such that $z(t, \xi_k, \mathbf{d}_k, \mathbf{v}) \in \mathcal{K}$ for all k and all $t \in [0, T]$. Using also the uniform continuity of $\Xi(\cdot)$ on compacts, one sees that there is some $K_2 \geq K_1$ such that

$$\int_0^T \Xi(|z(s, \xi_k, \mathbf{d}_k, \mathbf{v})|)\, ds \ \leq\ \int_0^T \Xi(|z(s, \xi, \mathbf{d}, \mathbf{v})|)\, ds + \varepsilon/2$$

for all $k \geq K_2$. Thus, (4.21) implies

$$\begin{aligned}\widetilde{V}_0(\xi_k, \mathbf{d}_k) &\leq \int_0^T \Xi(|z(s, \xi_k, \mathbf{d}_k, \mathbf{v})|)\, ds + \widetilde{V}_0(z(T, \xi_k, \mathbf{d}_k, \mathbf{v}), (\mathbf{d}_k)_T)\\ &\leq V_{\mathbf{v}}(\xi, \mathbf{d}) + \varepsilon/2 + \varepsilon/2 \ =\ \widetilde{V}_0(\xi, \mathbf{d}) + \varepsilon\end{aligned}$$

for all $k \geq K_2$. From this it follows that

$$\widetilde{V}_0(\xi, \mathbf{d}) \geq \limsup\, \widetilde{V}_0(\xi_k, \mathbf{d}_k) - \varepsilon.$$

Letting $\varepsilon \to 0$, one proves (4.26). □

For each $\xi \in \mathcal{E}$, define

$$V_0(\xi) := \sup_{\mathbf{d}\in\mathcal{M}_\Omega} \widetilde{V}_0(\xi,\mathbf{d}). \tag{4.28}$$

Note that the supremum is finite for each $\xi \in \mathcal{E}$; in fact, by (4.18) and (4.19), we have the upper bound

$$V_0(\xi) \le \mu_2(|\xi|) \qquad \forall\, \xi \in \mathcal{E}, \tag{4.29}$$

where μ_2 is as in inequality (4.9). Also observe that the same function V_0 results if the supremum in (4.28) is taken over the set $\mathcal{S}$. This follows from the fact that $\mathcal{M}_\Omega$ is dense in $\mathcal{S}$ and the continuity property of $\widetilde{V}_0$.

By sequential compactness of $\mathcal{S}$ and by continuity of $\widetilde{V}_0$, we get the following.

Corollary 4.12. *For any ξ in $\mathcal{E}$ there exists a (possibly relaxed) disturbance* $\mathbf{d}$ *such that $V_0(\xi) = \widetilde{V}_0(\xi,\mathbf{d})$, that is, $V_0(\xi) = \max_{\bar{\mathbf{d}}\in\mathcal{S}} \widetilde{V}_0(\xi,\bar{\mathbf{d}})$. Moreover, $V_0 : \mathcal{E} \to \mathbf{R}$ is continuous.*

Proof. Fix $\xi \in \mathcal{E}$. By definition of V_0, there exists a sequence of disturbances $\{\mathbf{d}_k(\cdot)\}$ such that $\widetilde{V}_0(\xi,\mathbf{d}_k) \nearrow V_0(\xi)$. By sequential compactness of $\mathcal{S}$, we can extract a subsequence $\{\mathbf{d}_{k_i}\}$, converging in $\mathcal{S}$ to some $\mathbf{d}$. By continuity of $\widetilde{V}_0$,

$$V_0(\xi) = \lim_{i\to\infty} \widetilde{V}_0(\xi,\mathbf{d}_{k_i}) = \widetilde{V}_0(\xi,\mathbf{d}),$$

proving the first statement of the corollary.

Now fix $\xi \in \mathcal{E}$. Take any sequence $\{\xi_k\} \in \mathcal{E}$, converging to ξ and such that the sequence $\{V_0(\xi_k)\}$ converges. Let $\mathbf{d}_k(\cdot)$ and $\mathbf{d} \in \mathcal{S}$ be maximizing disturbances for ξ_k and ξ, respectively, that is,

$$V_0(\xi_k) = \widetilde{V}_0(\xi_k,\mathbf{d}_k) \text{ and } V_0(\xi) = \widetilde{V}_0(\xi,\mathbf{d}).$$

Extracting a subsequence if necessary, let $\hat{\mathbf{d}}$ be a limit of $\{\mathbf{d}_k\}$ in $\mathcal{S}$. Then, by continuity of $\widetilde{V}_0$ and by definition of V_0, we have

$$V_0(\xi) \;\ge\; \widetilde{V}_0(\xi,\hat{\mathbf{d}}) \;=\; \lim_{k\to\infty} \widetilde{V}_0(\xi_k,\mathbf{d}_k) \;=\; \lim_{k\to\infty} V_0(\xi_k).$$

Consequently, if $\{\xi_i\} \subset \mathcal{E}$ is any sequence, converging to ξ, then

$$V_0(\xi) \;\ge\; \limsup_{i\to\infty} V_0(\xi_i),$$

proving upper semicontinuity of V_0.

On the other hand, again by continuity of $\widetilde{V}_0$, we have

$$\liminf_{k\to\infty} V_0(\xi_k) \ge \lim_{k\to\infty} \widetilde{V}_0(\xi_k,\mathbf{d}) = V_0(\xi),$$

showing the lower semicontinuity of V_0. Thus we conclude that V_0 is continuous on $\mathcal{E}$. □

Lemma 4.13. *There exists a $\mathcal{K}_\infty$-function $\underline{\alpha}$ such that*

$$\underline{\alpha}(|\xi|) \le V_0(\xi)$$

for all $\xi \in \mathcal{E}_1$.

Notice that if $|\xi| \ge 1.6\rho(|h(\xi)|)$ and $\xi \ne 0$, then $\xi \in \mathcal{E}$, because $|\xi| > |\xi|/1.6 \ge \rho(|h(\xi)|)$. To prove Lemma 4.13, we first prove a technical lemma.

Lemma 4.14. *Suppose $\Sigma : \dot{z} = g(z,d)$, $y = h(z)$ is a system of type* (2.14)*, and $p(\cdot)$ is a smooth function of class $\mathcal{K}_\infty$, such that the following conditions hold:*

- *$|g(\xi, d)| \leq 1$ for all $\xi \in \mathbf{X}$ and $d \in \Omega$,*
- *$|\nabla(p \circ |h|)(\xi) \cdot g(\xi, d)| \leq 1$ for all $d \in \Omega$ and all ξ with $|h(\xi)| \geq 1$,*
- *$p(s) \geq s$ for all $s > 0$.*

Pick any constant $a > 0$ and define $K_0 = p(1) + (2+a)/a + 1$. Then for each $\xi \in \mathbf{X}$ such that

$$|\xi| \geq (1+a)p(|h(\xi)|) \text{ and } |\xi| \geq K_0,$$

it holds that

$$|z(t, \xi, \mathbf{d})| > p(|h(z(t, \xi, \mathbf{d}))|)$$

for all $t \in [0, 1)$ and any $\mathbf{d} \in \mathcal{M}_\Omega$.

Proof. Fix a, ξ, and $\mathbf{d}$ as in the formulation of the lemma, and define

$$\theta := \min\{t \,:\, |z(t, \xi, \mathbf{d})| \leq p(|h(z(t, \xi, \mathbf{d}))|)\}$$

with the convention $\theta = +\infty$ if the inequality never holds for $t \geq 0$. Assume the lemma is false, so that $\theta < 1$.

Let $\eta = z(\theta, \xi, \mathbf{d})$, and let $\hat{\mathbf{d}}$ be the shift of $\mathbf{d}$ by θ, that is, $\hat{\mathbf{d}}(t) = \mathbf{d}(t + \theta)$. Since $|g(z, d)| \leq 1$ for all $z \in \mathbf{X}$ and all $d \in \Omega$, it holds that $|\eta| \geq |\xi| - \theta \geq K_0 - 1$. By the definitions of η and θ, one has

$$p(|h(\eta)|) = |\eta| \geq K_0 - 1, \tag{4.30}$$

so also $|h(\eta)| \geq p^{-1}(K_0 - 1) > 1$. Thus, $|h(z(s, \eta, \hat{\mathbf{d}}))| > 1$ for all s near zero.

Claim. $|h(z(s, \eta, \hat{\mathbf{d}}))| > 1$ for all $s \in [-1,\, 0]$.

Assume the claim is false. Then there must exist some $-1 \leq s_0 < 0$ so that

$$s_0 = \max\left\{s \leq 0 :\ |h(z(s, \eta, \hat{\mathbf{d}}))| \leq 1\right\}.$$

We have that for each $s \in (s_0,\, 0]$, $|h(z(s, \eta, \hat{\mathbf{d}}))| > 1$.

Recall that $|\nabla(p \circ |h|)(z) f(z, d)| \leq 1$ for all z with $|h(z)| \geq 1$ and all $d \in \Omega$.

Thus

$$\left|\frac{d}{ds} p\left(\left|h(z(s, \eta, \hat{\mathbf{d}}))\right|\right)\right| \leq 1 \quad \forall\, s \in (s_0,\, 0].$$

This, in turn, implies that

$$p\left(\left|h(z(s_0, \eta, \hat{d}))\right|\right) \geq p(|h(\eta)|) + s_0 \geq K_0 + s_0 - 1 > p(1),$$

and so, since p is strictly increasing, $|h(z(s_0, \eta, \hat{\mathbf{d}}))| > 1$, thus contradicting the definition of s_0. This proves the claim.

It follows from the claim that $|h(z(s, \xi, \mathbf{d}))| > 1$ for all $s \in [0, \theta]$. Thus,

$$\begin{aligned} p(|h(\eta)|) = |\eta| \ &\geq\ |\xi| - \theta \\ &\geq (1+a)p(|h(\xi)|) - \theta \geq (1+a)p(|h(\eta)|) - (1+a)\theta - \theta. \end{aligned}$$

(The last inequality used the fact that $\left|\frac{d}{ds} p(|h(z(s, \xi, \mathbf{d}))|)\right| \leq 1$ for all $s \in [0, \theta]$.) It follows that $p(|h(\eta)|) \leq \frac{2+a}{a}\theta$, so from (4.30) we know that

$$K_0 \leq 1 + p(|h(\eta)|) \leq 1 + \frac{2+a}{a}\theta,$$

contradicting the choice of K_0. This shows that it is impossible to have $\theta < 1$. □

Proof of Lemma 4.13. Recall that if $\xi \notin \mathcal{D} \cup \mathcal{B}$, then $\widetilde{f}(\xi, d, v) = \hat{f}(\xi, d)$ for any d and v, so that

$$\left|\nabla(1.5\rho \circ |h|)(\xi) \cdot \widetilde{f}(\xi, d, v)\right| = \frac{1.5\,|\nabla(\rho \circ |h|)(\xi) \cdot f(\xi, d)|}{1 + |f(\xi, d)|^2 + \kappa(\xi)} \leq 1,$$

where the last inequality follows from (4.4). Therefore the assumptions of Lemma 4.14 are satisfied with $p := 1.5\rho$ and $f(\xi, d) := \widetilde{f}(\xi, d, v) = \widehat{f}(\xi, d)$. By Lemma 4.14, we can find a constant K_0 such that if $|\xi| > K_0$ and $|\xi| \geq 1.6\rho(|h(\xi)|)$, then $z(t, \xi, \mathbf{d}, v) \notin \mathcal{D} \cup \mathcal{B}$ for all $t \in [0, 1)$. In particular, for such a ξ we will have $\theta_{\mathbf{d}}(\xi, \mathbf{v}) > 1$ and $|z(t, \xi, \mathbf{d}, \mathbf{v})| > |\xi| - 1$ for all positive $t < 1$. Hence, the inequality

$$\Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|) > \Xi(|\xi| - 1) \quad \forall t \in [0, 1)$$

holds for any $\mathbf{d} \in \mathcal{M}_\Omega$, $\mathbf{v} \in \mathcal{W}$, and any ξ such that

$$|\xi| > \max\{K_0 + 1,\ 1.6\rho(|h(\xi)|)\}\,. \tag{4.31}$$

Therefore, for any ξ as in (4.31) and any $\mathbf{d} \in \mathcal{M}_\Omega$, the following estimate holds:

$$V_0(\xi) \ \geq\ \widetilde{V}_0(\xi, \mathbf{d}) \ \geq\ \int_0^1 \Xi(|z(t, \xi, \mathbf{d}, \mathbf{v})|)dt \ \geq\ \Xi(|\xi| - 1).$$

Next, notice that V_0 is strictly positive on $\mathcal{E}$. Indeed, $|\widetilde{f}(\xi, d, v)| \leq 3$ for any $\xi \in \mathcal{E}$, $d \in \Omega$, and $v \in [-1, 1]^n$, so that

$$|z(s, \xi, \mathbf{d}, \mathbf{v}) - \xi| \leq \frac{1}{2}\text{dist}(\xi, \mathcal{D}) \quad \forall\, s \leq \frac{\frac{1}{2}\text{dist}(\xi, \mathcal{D})}{3},\ \mathbf{d} \in \mathcal{M}_\Omega,\ \mathbf{v} \in \mathcal{W}.$$

Therefore $\theta_{\mathbf{d}}(\xi, \mathbf{v}) \geq \frac{1}{6}\text{dist}(\xi, \mathcal{D})$ and $|z(s, \xi, \mathbf{d}, \mathbf{v})| \geq \xi - \frac{1}{2}\text{dist}(\xi, \mathcal{D})$ for all $s \leq \frac{1}{6}\text{dist}(\xi, \mathcal{D})$, all $\mathbf{d} \in \mathcal{M}_\Omega$, and $\mathbf{v} \in \mathcal{W}$. So, we have

$$V_{\mathbf{v}}(\xi, \mathbf{d}) \ \geq\ \int_0^{\text{dist}(\xi,\mathcal{D})/6} \Xi\left(|z(s, \xi, \mathbf{d}, \mathbf{v})|\right) ds \ \geq\ \frac{\text{dist}(\xi, \mathcal{D})}{6}\, \Xi(|\xi| - \text{dist}(\xi, \mathcal{D})/2).$$

This shows that

$$\inf_{\mathbf{d} \in \mathcal{M}_\Omega} \inf_{v \in \mathcal{W}} V_{\mathbf{v}}(\xi, \mathbf{d}) \geq \frac{\text{dist}(\xi, \mathcal{D})}{6}\, \Xi\!\left(|\xi| - \frac{\text{dist}(\xi, \mathcal{D})}{2}\right). \tag{4.32}$$

Thus also the $\sup_{\mathbf{d} \in \mathcal{M}_\Omega} \inf_{v \in \mathcal{W}} V_{\mathbf{v}}(\xi, \mathbf{d})$ satisfies (4.32), and hence the same inequality holds for V_0.

Since V_0 is lower semicontinuous, it attains its minimum on any compact set. For each positive l define

$$r_l := \frac{1}{l}\max\{K_0 + 1,\ 1.6\rho(|h(\xi)|)\} \quad \text{and} \quad m_l = \inf\{V_0(z)\ :\ z \in \mathcal{E}_1,\ r_l \leq |z| \leq r_1\}\ .$$

Since the sequence $\{m_l\}$ is nonincreasing and positive, and Ξ is of class $\mathcal{K}_\infty$, we can find a $\mathcal{K}_\infty$-function $\underline{\alpha}$ such that

$$\underline{\alpha}(s) < m_l \ \ \forall\, s \in [r_l, r_{l-1}],\ \forall\, l > 1$$

and

$$\underline{\alpha}(s) < \Xi(s-1) \forall\, s \geq \max\left\{K_0 + 1,\ 1.6\rho(|h(\xi)|)\right\}.$$

By construction, $\underline{\alpha}$ will be a lower bound for V_0 on $\mathcal{E}_1$. □

Combining Lemma 4.13 with (4.29), we get the following, using $\bar{\alpha} := \mu_2$:

$$\underline{\alpha}(|\xi|) \leq V_0(\xi) \leq \bar{\alpha}(|\xi|) \qquad \forall\, \xi \in \mathcal{E}_1. \tag{4.33}$$

The following lemma and corollary summarize the dissipation properties for V_0.

LEMMA 4.15. *For any $\xi \in \mathbf{X} \setminus (\mathcal{D} \cup \mathcal{B})$ and $\mathbf{d} \in \mathcal{M}_\Omega$, and any t_0 such that $z(t, \xi, \mathbf{d}) \notin \mathcal{D} \cup \mathcal{B}$ for all $t \in [0, t_0]$, the following dissipation inequality holds:*

$$V_0(z(t, \xi, \mathbf{d})) - V_0(\xi) \leq -\int_0^{t_0} \Xi(|z(t, \xi, \mathbf{d})|)\, dt.$$

Proof. Fix $\xi \in \mathbf{X} \setminus (\mathcal{D} \cup \mathcal{B})$, $\mathbf{d} \in \mathcal{M}_\Omega$, and any positive t_0 as in the formulation of the lemma. Let $\varepsilon > 0$ be given. Find $\mathbf{d}_1$ such that

$$V_0(z(t_0, \xi, \mathbf{d})) - \widetilde{V}_0(z(t_0, \xi, \mathbf{d}), \mathbf{d}_1) < \varepsilon,$$

and let $\mathbf{v}_1 \in \mathcal{W}$ be a control such that $\widetilde{V}_0(z(t_0, \xi, \mathbf{d}), \mathbf{d}_1) = V_{\mathbf{v}_1}(z(t_0, \xi, \mathbf{d}), \mathbf{d}_1)$. Let $\widetilde{\mathbf{d}}$ be defined by

$$\widetilde{\mathbf{d}}(t) = \begin{cases} \mathbf{d}(t) & \text{if } 0 \leq t \leq t_0, \\ \mathbf{d}_1(t - t_0) & \text{if } t > t_0. \end{cases}$$

Then $z(t, z(t_0, \xi, \mathbf{d}), \mathbf{d}_1, \mathbf{v}_1) = z(t + t_0, \xi, \widetilde{\mathbf{d}}, \widetilde{\mathbf{v}})$ for all $t \geq 0$. By assumption,

$$z(t, \xi, \mathbf{d}) \notin \mathcal{D} \cup \mathcal{B} \qquad \forall\, t \in [0, t_0],$$

and therefore we have

$$z(t, \xi, \mathbf{d}) = z(t, \xi, \mathbf{d}, \mathbf{v}) \qquad \forall\, t \in [0, t_0], \quad \forall\, \mathbf{v} \in \mathcal{W}. \tag{4.34}$$

Notice also that (4.34) implies that $\theta_{\widetilde{\mathbf{d}}}(\xi, \mathbf{v}) > t_0$ for all $\mathbf{v} \in \mathcal{W}$ and

$$\begin{aligned} \widetilde{V}_0(\xi, \widetilde{\mathbf{d}}) &= \min_{\mathbf{v} \in \mathcal{W}} \int_{t_0}^{\theta_{\widetilde{\mathbf{d}}}(\xi, \mathbf{v})} \Xi\left(\left|z(s, \xi, \widetilde{\mathbf{d}}, \mathbf{v})\right|\right) ds \\ &= \int_0^{t_0} \Xi\left(\left|z(s, \xi, \widetilde{\mathbf{d}})\right|\right) ds + \min_{\mathbf{v} \in \mathcal{W}} \int_{t_0}^{\theta_{\widetilde{\mathbf{d}}}(\xi, \mathbf{v})} \Xi\left(\left|z(s, \xi, \widetilde{\mathbf{d}}, \mathbf{v})\right|\right) ds \\ &= \int_0^{t_0} \Xi(|z(s, \xi, \mathbf{d})|) ds + \min_{\mathbf{v} \in \mathcal{W}} \int_0^{\theta_{\mathbf{d}_1}(z(t_0, \xi, \mathbf{d}), \mathbf{v})} \Xi(|z(s, z(t_0, \xi, \mathbf{d}), \mathbf{d}_1, \mathbf{v})|) ds \\ &= \int_0^{t_0} \Xi(|z(s, \xi, \mathbf{d})|) ds + \widetilde{V}_0(z(t_0, \xi, \mathbf{d}), \mathbf{d}_1). \end{aligned}$$

Consequently, one has

$$\widetilde{V}_0(z(t_0, \xi, \mathbf{d}), \mathbf{d}_1) = \widetilde{V}_0(\xi, \widetilde{\mathbf{d}}) - \int_0^{t_0} \Xi\left(|z(s, \xi, \mathbf{d})|\right) ds.$$

Thus,

$$\begin{aligned} V_0(z(t_0,\xi,\mathbf{d})) &\le \widetilde{V}_0(z(t_0,\xi,\mathbf{d}),\mathbf{d}_1)+\varepsilon \\ &= \widetilde{V}_0(\xi,\widetilde{\mathbf{d}}) - \int_0^{t_0} \Xi(|z(s,\xi,\mathbf{d})|)\,ds + \varepsilon \\ &\le V_0(\xi) - \int_0^{t_0} \Xi(|z(s,\xi,\mathbf{d})|)\,ds + \varepsilon. \end{aligned}$$

Letting $\varepsilon \to 0$, we get the desired inequality. □

Thus, we have proven that the UOSS dissipation inequality holds for V_0 along the trajectories of the slower system $\hat{\Sigma}$ which are entirely contained in $\mathbf{X}\setminus(\mathcal{D}\cup\mathcal{B})$. It follows immediately that the same estimate holds along the trajectories of the original system Σ.

Corollary 4.16. *For any $\xi\in\mathbf{X}\setminus(\mathcal{D}\cup\mathcal{B})$ and $\mathbf{d}\in\mathcal{M}_\Omega$, and any t_0 such that $x(t,\xi,\mathbf{d})\notin\mathcal{D}\cup\mathcal{B}$ for all $t\in[0,t_0]$, the following dissipation inequality holds:*

$$V_0(x(t,\xi,\mathbf{d})) - V_0(\xi) \le -\int_0^{t_0} \Xi(|x(t,\xi,\mathbf{d})|)\,dt. \tag{4.35}$$

Proof. Pick an initial state $\xi\in\mathbf{X}\setminus(\mathcal{D}\cup\mathcal{B})$, a disturbance $\mathbf{d}\in\mathcal{M}_\Omega$, and an appropriate t_0. Then

$$\begin{aligned} &V_0(x(t_0,\xi,\mathbf{d})) - V_0(x(\xi)) \\ &= V_0(z(\sigma_{\xi,\mathbf{d}}(t_0),\xi,\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1})) - V_0(z(\sigma_{\xi,\mathbf{d}}(t_1),\xi,\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1})) \\ &\le -\int_0^{\sigma_{\xi,\mathbf{d}}(t_0)} \Xi\left(\left|z(s,\xi,\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1})\right|\right)ds \\ &= -\int_0^{t_0} \Xi\left(\left|z(\sigma_{\xi,\mathbf{d}}(t),\xi,\mathbf{d}\circ{\sigma_{\xi,\mathbf{d}}}^{-1})\right|\right)d\sigma_{\xi,\mathbf{d}}(t) \\ &= -\int_0^{t_0} \Xi(|x(t,\xi,\mathbf{d})|)\,\frac{d}{dt}\sigma_{\xi,\mathbf{d}}(t)\,dt \\ &= -\int_0^{t_0} \Xi(|x(t,\xi,\mathbf{d})|)[1+|f(x(t,\xi,\mathbf{d}),\mathbf{d}(t))|^2+\kappa(x(s,\xi,\mathbf{d}))]\,dt \\ &\le -\int_0^{t_0} \Xi(|x(t,\xi,\mathbf{d})|)\,dt. \qquad \square \end{aligned}$$

4.4. Some definitions and facts from nonsmooth analysis.

Definition 4.17. *A vector $\zeta\in\mathbf{R}^n$ is a* proximal subgradient *(respectively,* proximal supergradient*) of the function $V:\mathbf{R}^n\to(-\infty,+\infty]$ at x if there exists some positive σ such that, for all x' in some neighborhood of x,*

$$V(x') \ge V(x) + \zeta\cdot(x'-x) - \sigma|x'-x|^2 \tag{4.36}$$

$$(\textit{correspondingly, } V(x') \le V(x) + \zeta\cdot(x'-x) + \sigma|x'-x|^2). \tag{4.37}$$

The (possibly empty) set of all proximal subgradients (respectively, supergradients) of V at x is called the proximal subdifferential *and is denoted $\partial_P V(x)$ (respectively,* proximal superdifferential*, denoted $\partial^P V(x)$). Note that the definitions imply that if* *the function V is differentiable at x, then both the subdifferential and superdifferential sets must be subsets of the $\{\nabla V(x)\}$.*

LEMMA 4.18. *Let Σ be a system of type* (2.14)*, ξ be a vector in $\mathbf{X}$, $\bar{d} \in \Omega$, and $V : \mathbf{X} \to \mathbf{R}$. Then, if there exist some continuous $\alpha_\xi : \mathbf{R}_{\geq 0} \to \mathbf{R}$ and $\varepsilon > 0$ such that the following inequality holds for all $\tau < \varepsilon$,*

$$V(x(\tau,\xi,\mathbf{d})) - V(\xi) \;\leq\; \int_0^\tau \alpha_\xi(t)dt \tag{4.38}$$

(where $\mathbf{d}$ is the constant disturbance equal to $\bar{d}$), then for any $\zeta \in \partial_P V(\xi)$ the proximal form of inequality (4.38) *holds:*

$$\zeta \cdot f(\xi,\bar{d}) \leq \alpha_\xi(0). \tag{4.39}$$

Proof. It follows from (4.36) and (4.38) that, for all τ close enough to 0 we have

$$\int_0^\tau \alpha_\xi(t)dt \;\geq\; V(x(\tau,\xi,\mathbf{d})) - V(\xi) \;\geq\; \zeta \cdot (x(\tau,\xi,\mathbf{d}) - \xi) - \sigma|x(\tau,\xi,\mathbf{d}) - \xi|^2.$$

Dividing by τ and passing to the limit as τ tends to 0, we get (4.39). □

4.5. Smoothing out a continuous Lyapunov function. The next result shows how to approximate a continuous function V by a locally Lipschitz one in a weak C^1 sense. The function V is assumed to be bounded below, or up to a translation by a constant, nonnegative.

LEMMA 4.19. *Let $\Sigma : \dot{x} = f(x,d)$ be a system, with $x \in \mathbf{X} = \mathbf{R}^n$ and $d \in \Omega$, a compact metric space, so that $f(x,d)$ is locally Lipschitz in x uniformly on d and jointly continuous in x and d. Assume that we are given*

- *an open subset $\mathcal{O}$ of $\mathbf{X}$;*
- *a continuous, nonnegative function $V : \mathcal{O} \to \mathbf{R}$ satisfying*

  $$\zeta \cdot f(x,d) \leq \Theta(x,d) \;\; \forall x \in \mathcal{O},\; \zeta \in \partial_P V(x),\; d \in \Omega \tag{4.40}$$

  with some continuous function $\Theta : \mathcal{O} \times \Omega \to \mathbf{R}$;
- *two positive, continuous functions Υ_1 and Υ_2 on $\mathcal{O}$.*

Then there exists a function $\widetilde{V} : \mathcal{O} \to \mathbf{R}$, locally Lipschitz on $\mathcal{O}$, such that

$$0 \leq V(x) - \widetilde{V}(x) \leq \Upsilon_1(x) \;\; \forall x \in \mathcal{O} \tag{4.41}$$

and

$$L_{f_d}\widetilde{V}(x) \leq \Theta(x,d) + \Upsilon_2(x) \quad \forall d \in \Omega \text{ and for almost all } x \in \mathcal{O}, \tag{4.42}$$

where f_d is the vector field defined by $f_d = f(\cdot,d)$.

Note that, by Rademacher's theorem, the directional derivative $L_{f_d}\widetilde{V}(x)$, given by $\nabla V(x) \cdot f(x,d)$, is defined for almost all x because $\widetilde{V}$ is locally Lipschitz. The proof will follow closely along the lines of the proof of the similar result for CLFs, found in [9] or [40].

We will first prove a "local" version of the result. For any K which is a compact subset of $\mathcal{O}$ and $r > 0$, we introduce the following notations:

- $\bar{B}_r(K) := \{x \in \mathbf{X} : \exists \xi \in K \text{ with } |x - \xi| \leq r\}$, the closed r-fattening of K, for $r > 0$,
- $\beta(K) := \sup_{x \in K} V(x)$,

- $m_K := \frac{1}{4} \min \{\Upsilon_2(x) : x \in K\}$,
- $\ell_K :=$ a Lipschitz constant for f with respect to x in K, that is,

$$|f(x, d) - f(x', d)| \leq \ell_K |x - x'| \;\; \forall d \in \Omega, \; \forall x \in K,$$

- $\omega_K(\cdot) :=$ the modulus of continuity of V on K, that is,

$$\omega_K(\delta) := \sup \{V(x) - V(x') \; : \; |x - x'| \leq \delta; x, x' \in K\},$$

- $\pi_K(\cdot) :=$ the modulus of continuity of Θ on $K \times \Omega$, that is,

$$\begin{aligned}\pi_K(\delta) := \sup\{\Theta(x, d) - \Theta(x', d') \; : \; &|x - x'| + \operatorname{dist}(d, d') \leq \delta;\\ &x, x' \in K, d, d' \in \Omega\}.\end{aligned}$$

To approximate the given continuous function V by a locally Lipschitz one, we would like to use the notion of "Iosida–Moreau inf-convolution," well known in convex analysis. Fix a parameter $\alpha \in (0, 1]$. Suppose for the moment that V is defined on the whole $\mathbf{X}$. Define

$$V_\alpha(x) := \min_{y \in \mathbf{X}} \left[V(y) + \frac{1}{2\alpha^2} |y - x|^2 \right].$$

For each fixed x, the set of points y where the minimum is attained is nonempty because V is bounded below. Denote one of them by $y_\alpha(x)$.

Fix a compact K and $\alpha \in (0, 1]$, let

$$K_\alpha := \bar{B}_{\alpha\sqrt{2\beta(K)}}(K).$$

The following four claims summarize some of the useful properties of V_α, proven in [9], [40] (or see the primary sources such as [8]).

Claim 1. For all $x \in K$,

$$|y_\alpha(x) - x|^2 \; \leq \; \min \left\{ 2\alpha^2 \beta(K), \; 2\alpha^2 \omega_{K_\alpha} \left(\alpha \sqrt{2\beta(K)} \right) \right\}.$$

Proof of Claim 1. By definition of V_α and $\beta(K)$, we have

$$\frac{1}{2\alpha^2} \, |y_\alpha(x) - x|^2 \leq V(x) - V(y_\alpha(x)) \leq V(x) \leq \beta(K), \tag{4.43}$$

so that $|y_\alpha(x) - x|^2 \leq 2\alpha^2 \beta(K)$. On the other hand, the first inequality in (4.43) implies also that

$$\begin{aligned}|y_\alpha(x) - x|^2 &\leq 2\alpha^2 (V(x) - V(y_\alpha(x)))\\ &\leq 2\alpha^2 \omega_{K_\alpha} (|y_\alpha(x) - x|) \; \leq \; 2\alpha^2 \omega_{K_\alpha} \left(\alpha \sqrt{2\beta(K)} \right),\end{aligned}$$

proving the claim.

Let

$$\zeta_\alpha(x) = \frac{x - y_\alpha(x)}{\alpha^2}.$$

Claim 2. For any $x \in \mathbf{X}$,

$$\zeta_\alpha(x) \in \partial_P V(y_\alpha(x)) \;\text{ and} \tag{4.44}$$

$$\zeta_\alpha(x) \in \partial^P V_\alpha(x). \tag{4.45}$$

Claim 3. For any $x \in K$,

$$V_\alpha(x) \le V(x) \le V_\alpha(x) + \omega_{K_\alpha}\left(\alpha\sqrt{2\beta(K)}\right).$$

Claim 4. V_α is locally Lipschitz.

Now recall that, in our setting, V is defined on an open subset $\mathcal{O}$ of $\mathbf{X}$, so, we can't minimize the expression in the definition of V_α over the whole state space. However, for any compact subset K of $\mathcal{O}$ we can choose α small enough so that

$$K_\alpha \subset \mathcal{O},$$

hence, for any x in K we could define V_α minimizing over $y \in \mathcal{O}$, and the same function V_α results on K.

LEMMA 4.20. *Assume that* V *satisfies* (4.40), *a compact* K *is fixed, and* $\alpha \in (0,1]$ *is chosen to satisfy*

- $K_\alpha = \bar{B}_{\alpha\sqrt{2\beta(K)}}(K) \subset \mathcal{O}$,
- $\pi_{K_\alpha}\left(\alpha\sqrt{2\beta(K)}\right) \le m_K$, *and*
- $2\ell_{K_\alpha}\omega_{K_\alpha}\left(\alpha\sqrt{2\beta(K)}\right) \le m_K$.

Then the function

$$V_\alpha(x) := \inf_{y\in\mathcal{O}}\left[V(y) + \frac{1}{2\alpha^2}|y-x|^2\right] \tag{4.46}$$

will possess the following property:

$\forall\ x \in K, \forall\ d \in \Omega$, *and* $\forall\ \zeta \in \partial_P V_\alpha(x)$,

$$\zeta \cdot f(x,d) \le \Theta(x,d) + 2m_K.$$

Proof. Fix any $x \in K$. The choice of α ensures that the infimum in (4.46) is a minimum, and it is achieved at some $y_\alpha(x) \in K_\alpha$. The definition of $\zeta_\alpha(x)$ and Claim 1 imply that

$$|\zeta_\alpha(x)||y_\alpha(x)-x| = \frac{|y_\alpha(x)-x|^2}{\alpha^2} \le 2\omega_{K_\alpha}\left(\alpha\sqrt{2\beta(K)}\right). \tag{4.47}$$

Using again the fact that $y_\alpha(x) \in K_\alpha$,

$$|f(x,d) - f(y_\alpha(x),d)| \le \ell_{K_\alpha}|x - y_\alpha(x)|.$$

Combining the last inequality with (4.47) we obtain

$$|\zeta_\alpha(x)||f(x,d) - f(y_\alpha(x),d)| \le 2\ell_{K_\alpha}\omega_{K_\alpha}\left(\alpha\sqrt{2\beta(K)}\right). \tag{4.48}$$

Now, by Claim 2, $\zeta_\alpha(x) \in \partial_P V(y_\alpha(x))$. Hence,

$$\begin{aligned}\zeta_\alpha(x)\cdot f(y_\alpha(x),d) &\le \Theta(y_\alpha(x),d)\\ &\le \Theta(x,d) + \pi_{K_\alpha}(|x-y_\alpha(x)|)\\ &\le \Theta(x,d) + \pi_{K_\alpha}\left(\alpha\sqrt{2\beta(K)}\right)\\ &\le \Theta(x,d) + m_K.\end{aligned}$$

So, by the last inequality, (4.48), and the choice of α, we have

$$\begin{aligned}\zeta_\alpha(x)\cdot f(x,d) &= \zeta_\alpha(x)\cdot f(y_\alpha(x),d)+\zeta_\alpha(x)\cdot(f(x,d)-f(y_\alpha(x),d))\\ &\le \Theta(x,d)+m_K+m_K.\end{aligned} \tag{4.49}$$

Next, we show that $\partial_P V_\alpha(x)\subseteq\{\zeta_\alpha\}$ for all $x\in K$. Pick any $\zeta\in\partial_P V_\alpha(x)$. By definition of the proximal subgradient, the inequality

$$\zeta\cdot(y-x)\le V_\alpha(y)-V_\alpha(x)+o(|y-x|) \tag{4.50}$$

holds for all y near x. Since $\zeta_\alpha(x)$ is a proximal supergradient of V_α at x, we also have

$$-\zeta_\alpha(x)\cdot(y-x)\le -V_\alpha(y)+V_\alpha(x)+o(|y-x|). \tag{4.51}$$

Adding (4.50) and (4.51), we get

$$(\zeta-\zeta_\alpha(x))\cdot(y-x)\le o(|y-x|) \tag{4.52}$$

for all y sufficiently close to x. Substituting $y=x+h(\zeta-\zeta_\alpha(x))$ in (4.52) and letting h tend to 0, we arrive at $\zeta=\zeta_\alpha(x)$. □

Now we are ready to prove the main lemma of the section.

Proof of Lemma 4.19. For every $x\in\mathcal{O}$ find an $r_x>0$ small enough so that $\bar{B}_{r_x}(x)\subset\mathcal{O}$. Then the collection of open balls $\{B_{r_x}(x),\, x\in\mathcal{O}\}$ forms an open covering of $\mathcal{O}$. By paracompactness of $\mathcal{O}$ we can find a locally finite refinement $\{B_i, i\in\mathbb{N}\}$ of $\{B_{r_x}(x),\, x\in\mathcal{O}\}$ (cf. [6, Lemma 4.1]). Moreover, since $\cup_{x\in\mathcal{O}}B_{r_x}=\mathcal{O}$, we also have $\cup_i B_i=\mathcal{O}$. Let $\{\varphi_i, i\in\mathbb{N}\}$ be a partition of unity, subordinate to $\{B_i\}$. For each index i, let

$$\mathcal{J}_i=\{j\in\mathbb{N}\,:\,B_i\cap B_j\ne\emptyset\}\,.$$

Notice that $\mathcal{J}_i$ is finite for all i, because of local finiteness of the covering $\{B_i\}$. For every $i\in\mathbb{N}$ define

$$M_i:=\sup|\nabla\varphi_i(x)||f(x,d)|,$$

where the supremum is taken over all $x\in\cup_{j\in\mathcal{J}_i}\bar{B}_j$ and all $d\in\Omega$; and

$$N_i=\max_{j\in\mathcal{J}_i}\operatorname{card}(\mathcal{J}_j),$$

that is, N_i denotes the maximum cardinality of $\mathcal{J}_j$ for j such that B_j intersects B_i. Next, for each compact $K=\bar{B}_i$, $i\in\mathbb{N}$, choose an α_i as in the formulation of Lemma 4.20, satisfying the following two additional conditions:

$$N_i\,\omega_{K_{\alpha_i}}\left(\alpha_i\sqrt{2\beta(\bar{B}_i)}\right)M_i<\frac{1}{2}\inf_{x\in\bar{B}_i}\Upsilon_2(x) \tag{4.53}$$

and

$$\omega_{K_{\alpha_i}}\left(\alpha_i\sqrt{2\beta(\bar{B}_i)}\right)<\frac{1}{2}\inf_{x\in\bar{B}_i}\Upsilon_1(x) \tag{4.54}$$

(where K_{α_i} denotes $B_{\alpha_i\sqrt{2\beta(\bar{B}_i)}}(\bar{B}_i)$).

Next, for each $i \in \mathbb{N}$, define V_{α_i} on $\bar{B}_i$ as in (4.46) (this can be done, because, by the choice of α_i, $K_{\alpha_i} \subseteq \mathcal{O}$). Since V_{α_i} is locally Lipschitz, it is differentiable almost everywhere by Rademacher's theorem. Lemma 4.20 implies that for almost all $x \in \bar{B}_i$ and all $d \in \Omega$

$$L_{f_d} V_{\alpha_i}(x) \leq \Theta(x,d) + \Upsilon_2(x)/2.$$

Define

$$\widetilde{V} := \sum_{i=1}^{+\infty} V_{\alpha_i} \varphi_i.$$

Strictly speaking, this does not make sense, because V_{α_i} is not defined outside $\bar{B}_i$, but that does not matter because φ_i vanishes outside B_i anyway.

Since each $V_{\alpha_i}(x) \leq V(x)$ for all $x \in B_i$ and φ_i's add up to 1, the definition of $\widetilde{V}$ shows that $\widetilde{V}(x) \leq V(x)$ for all $x \in \mathcal{O}$. It is also clear from the definition of $\widetilde{V}$ that $\widetilde{V}$ is locally Lipschitz on $\mathcal{O}$.

We claim that for almost all $x \in \mathcal{O}$ and all $d \in \Omega$ the following hold:

1. $0 \leq V(x) - \widetilde{V}(x) \leq \Upsilon_1(x)$;
2. $\nabla \widetilde{V}(x) \cdot f(x,d) \leq \Theta(x,d) + \Upsilon_2(x)$.

Take any $x \in \mathcal{O}$ and find $i \in \mathbb{N}$ such that $x \in B_i$. Define also $\mathcal{J}_x := \{j \in \mathbb{N} : x \in B_j\}$. Note that $\mathcal{J}_x \subseteq \mathcal{J}_i$ and that $\widetilde{V}(x) := \sum_{j \in J_x} V_{\alpha_j}(x)\varphi_j(x)$. Then Claim 3 and the choice of α_j (condition (4.54)) imply

$$V(x) - V_{\alpha_j}(x) \leq \omega_{K_{\alpha_j}} \left(\alpha_j \sqrt{2\beta(\bar{B}_j)} \right) \leq \Upsilon_1(x)/2$$

for all $j \in \mathcal{J}_x$. Thus,

$$\begin{aligned} V(x) - \widetilde{V}(x) &= V(x) - \sum_{j=1}^{+\infty} V_{\alpha_j} \varphi_j \\ &= V(x) - \sum_{j \in \mathcal{J}_i} V_{\alpha_j} \varphi_j \\ &\leq V(x) - \min_{j \in \mathcal{J}_x} \left\{ V_{\alpha_j}(x) \right\} \left(\sum_{j \in \mathcal{J}_x} \varphi_j(x) \right) \\ &= V(x) - \min_{j \in \mathcal{J}_x} \left\{ V_{\alpha_j}(x) \right\} \\ &\leq \frac{\Upsilon_1(x)}{2}, \end{aligned}$$

proving the first statement.

To prove the second statement, write

$$\begin{aligned} L_{f_d} \widetilde{V}(x) &= \sum_{j \in \mathcal{J}_x} L_{f_d} V_{\alpha_j}(x) \, \varphi_j(x) + \sum_{j \in \mathcal{J}_x} V_{\alpha_j}(x) \, L_{f_d} \varphi_j(x) \\ &= \sum_{j \in \mathcal{J}_x} L_{f_d} V_{\alpha_j}(x) \, \varphi_j(x) + \sum_{j \in \mathcal{J}_x} V_{\alpha_j}(x) \, L_{f_d} \varphi_j(x) - V(x) \sum_{j \in \mathcal{J}_x} L_{f_d} \varphi_j(x) \qquad (4.55) \\ &\leq \max_{j \in \mathcal{J}_x} L_{f_d} V_{\alpha_j}(x) \sum_{j \in \mathcal{J}_x} \varphi_j(x) + \sum_{j \in \mathcal{J}_x} \left(V_{\alpha_j}(x) - V(x) \right) \, L_{f_d} \varphi_j(x) \end{aligned}$$

$$\le \Theta(x,d) + \Upsilon_2(x)/2 + \sum_{j\in\mathcal{J}_i} \left|V_{\alpha_j}(x) - V(x)\right| |L_{f_d}\varphi_j(x)|$$

$$\le \Theta(x,d) + \Upsilon_2(x)/2 + \sum_{j\in\mathcal{J}_i} M_j \omega_{K_{\alpha_j}}\left(\alpha_j\sqrt{2\beta(\bar{B}_j)}\right) \tag{4.56}$$

$$\le \Theta(x,d) + \Upsilon_2(x)/2 + \sum_{j\in\mathcal{J}_i} \frac{\Upsilon_2(x)}{2N_j} \tag{4.57}$$

$$\le \Theta(x,d) + \Upsilon_2(x)/2 + \Upsilon_2(x)/2 \tag{4.58}$$

$$\le \Theta(x,d) + \Upsilon_2(x),$$

where the equality (4.55) follows from the fact that $\sum_{j\in\mathcal{J}_x} L_{f_d}\varphi_j(x) = 0$, inequality (4.56) follows from Lemma 4.20 and Claim 3; inequality (4.57) follows from the choice of α_i (condition (4.53)); and (4.58) is implied by the fact that

$$\operatorname{card}\mathcal{J}_i \le N_j \quad \forall j \in \mathcal{J}_i.$$

This completes the proof of the lemma. □

The lemma we have just proved will provide the "continuous ⇒ locally Lipschitz away from zero" step in the smoothing process. To obtain a smooth Lyapunov function, we will use the following simple smoothing result. The proof is given, for the special case when α does not depend on d, in [25], but the general case (α depends on d) is proved in exactly the same manner, so we omit the proof here.

Lemma 4.21. *Let $\mathcal{O}$ be an open subset of $\mathbf{R}^n$, and let Ω be a compact subset of $\mathbf{R}^l$, and assume the following as given:*

- *a locally Lipschitz function $\Phi : \mathcal{O} \to \mathbf{R}$;*
- *a continuous map $f : \mathbf{R}^n \times \Omega \to \mathbf{R}^n$, $(x, d) \mapsto f(x, d)$ which is locally Lipschitz on x uniformly on d;*
- *a continuous function $\alpha : \mathcal{O} \times \Omega \to \mathbf{R}$ and continuous functions $\mu, \nu : \mathcal{O} \to \mathbf{R}_{>0}$*

such that for each $d \in \Omega$,

$$L_{f_d}\Phi(\xi) \le \alpha(\xi, d) \quad \textit{almost everywhere } \xi \in \mathcal{O}, \tag{4.59}$$

where f_d is the vector field defined by $f_d = f(\cdot, d)$ (recall that $\nabla\Phi$ is defined almost everywhere, since Φ is locally Lipschitz by Rademacher's theorem). Then there exists a smooth function $\Psi : \mathcal{O} \to \mathbf{R}$ such that

$$|\Phi(\xi) - \Psi(\xi)| < \mu(\xi) \quad \forall \xi \in \mathcal{O}$$

and for each $d \in \Omega$,

$$L_{f_d}\Psi(\xi) \le \alpha(\xi, d) + \nu(\xi) \quad \forall \xi \in \mathcal{O}.$$

The next result immediately follows by Lemma 4.21.

Corollary 4.22. *Under the assumptions of Lemma 4.19 there also exists a smooth function $\hat{V}$ on $\mathcal{O}$, satisfying inequalities*

$$\left|V(x) - \hat{V}(x)\right| \le \Upsilon_1(x) \ \forall x \in \mathcal{O} \tag{4.60}$$

and

$$L_{f_d}\hat{V}(x) \le \Theta(x,d) + \Upsilon_2(x) \quad \forall d \in \Omega, \ \forall x \in \mathcal{O}. \tag{4.61}$$

Proof. Suppose that $\mathcal{O}$, V, Υ_1, Υ_2 are given and the assumptions of Lemma 4.19 hold. Replacing Υ_1 by $\Upsilon_1/2$ and Υ_2 by $\Upsilon_2/2$, and applying Lemma 4.19, we can find a locally Lipschitz function $\widetilde{V}$, defined on $\mathcal{O}$ and satisfying (4.41) and (4.61) with $\Upsilon_1/2$ and $\Upsilon_2/2$ instead of Υ_1 and Υ_2. Next, Lemma 4.21, applied with the same $\mathcal{O}$ and with $\Phi := \widetilde{V}$, $\alpha := \Theta + \Upsilon_2/2$, $\mu := \Upsilon_1/2$, and $\nu := \Upsilon_2/4$, furnishes a smooth function $\hat{V} := \Psi$ as needed. □

In section 4.3 we have constructed a function V_0, satisfying inequalities (4.33) on $\mathcal{E}_1$ and (4.35) along all trajectories of Σ, contained in $\mathcal{E}_1$. Therefore, by Lemma 4.18, the proximal inequality (4.39) holds for V_0 at any interior point of $\mathcal{E}_1$. Then Corollary 4.22, applied with $\mathcal{O} := \text{int } \mathcal{E}_1$, $\Upsilon_1(\cdot) := \underline{\alpha}(|\cdot|)/2$, $\Upsilon_2(\cdot) := \Xi(|\cdot|)/2$, $\Theta(x,d) := \Xi(|x|)$, provides a smooth function

$$V_1 : \text{int } \mathcal{E}_1 \to \mathbf{R}_{>0}; \quad V_1 := \hat{V}_0,$$

satisfying the following two conditions for all $\xi \in \text{int } \mathcal{E}_1$:

$$\underline{\alpha}(|\xi|)/2 \le V_1(\xi) \le \bar{\alpha}(|\xi|) + \underline{\alpha}(|\xi|)/2$$

(we will replace $\underline{\alpha}(|\cdot|)/2$ by $\underline{\alpha}(\cdot)$ and $\bar{\alpha}(|\cdot|) + \underline{\alpha}(|\cdot|)/2$ by $\bar{\alpha}(|\cdot|)$ from now on to avoid cluttering the notation),

$$L_{f_d} V_1(\xi) \le -\Xi_1(|\xi|) \quad \forall d \in \Omega, \tag{4.62}$$

where $\Xi_1(\cdot) \equiv \Xi(\cdot)/2$.

4.6. Extending to the rest of X and smoothing at the origin. To construct a UOSS-Lyapunov-like function defined on the whole $\mathbf{X}$, we must "patch" V_1 with some smooth, proper, and positive definite function such that the dissipation inequality still holds.

LEMMA 4.23. *Suppose Σ is a system of type* (2.14), *and ρ is a function of class $\mathcal{K}_\infty$. Define $\mathcal{E}_1 := \{x \in \mathbf{X} : |x| > 2\rho(|h(x)|)\}$ and suppose that $V_1 : \mathcal{E}_1 \to \mathbf{R}_{\ge 0}$ is a smooth function satisfying, with some suitable $\mathcal{K}_\infty$ functions, the inequality*

$$\underline{\alpha}(|\xi|) \le V_1(\xi) \le \bar{\alpha}(|\xi|) \tag{4.63}$$

and inequality (4.62) *on $\mathcal{E}_1$. Then there exist a Lyapunov-like function V_2 for Σ, smooth away from the origin, and a class $\mathcal{K}_\infty$ function Φ, such that*

$$V_2(x) = \Phi \circ V_1(x) \quad \forall\, x \text{ such that } |x| > 3\rho(|h(x)|),$$

and the following dissipation inequality holds with some $\check{\alpha}_3 \in \mathcal{K}_\infty$, $\check{\gamma} \in \mathcal{K}$:

$$\nabla V_2(\xi) \cdot f(\xi, d) \le -\check{\alpha}_3(|\xi|) + \check{\gamma}(|h(\xi)|) \qquad \forall \xi \ne 0, \quad \forall d \in \Omega. \tag{4.64}$$

Proof. Let

$$\mathcal{E}_2 := \{\xi : |\xi| > 3\rho(|h(\xi)|)\}.$$

Since the sets $\{\xi : |\xi| \ge 3\rho(|h(\xi)|)\}$ and $\{\xi : |\xi| \le 2\rho(|h(\xi)|)\}$ are disjoint and closed in the topology of $\mathbf{X} \setminus \{0\}$, one can find a smooth function $\phi : \mathbf{X} \setminus \{0\} \to [0,1]$ with the property that

$$\phi(\xi) = \begin{cases} 1 & \text{if } |\xi| \ge 3\rho(|h(\xi)|), \\ 0 & \text{if } |\xi| \le 2\rho(|h(\xi)|) \end{cases}$$

and ϕ is nonzero elsewhere. It is easy to see that $|\nabla\phi(x)|$ is bounded above by a $\mathcal{K}$-function ν_2 of $|x|$ outside the unit ball centered at 0. One can also find a smooth, strictly increasing function $\nu_1 : [0,1] \to \mathbf{R}_{\geq 0}$, such that $\nu_1(0) = 0$ and $|\nabla\phi(x)| \leq \frac{1}{\nu_1(|x|)}$ for all x such that $0 < |x| \leq 1$.

Let ν_3 be a $\mathcal{K}$-function such that $\max_{d\in\Omega} |f(\xi,d)| \leq \nu_3(|\xi|)$. Take any smooth function $\pi_1 \colon [0,\bar{\alpha}(1)] \to \mathbf{R}_{\geq 0}$ with $\pi_1'(s) > 0$ for all $s \in (0,\bar{\alpha}(1))$, such that

$$\frac{\pi_1(\bar{\alpha}(r))}{\nu_1(r)} < s(r), \quad \frac{\pi_1(r^2)}{\nu_1(r)} < s(r)$$

for some $\mathcal{K}$-function s and all $0 < r \leq 1$. Let π_2 be any $\mathcal{K}$-function such that $\pi_2(r) \leq \pi_1'(r)$ for all nonnegative $r \leq 1$.

Let $\Phi(r) = \int_0^r \pi_2(r_1)dr_1$. Then $\Phi(r) \leq \pi_1(r)$ for all $r \leq 1$, so that $\frac{\Phi(\bar{\alpha}(r))}{\nu_1(r)} < s(r)$ and $\frac{\Phi(r^2)}{\nu_1(r)} < s(r)$ for all $r \in (0,1]$.

Now let

$$V_2(\xi) = \phi(\xi)\Phi(V_1(\xi)) + (1-\phi(\xi))\Phi(|\xi|^2). \tag{4.65}$$

If $|\xi| > 3\rho(|h(\xi)|)$, then $V_2 \equiv \Phi\circ V_1$ in a neighborhood of ξ, so that, for all $d \in \Omega$,

$$\nabla V_2(\xi)\cdot f(\xi,d) = \Phi'(V_1(\xi))[\nabla V_1(\xi)\cdot f(\xi,d)] \leq -\pi_2(\underline{\alpha}(|\xi|))\Xi_1(|\xi|). \tag{4.66}$$

On the other hand, if $|\xi| \leq 3\rho(|h(\xi)|)$, then

$$\begin{aligned}\nabla V_2(\xi)\cdot f(\xi,d) = {} & [\nabla\phi(\xi)\cdot f(\xi,d)]\,\Phi(V_1(\xi)) + \phi(\xi)\,\Phi'(V_1(\xi))\,[\nabla V_1(\xi)\cdot f(\xi,d)] \\ & -[\nabla\phi(\xi)\cdot f(\xi,d)]\,\Phi(|\xi|^2) + (1-\phi(\xi))\,2\Phi'(|\xi|^2)\,[\xi\cdot f(\xi,d)] \\ \leq {} & |\nabla\phi(\xi)|\,|f(\xi,d)|\;\Phi(V_1(\xi)) + |\nabla\phi(\xi)|\,|f(\xi,d)|\;\Phi(|\xi|^2) \\ & + 2\,|(1-\phi(\xi))|\;\Phi'(|\xi|^2)\;|\xi|\,|f(\xi,d)| \\ \leq {} & |\nabla\phi(\xi)|\,|f(\xi,d)|\;\Phi(\bar{\alpha}(|\xi|)) + |\nabla\phi(\xi)|\,|f(\xi,d)|\;\Phi(|\xi|^2) \\ & +2\Phi'(|\xi|^2)\;|\xi|\,|f(\xi,d)|\,,\end{aligned}$$

where the first inequality follows from the fact that $\phi(\xi)\,\Phi'(V_1(\xi))\,[\nabla V_1(\xi)\cdot f(\xi,d)] \leq 0$. Next, the definition of Φ provides the following bounds for the three terms in the right-hand side of the last inequality:

$$\Phi(\bar{\alpha}(|\xi|))\,|\nabla\phi(\xi)|\,|f(\xi,d)| \leq \max\{s(|\xi|), \Phi(\bar{\alpha}(|\xi|))\nu_2(|\xi|)\}\,\nu_3(|\xi|),$$

$$2\Phi'(|\xi|^2)[|\xi|\,|f(\xi,d)|] \leq 2\pi_2(|\xi|^2)\,|\xi|\,\nu_3(|\xi|),$$

$$\Phi(|\xi|^2)\,|\nabla\phi(\xi)|\,|f(\xi,d)| \leq \max\left\{s(|\xi|), \Phi(|\xi|^2)\nu_2(|\xi|)\right\}\nu_3(|\xi|).$$

Define $\breve{\alpha}_3$, $\breve{\gamma}$, $\breve{\alpha}_1$, and $\breve{\alpha}_2$ by

$$\breve{\alpha}_3(r) := \pi_2(\underline{\alpha}(r))\Xi_1(r),$$

$$\begin{aligned}\breve{\gamma}(r) := {} & 2\pi_2((3\rho(r))^2)3\rho(r)\nu_3(3\rho(r)) \\ & + \max\left\{s(3\rho(r)), \Phi((3\rho(r))^2)\nu_2(3\rho(r))\right\}\nu_3(3\rho(r)) + \breve{\alpha}_3(3\rho(r)),\end{aligned}$$

$$\check{\alpha}_1(r) = \min\left\{\Phi(r^2), \Phi \circ \underline{\alpha}(r)\right\},$$

and

$$\check{\alpha}_2(r) = \max\left\{\Phi(r^2), \Phi \circ \bar{\alpha}(r)\right\}.$$

Then inequalities

$$\check{\alpha}_1(|x|) \le V_2(x) \le \check{\alpha}_2(|x|) \quad \forall\, x \neq 0 \tag{4.67}$$

and (4.64) hold for V_2. Define $V_2(0) := 0$. Then inequality (2.3) holds for V_2 with $\alpha_1 := \check{\alpha}_1$, $\alpha_2 := \check{\alpha}_2$ on $\mathbf{X}$, and in particular implies that V_2 is continuous as 0. Then, V_2 is a UOSS-Lyapunov-like function for Σ, smooth away from the origin. □

Recall that $V_2(\xi) \equiv \Phi(V_1(\xi))$ for all ξ with $|\xi| > 3\rho(|h(\xi)|)$. Therefore inequality (4.62) implies that

$$\nabla V_2 \cdot f(\xi, d) \le -\alpha_3(|\xi|)\ \forall\, |\xi| > 3\rho(|h(\xi)|),\ \forall d \in \Omega. \tag{4.68}$$

Proposition 4.24. *Suppose that a system Σ of type* (2.14) *admits a continuous UOSS-Lyapunov-like function V_2, smooth away from* 0 *and satisfying inequalities* (4.67), (4.64)*, and* (4.68)*. Then Σ admits a UOSS-Lyapunov function.*

The basic idea used to obtain a Lyapunov function, smooth on the whole $\mathbf{X}$, is composing V_2 with some appropriately chosen $\mathcal{K}_\infty$-function β. This technique was previously utilized in [25]. We will need the following generalization of Lemma 4.3 from [25], where this function β is constructed. In our setting we also need the derivative of β to be of class $\mathcal{K}_\infty$, which was not required in [25]. The proof is only a slight modification of the proof of the mentioned lemma.

Lemma 4.25. *Assume that $V : \mathbf{R}^n \longrightarrow \mathbf{R}_{\ge 0}$ is C^0, positive definite, and the restriction $V|_{\mathbf{R}^n\setminus\{0\}}$ is C^∞.*

Then, given any $m \in \mathbb{N}$ there exists a $\mathcal{K}_\infty$-function β_m, smooth on $(0,\infty)$, satisfying the following conditions:

- *$\beta_m^{(i)}(t) \to 0$ as $t \to 0^+$ for each $i = 0, 1, \ldots$,*
- *$\beta_m^{(i)} \in \mathcal{K}_\infty\ \forall i \le m$,*
- *$W_m := \beta \circ V$ is a C^∞ function on all of $\mathbf{R}^n$.*

We now return to the proof of Proposition 4.24.

Proof. Take the function V_2, and apply Lemma 4.25 to get a function β_1, with derivative in class $\mathcal{K}$, such that

$$V_3 := \beta_1 \circ V_2$$

is smooth on $\mathbf{X}$. Let $\alpha_1(\cdot) := \beta_1(\check{\alpha}_1(\cdot)/2)$, $\alpha_2(\cdot) := \beta_1(\check{\alpha}_2(\cdot) + \check{\alpha}_1(\cdot)/2)$. Then $\alpha_i \in \mathcal{K}_\infty$ and

$$\alpha_1(|x|) \le V_3(x) \le \alpha_2(|x|).$$

Furthermore, for any $x \in \mathbf{X} \setminus \{0\}$ we have

$$\begin{aligned} L_{f_d} V_3(x) &= \beta_1'(V_2(x))\,[L_{f_d} V_2]\,(x) \\ &\le \beta_1'(V_2(x))\,(-\alpha_3(|x|)/2 + \check{\gamma}(|h(x)|)) \\ &\le -\beta_1'(\check{\alpha}_1(|x|)/2)\check{\alpha}_3(|x|)/2 + \beta_1'(\check{\alpha}_2(|x|) + \check{\alpha}_1(|x|)/2)\check{\gamma}(|h(x)|). \end{aligned}$$

Recall that, because of (4.68), the $\check{\gamma}$ term in the last estimate can be dropped if $|x| > 3\rho(|h(x)|)$, so that we can write

$$L_{f_d} V_3(x) \leq -\beta_1'(\check{\alpha}_1(|x|)/2)\check{\alpha}_3(|x|)/2 + \beta_1'(\check{\alpha}_2(3\rho(|h(x)|)) + \check{\alpha}_1(3\rho(|h(x)|))/2)\check{\gamma}(|h(x)|).$$

Thus V_3 is a smooth UOSS-Lyapunov function, satisfying the dissipation inequality with

$$\alpha_3(\cdot) := \beta_1'(\check{\alpha}_1(\cdot)/2)\,\check{\alpha}_3(\cdot)/2$$

and

$$\gamma(\cdot) := \beta_1'(\check{\alpha}_2(3\rho(\cdot)) + \check{\alpha}_1(3\rho(\cdot))/2)\,\check{\gamma}(\cdot).$$

This completes the construction. □

5. Norm-observers.

As conjectured in [45] and proved in this presentation, every IOSS system admits an IOSS-Lyapunov function. One of the main motivations for the notion of IOSS, and for deriving Lyapunov characterizations, is the fact that a Lyapunov function enables us to get insights into the behavior of control systems. In particular, it may be useful to have an estimate of how far the system is from the equilibrium at any given time, and in some situations this "norm-estimate" is sufficient for the design of a stabilizer. We next provide a construction for a norm-observer in the most general case—for systems of type (2.1), assuming that we have a smooth UIOSS-Lyapunov function at our disposal.

5.1. Exponential decay Lyapunov functions.

DEFINITION 5.1. *Let Σ be a system of type* (2.1). *A C^1-function $V : \mathbf{X} \to \mathbf{R}_{\geq 0}$ is an* exponential decay UIOSS-Lyapunov function *for Σ if it satisfies* (2.3) *with some α_1 and α_2, and the following version of inequality* (2.4)*:*

$$\text{(5.1)}\quad \nabla V(x) \cdot f(x,u,w) \leq -V(x) + \sigma_1(|u|) + \sigma_2(|h(x)|) \quad \forall x \in \mathbf{X},\ u \in \mathbb{U},\ w \in \Gamma$$

holds with some σ_1 and $\sigma_2 \in \mathcal{K}$.

LEMMA 5.2. *Suppose V is a UIOSS-Lyapunov function for a system Σ of type* (2.1), *satisfying inequality* (2.4). *Then there exists a $\mathcal{K}_\infty$-function ρ such that a function $W := \rho \circ V$ is an exponential decay UIOSS-Lyapunov function for Σ.*

Proof. Assume that system (2.1) admits a UIOSS-Lyapunov function with α_i ($i = 1, 2$) as in (2.3) and with $\alpha, \sigma_1, \sigma_2$ as in (2.4). Replacing α by $\alpha \circ \alpha_2^{-1}$, we have

$$\text{(5.2)}\qquad \frac{d}{dt}V(x(t,\xi,\mathbf{u})) \;\leq\; -\alpha(V(x(t,\xi,\mathbf{u}))) + \sigma_1(|\mathbf{u}(t)|) + \sigma_2(|y(t,\xi,\mathbf{u})|)$$

for almost all $t \in [0,\ t_{\max}(\xi, u))$. According to Lemma 12 in [32], there exists some function $\rho \in \mathcal{K}_\infty$ which can be extended as a C^1-function to a neighborhood of $[0, \infty)$ such that $\rho'(r)\frac{\alpha(r)}{2} \geq \rho(r)$ for all $r \geq 0$. Consider the function $W(\xi) := \rho(V(\xi))$. Observe that W is again proper and positive definite. Along any trajectory $x(t) := x(t,\xi,\mathbf{u})$ (with $y(t) := y(t,\xi,\mathbf{u})$), at any point where (2.4) holds, one has that $\frac{d}{dt}W(x(t)) = \rho'(V(x(t)))\frac{d}{dt}V(x(t))$ is upper bounded by

$$-\rho'(V(x(t)))\frac{\alpha(V(x(t)))}{2} + \rho'(V(x(t)))\left(-\frac{\alpha(V(x(t)))}{2} + \sigma_1(|\mathbf{u}(t)|) + \sigma_2(|y(t)|)\right),$$

which in turn is bounded by

$$-\rho(V(x(t)) + \rho'(V(x(t)))\left(-\frac{\alpha(V(x(t)))}{2} + \sigma_1(|\mathbf{u}(t)|) + \sigma_2(|y(t)|)\right). \tag{5.3}$$

Observe that when $V(x(t)) \geq \alpha^{-1}(2\sigma_1(|\mathbf{u}(t)|) + 2\sigma_2(|y(t)|))$ it holds that

$$\rho'(V(x(t)))\left(-\frac{\alpha(V(x(t)))}{2} + \sigma_1(|\mathbf{u}(t)|) + \sigma_2(|y(t)|)\right) \leq 0, \tag{5.4}$$

while if instead $V(x(t)) \leq \alpha^{-1}(2\sigma_1(|\mathbf{u}(t)|) + 2\sigma_2(|y(t)|))$, then

$$\rho'(V(x(t)))\,(\sigma_1(|\mathbf{u}(t)|) + \sigma_2(|y(t)|)) \;\leq\; \hat{\sigma_1}(|\mathbf{u}(t)|) + \hat{\sigma_2}(|y(t)|) \tag{5.5}$$

for some $\mathcal{K}_\infty$-functions $\hat{\sigma_1}$ and $\hat{\sigma_2}$ (using here the fact that $\rho'(s)$ is a continuous function). Combining (5.4) and (5.5), one concludes from the estimate (5.3) on $\frac{d}{dt}W(x(t))$ that

$$\frac{d}{dt}W(x(t)) \;\leq\; -W(x(t)) + \hat{\sigma_1}(|\mathbf{u}(t)|) + \hat{\sigma_2}(|y(t)|)$$

for almost all $t \in [0, t_{\max})$. □

5.2. Construction of a norm-observer.

PROPOSITION 5.3. *Suppose that a system Σ admits an exponential decay UIOSS-Lyapunov function V. Then the pair $(\Sigma_{n.o}, k)$, where*

$$\Sigma_{n.o}: \quad \dot{p} = -p + \sigma_1(|u|) + \sigma_2(|y|), \tag{5.6}$$

with σ_1 and σ_2 as in (5.1) *and $k(\cdot,\cdot)$ defined by $k(s,r) = s$, is a norm-estimator for Σ.*

Proof. Assume without loss of generality that the function α_2 in the definition of V satisfies $r \leq \alpha_2(r)$ for all nonnegative r.

The system (5.6) is ISS with respect to u and y, since it can be seen as an asymptotically stable linear system driven by the input $(\sigma_1(|u|), \sigma_2(|y|))$, so, inequality (2.7) obviously holds. Pick any initial states ξ, ζ of Σ and (5.6), respectively, any control $\mathbf{u}$, and any disturbance $\mathbf{w}$. Consider the resulting trajectory $(x(t), p(t))$ of the composite system. Property (5.1) implies that

$$\frac{d}{dt}(V(x(t)) - p(t)) \leq -(V(x(t)) - p(t)) \tag{5.7}$$

for almost all $t \in [0, t_{\max}(\xi, \mathbf{u}, \mathbf{w}))$. Thus

$$V(x(t)) \;\leq\; p(t) + e^{-t}(V(\xi) - \zeta) \;\leq\; |p(t)| + 2e^{-t}\alpha_2(|\xi| + |\zeta|)$$

(using $r \leq \alpha_2(r)$). This can be written as (2.8) with $\rho := \alpha_1^{-1}(2(\cdot))$ and $\beta(s,t) := \alpha_1^{-1}(4e^{-t}\alpha_2(s))$. □

Implication 2 ⇒ 3 of Theorem 2.4 now follows from Lemma 5.2 and Proposition 5.3.

We now turn to the proof of implication 3 ⇒ 1 of Theorem 2.4.

Proof. Assume that $(\Sigma_{n.o}, k)$ is some norm-estimator for Σ. Choose any initial state ξ for Σ, any input $\mathbf{u}$, disturbance $\mathbf{w}$, and the special initial state $\zeta = 0$ for $\Sigma_{n.o}$. Then inequality (2.7) becomes

$$|k(p(t, 0, \mathbf{u}, \mathbf{y}_{\xi,\mathbf{u},\mathbf{w}}), y(t, \xi, \mathbf{u}, \mathbf{w}))| \;\leq\; \hat{\gamma}_1\left(\left\|\mathbf{u}|_{[0,t]}\right\|\right) + \hat{\gamma}_2\left(\left\|\mathbf{y}_{\xi,\mathbf{u},\mathbf{w}}|_{[0,t]}\right\|\right) \tag{5.8}$$

for all $t \in [0, t_{\max})$, with some class-$\mathcal{K}$ functions $\hat{\gamma}_1$ and $\hat{\gamma}_2$ (the $\mathcal{KL}$-term vanishes because $\zeta = 0$). Then, combining (5.8) with the estimate (2.8) we get

$$\begin{aligned} |x(t,\xi,\mathbf{u},\mathbf{w})| &\le \beta(|\xi|,t) + \rho(|k(p(t,0,\mathbf{u},\mathbf{y}_{\xi,\mathbf{u},\mathbf{w}}), y(t,\xi,\mathbf{u},\mathbf{w}))|) \\ &\le \max\left\{2\beta(|\xi|,t),\, 4\rho(\hat{\gamma}_1(\left\|\mathbf{u}|_{[0,t]}\right\|)),\, 4\rho(\hat{\gamma}_1(\left\|y_{\xi,\mathbf{u},\mathbf{w}}|_{[0,t]}\right\|))\right\}. \end{aligned}$$

This proves the UIOSS property for Σ. □

6. Pointers for future research.

6.1. Integral variants of UIOSS. The UIOSS property gives uniform estimates of the states in terms of the uniform bounds on outputs and essential bounds on controls. A natural question to ask is what property will result if instead of the uniform (or essential) bounds we use other "finite energy" concepts, such as, for example, L^γ-type norms (defined by $\left\|g|_{[\sigma,\tau]}\right\|_\gamma = \int_\sigma^\tau \gamma(|g(t)|)dt$) of inputs and/or outputs, where the "γ's" for inputs and outputs are some appropriately chosen functions of class $\mathcal{K}_\infty$, which depend on the system. For systems without controls, the iiUOSS property provides a "finite energy outputs ⇒ finite energy state" characterization, which is "almost" equivalent to UOSS (see Theorem 2.16). Searching for the *uniform* estimate of the states in terms of L^γ-norms of inputs and outputs leads to the following definition.

DEFINITION 6.1. *A system of type* (2.1) *is* uniformly integral input-output-to-state stable *(UiIOSS) if there exist functions $\alpha_x \in \mathcal{K}_\infty$, $\beta \in \mathcal{KL}$, γ_1, and $\gamma_2 \in \mathcal{K}$ such that*

$$\alpha_x\left(|x(t,\xi,\mathbf{w},\mathbf{u})|\right) \le \beta(|\xi|,t) + \int_0^t \left(\gamma_1(|\mathbf{u}(s)|) + \gamma_2\left(|y(s,\xi,\mathbf{w},\mathbf{u})|\right)\right) ds \tag{6.1}$$

for all $\xi \in \mathbf{X}$, all $\mathbf{w}$ and $\mathbf{u}$, and all $t \in [0, t_{\max}(\xi,\mathbf{u},\mathbf{w}))$.

This general definition may be adjusted in obvious manners to all the particular cases of system (2.1).

Remark 6.1. It is easy to see that estimate (6.1) is equivalent to the following estimate (with different bounding functions, of course):

$$\begin{aligned} |x(t,\xi,\mathbf{w},\mathbf{u})| \le \max\Bigg\{&\beta(|\xi|,t),\, \gamma\left(\int_0^t (\gamma_1(|\mathbf{u}(s)|))ds\right), \\ &\gamma\left(\int_0^t \gamma_2(|y(s,\xi,\mathbf{w},\mathbf{u})|)ds\right)\Bigg\}, \end{aligned} \tag{6.2}$$

In the particular case of systems without outputs and disturbances, a Lyapunov characterization of the UiIOSS property, reduced to integral input-to-state stability (iISS), was obtained in [4]. By repeating the proof of the implication 1 ⇒ 2 of Theorem 1 in [4] one can show that a system of type (2.1) will be UiIOSS if it admits a smooth, proper Lyapunov function $V : \mathbf{X} \to R_{\ge 0}$, satisfying inequality (2.4) with some σ_1 and σ_2 of class $\mathcal{K}$, and a *positive definite* function α. Whether or not this sufficient condition is also necessary for UiIOSS is not known. Notice, however, that this condition is weaker than the corresponding property for UIOSS, as the dissipation condition for a UIOSS-Lyapunov function requires α to be of class $\mathcal{K}_\infty$. Thus, any UIOSS system will also be UiIOSS. The converse implication is not true, as demonstrated in the following example.

Example 6.2. The construction is similar to the one used in Remark 3.1, so, we recall that $\phi_\varepsilon(\cdot)$ denotes a C^∞-bump function as in (3.12), and $1_A(\cdot)$ is the indicator function of a set A.

Choose $\varepsilon_f = 0.1$ and any $\varepsilon_h < (1-\varepsilon_f)e^{-1}$, and consider the autonomous system

$$\Sigma_1 : \quad \dot{x} = f(x); \quad y = h(x)$$

with

$$f_1(x) = x\left[1_{(-\infty,-1]}(x)(1-\phi_{\varepsilon_f}(x+1)) + 1_{[1,+\infty)}(x)(1-\phi_{\varepsilon_f}(x-1))\right] \\ -x\left[1_{(-1,1)}(x)(1-\phi_{\varepsilon_f}(x+1))(1-\phi_{\varepsilon_f}(x-1))\right],$$

and

$$h_1(x) = 1 - \phi_{\varepsilon_h}(x)$$

evolving in $\mathbf{X} = \mathbf{R}$. We claim that Σ_1 is iOSS but not OSS.

Consider also an autonomous system Σ_2 on $\mathbf{R}$ with

$$f_2(x) = x; \qquad h_2(x) \equiv 1. \tag{6.3}$$

This system cannot serve as a counterexample, because $h_2(0) \neq 0$. However, its behavior away from 0 is identical to that of Σ_1, so that considering it will simplify the presentation.

Let $x_i(t,\xi)$, $i = 1,2$, denote the solutions of Σ_i, starting at ξ, and let $y_i(t,\xi)$ denote the corresponding output trajectories. It is easy to see that both Σ_1 and Σ_2 are forward complete.

Since the dynamics of Σ_1 and Σ_2 are odd functions and outputs are even (but only the magnitudes of outputs are involved in the estimates), we need only to consider trajectories starting from the positive initial states, as the same estimates will work for trajectories contained in the other half-line.

Since $f_1(x) < f_2(x)$ for all $x \in (0, 1+\varepsilon_f)$ and $f_1(x) = f_2(x)$ for $x \geq 1+\varepsilon_f$, we have, for all $t > 0$, $x_1(t,\xi) < x_2(t,\xi)$ for any $\xi \in (0, 1+\varepsilon_f)$, and $x_1(t,\xi) = x_2(t,\xi) = \xi e^t$ if $|\xi| \geq 1+\varepsilon_f$. In particular, this shows that Σ_1 is not OSS, because the trajectory diverges to $+\infty$, but $h_1(\xi e^t)$ is bounded.

However, observe that when $t \leq |\xi|$,

$$|x_2(t,\xi)| \leq |\xi|\, e^{|\xi|} \leq e^{2|\xi|}\, |\xi|\, e^{-\frac{t}{1+|\xi|}},$$

whereas when $t > |\xi|$, we have

$$|x_2(t,\xi)| \leq te^t.$$

Letting

$$\beta(r,t) := re^{2r-\frac{t}{1+r}}; \quad \gamma_2 := Id; \quad \gamma(t) := te^t,$$

and noticing that

$$\int_0^t \gamma_2(|y_2(s,\xi)|)ds = t,$$

we conclude $x_2(t,\xi)$ satisfies estimate (6.2).

Now observe that

- If $\xi \geq 1$, then $1 \leq x_1(t,\xi) \leq x_2(t,\xi)$ for all $t \geq 0$, so that $y_1(t,\xi) = y_2(t,\xi) = 1$ and $x_1(t,\xi)$ satisfies (6.2).

- If $\xi \in [0, 1 - \varepsilon_f]$, then

$$x_1(t, \xi) = \xi e^{-t} \leq \beta(|\xi|, t).$$

- Finally, if $\xi \in (1 - \varepsilon_f, 1)$, then
  - for all $t \geq 0$ it holds that $|x_1(t, \xi)| < 1$;
  - for all $t \in [0, 1]$ it holds that $|x_1(t, \xi)| > (1 - \varepsilon_f)e^{-1} > \varepsilon_h$, so that $y_1(t, \xi) = 1 = y_2(t, \xi)$.

  Therefore $x_1(t, \xi)$ satisfies (6.2) for all $t \leq 1$ and

$$\int_0^t \gamma_2(|y_2(s, \xi)|)ds \geq 1 \geq x_1(t, \xi) \quad \forall\, t \geq 1.$$

This shows that $x_1(t, \xi)$ satisfies (6.2) for all ξ and $t \geq 0$, thus, Σ_1 is, indeed, iOSS.

It could also be of interest to consider yet another property, providing a uniform estimate for the state in terms of the uniform norm of the output and L^γ-norm of the input.

Definition 6.2. *A system of type* (2.1) *satisfies the U(iI)OSS property if there exist functions $\alpha_x \in \mathcal{K}_\infty$, $\beta \in \mathcal{KL}$, γ_1, and $\gamma_2 \in \mathcal{K}$ such that*

$$\alpha_x\left(|x(t, \xi, \mathbf{w}, \mathbf{u})|\right) \leq \beta(|\xi|, t) + \int_0^t \gamma_1(|\mathbf{u}(s)|)ds + \gamma_2\left(\left\|y_{\xi,\mathbf{w},\mathbf{u}}|_{[0,t]}\right\|\right) \tag{6.4}$$

for all $\xi \in \mathbf{X}$, all $\mathbf{w}$ and $\mathbf{u}$, and all $t \in [0, t_{\max}(\xi, \mathbf{u}, \mathbf{w}))$.

Deriving a Lyapunov characterization for this property may be a challenge. It would be logical to expect that a good candidate for a U(iI)OSS-Lyapunov function can be a proper function $V : \mathbf{X} \to \mathbf{R}_{\geq 0}$, satisfying inequality (2.4) with α being either positive definite or class $\mathcal{K}_\infty$. However, we can see right away that neither of these two possibilities is the right guess. Indeed, notice that both OSS and iISS are particular cases of the U(iI)OSS property. If a U(iI)OSS-Lyapunov function would satisfy (2.4) with α of class $\mathcal{K}_\infty$ (as required by OSS), it would follow that every disturbance-free U(iI)OSS system without outputs is ISS, which is not true (see [4] for an example of an iISS system, which is not ISS). On the other hand, if a U(iI)OSS-Lyapunov function satisfied (2.4) with α positive definite (as required by iISS), it would imply that having such a dissipation function is sufficient for OSS, which is not so, because this would mean that every iOSS system is OSS.

6.2. Incremental IOSS. As mentioned in the introduction, the detectability property for nonlinear systems is not equivalent to zero-detectability. In searching for a correct notion for nonlinear detectability one could think of the following generalization of UIOSS.

Definition 6.3. *A system* (2.1) *is incrementally uniformly input-output to state stable (ΔUIOSS) if there exists some $\beta \in \mathcal{KL}$ and $\gamma_1, \gamma_2 \in \mathcal{K}$ such that, for every two initial states ξ_1 and ξ_2, any two controls $\mathbf{u}_1$ and $\mathbf{u}_2$, and any disturbance $\mathbf{w}$,*

$$\begin{aligned} |x(t, \xi_1, \mathbf{u}_1, \mathbf{w}) - x(t, \xi_2, \mathbf{u}_2, \mathbf{w})| \leq \max\big\{&\beta(|\xi_1 - \xi_2|, t), \\ &\gamma_1\left(\left\|(\mathbf{u}_1 - \mathbf{u}_2)|_{[0,t]}\right\|\right), \gamma_2\left(\left\|(\mathbf{y}_{\xi_1,\mathbf{u}_1,\mathbf{w}} - \mathbf{y}_{\xi_2,\mathbf{u}_2,\mathbf{w}})|_{[0,t]}\right\|\right)\big\} \end{aligned} \tag{6.5}$$

for all t in the common domain of definition.

Deriving a right Lyapunov characterization for this property may lead to a construction of a full-order observer.

REFERENCES

[1] F. Albertini and E. D. Sontag, *Continuous control-Lyapunov functions for asymptotically controllable time-varying systems*, Internat. J. Control, 72 (1999), pp. 1630–1641.

[2] D. Angeli, *Intrinsic robustness of global asymptotic stability*, Systems Control Lett., 38 (1999), pp. 297–307.

[3] D. Angeli and E. D. Sontag, *Forward completeness, unboundedness observability, and their Lyapunov characterizations*, Systems Control Lett., 38 (1999), pp. 209–217.

[4] D. Angeli, E. D. Sontag, and Y. Wang, *A characterization of integral input to state stability*, IEEE Trans. Automat. Control, 45 (2000), pp. 1082–1097.

[5] Z. Artstein, *Relaxed controls and the dynamics of control systems*, SIAM J. Control Optim., 16 (1978), pp. 689–701.

[6] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd ed., Academic Press, Orlando, FL, 1986.

[7] P. D. Christofides and A. Teel, *Singular perturbations and input-to-state stability*, IEEE Trans. Automat. Control, 41 (1996), pp. 1645–1650.

[8] F. Clarke, Y. Ledyaev, R. Stern, and P. Wolensky, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[9] F. Clarke, Y. S. Ledyaev, E. Sontag, and A. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[10] J. W. Helton and M. R. James, *Extending H_∞ Control to Nonlinear Systems*, SIAM, Philadelphia, 1999.

[11] J. Hespanha and A. Morse, *Certainty equivalence implies detectability*, Systems Control Lett., 36 (1999), pp. 1–13.

[12] D. Hill and P. Moylan, *Dissipative dynamical systems: Basic input-output and state properties*, J. Franklin Institute, 5 (1980), pp. 327–357.

[13] D. J. Hill and P. Moylan, *Dissipative nonlinear systems: Basic properties and stability analysis*, in Proceedings of the 31st IEEE Conference on Decision and Control, Tucson, AZ, IEEE, Piscataway, NJ, 1992, pp. 3259–3264.

[14] X. M. Hu, *On state observers for nonlinear systems*, Systems Control Lett., 17 (1991), pp. 645–473.

[15] A. Isidori, *Global almost disturbance decoupling with stability for non-minimum-phase single-input single-output nonlinear systems*, Systems Control Lett., 28 (1996), pp. 115–122.

[16] A. Isidori, *Nonlinear Control Systems* II, Springer-Verlag, London, 1999.

[17] M. James, *A partial differential inequality for dissipative nonlinear systems*, Systems Control Lett., 21 (1993), pp. 315–320.

[18] Z.-P. Jiang and L. Praly, *Preliminary results about robust Lagrange stability in adaptive nonlinear regulation*, Internat. J. Control, 6 (1992), pp. 285–307.

[19] Z.-P. Jiang, A. Teel, and L. Praly, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1994), pp. 95–120.

[20] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.

[21] M. Krichman and E. D. Sontag, *A version of a converse Lyapunov theorem for input-output to state stability*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, IEEE, Piscataway, NJ, 1998, pp. 4121–4126.

[22] M. Krstić and H. Deng, *Stabilization of Uncertain Nonlinear Systems*, Springer-Verlag, London, 1998.

[23] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley and Sons, New York, 1995.

[24] D. Liberzon, E. D. Sontag, and Y. Wang, *On integral-input-to-state stabilization*, in Proceedings of the 1999 American Control Conference, San Diego, 1999, pp. 1598–1602.

[25] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[26] W. M. Lu, *A class of globally stabilizing controllers for nonlinear systems*, Systems Control Lett., 25 (1995), pp. 13–19.

[27] W. M. Lu, *A state-space approach to parameterization of stabilizing controllers for nonlinear systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1576–1588.

[28] R. Marino and P. Tomei, *Nonlinear output feedback tracking with almost disturbance decoupling*, IEEE Trans. Automat. Control, 44 (1999), pp. 18–28.

[29] F. Mazenc, L. Praly, and W. Dayawansa, *Global stabilization by output feedback: Examples and counterexamples*, Systems Control Lett., 23 (1994), pp. 119–125.

[30] A. S. Morse, *Control using logic-based switching*, in Trends in Control: A European Perspective, A. Isidori, ed., Springer-Verlag, London, 1995, pp. 69–114.

[31] D. J. Pan, Z. Z. Han, and Z. J. Zhang, *Bounded-input-bounded-output stabilization of nonlinear systems using state detectors*, Systems Control Lett., 21 (1993), pp. 189–198.

[32] L. Praly and Y. Wang, *Stabilization in spite of matched unmodelled dynamics and an equivalent definition of input-to-state stability*, Math. Control Signals Systems, 9 (1996), pp. 1–33.

[33] R. Sepulchre, M. Jankovic, and P. V. Kokotović, *Integrator forwarding: A new recursive nonlinear robust design*, Automatica J. IFAC, 33 (1997), pp. 979–984.

[34] E. D. Sontag, *Conditions for abstract nonlinear regulation*, Inform. Control, 51 (1981), pp. 105–127.

[35] E. D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.

[36] E. D. Sontag, *Some connections between stabilization and factorization*, in Proceedings of the IEEE Conference on Decision and Control, Tampa, FL, 1989, IEEE, Piscataway, NJ, 1989, pp. 990–995.

[37] E. D. Sontag, *On the input-to-state stability property*, Euro. J. Control, 1 (1995), pp. 24–36.

[38] E. D. Sontag, *Comments on integral variants of input-to-state stability*, Systems Control Lett., 34 (1998), pp. 93–100.

[39] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, 2nd ed., Springer-Verlag, New York, 1998.

[40] E. D. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Proceedings of the NAYO ASI Nonlinear Analysis, Differential Equations and Control, Montreal, 1998, F. Clarke and R. Stern, eds., Kluwer, Dordrecht, The Netherlands, 1999, pp. 551–598.

[41] E. D. Sontag, *The ISS philosophy as a unifying framework for stability-like behavior*, in Nonlinear Control in the Year 2000, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Lecture Notes in Control and Inform. Sci. 258, Springer-Verlag, Berlin, 2000.

[42] E. D. Sontag and Y. Wang, *On characterizations of the input-to-state stability property*, Systems Control Lett., 24 (1995), pp. 351–359.

[43] E. D. Sontag and Y. Wang, *Detectability of nonlinear systems*, in Proceedings of the Conference on Information Science and Systems (CISS 96), Princeton, NJ, 1996, pp. 1031–1036.

[44] E. D. Sontag and Y. Wang, *New characterizations of the input to state stability property*, IEEE Trans. Automat. Control, 41 (1996), pp. 1283–1294.

[45] E. D. Sontag and Y. Wang, *Output-to-state stability and detectability of nonlinear systems*, Systems Control Lett., 29 (1997), pp. 279–290.

[46] E. D. Sontag and Y. Wang, *Notions of input to output stability*, Systems Control Lett., 38 (1999), pp. 351–359.

[47] E. D. Sontag and Y. Wang, *Lyapunov characterizations of input to output stability*, SIAM J. Control Optim., 39 (2000), pp. 226–249.

[48] J. Tsinias, *Sontag's "input to state stability condition" and global stabilization using state detection*, Systems Control Lett., 20 (1993), pp. 219–226.

[49] J. Tsinias, *Input to state stability properties of nonlinear systems and applications to bounded feedback stabilization using saturation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 57–85.

[50] A. van der Schaft, *L^2-gain analysis of nonlinear systems and nonlinear state feedback H^∞-control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

[51] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[52] J. C. Willems, *Mechanisms for the stability and instability in feedback systems*, Proc. IEEE, 64 (1976), pp. 24–35.

SIAM J. Control Optim.
Vol. 39, No. 6, pp. 1929–1951

RELATIVE FLATNESS AND FLATNESS OF IMPLICIT SYSTEMS*

PAULO SÉRGIO PEREIRA DA SILVA† AND CARLOS CORRÊA FILHO‡

Abstract. In this work we define the concept of *relative flatness* of a system with respect to a subsystem. The subsystem associated to a set of outputs of a system is constructed, and called here *output subsystem.* It is shown that the relative flatness of a system with respect to the output subsystem implies the flatness of the corresponding implicit system obtained by setting these outputs to zero. A sufficient condition of relative flatness based on a *relative derived flag* is presented. Based on these results, a sufficient condition for the flatness of a class of nonlinear implicit systems is obtained.

Key words. nonlinear systems, implicit systems, time-varying systems, flatness, relative flatness, feedback linearization

AMS subject classifications. 93C05, 93B18, 93B29.

PII. S0363012998349832

1. Introduction and motivation.

The aim of this paper is to present the notion of *relative flatness* with respect to a subsystem. We show that this concept may be useful for control systems theory, in particular for studying the structure of nonlinear implicit systems. Our approach is based on the infinite dimensional geometric setting recently introduced in control theory [17, 40, 19] in combination with the ideas presented in [50, 48, 53]. Our sufficient conditions for flatness of implicit systems may be regarded as a generalization of the conditions obtained in [50] for explicit systems. Our setting has some connections with the ideas of [47], which has considered a different class of implicit systems.

Feedback linearization is an important problem in nonlinear control theory. This problem was completely solved in the static-state feedback case [25, 23] but necessary and sufficient conditions for feedback linearizability by dynamic state feedback are not yet known (see [5, 48, 6, 20, 51, 53, 1, 44, 52, 22, 41, 54]).

The notion of differential flatness was introduced by Fliess et al. [16, 18] and is strongly related to the problem of feedback linearization. This concept corresponds to a complete and finite parametrization of all solutions of a control system by a differentially independent family of functions called flat output.

Linear *singular* (or implicit) systems are an important class of control systems and many papers and books on this subject are found in the literature [4, 31].[1] Solvability of nonlinear implicit differential equations is considered in [2, 43]. Other problems like controllability [29], stabilization [32, 7], canonical forms [45], and feedback control [8] have already been considered.

*Received by the editors December 23, 1998; accepted for publication (in revised form) October 26, 2000; published electronically April 24, 2001. A preliminary version of this paper appeared in *Proceedings of the 4th IFAC Nonlinear Control Systems Design Symposium*, Enschede, The Netherlands, Vol. 2, 1998, Pergamon, New York, 1998, pp. 516–522.

http://www.siam.org/journals/sicon/39-6/34983.html

†Escola Politécnica da USP - PTC, São Paulo, SP 05508-900, Brazil (paulo@lac.usp.br). The work of this author was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under grant 300492/95-2 and Fundação de Amparo à Pesquisa do Estado de São Paulo FAPESP under grant 97 / 04668-1.

‡Faculdade SENAC de Ciências Exatas e Tecnologia, Rua Galvão Bueno, 430, São Paulo, SP, Brazil (carlos@cei.sp.senac.br).

[1] Note that the *module theoretic* approach of [15] is also valid for implicit systems.

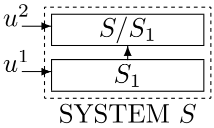


FIG. 1.1. *Structure of a system S with respect to a subsystem S_1.*

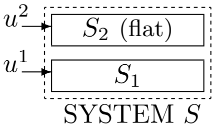


FIG. 1.2. *Structure of a system S that is relatively flat with respect to a subsystem S_1. Note that S_2 is flat.*

Feedback linearization of implicit systems has been studied for instance in [30, 26]. These works consider the problem of finding a state transformation and a state feedback such that the closed loop system is a linear singular system. In this work we tackle the problem of finding sufficient conditions for flatness of a class of time-varying implicit systems of the form[2]

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{1.1a}$$

$$y(t) = h(t, x(t), u(t)) = 0, \tag{1.1b}$$

where $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^p$, $u(t) \in \mathbb{R}^m$ and all the components of $f(x)$ and $g(x)$ are analytical functions of x. We stress that a set of implicit differential equations of arbitrary order can be put into the form (1.1a)–(1.1b) [39].

We now present, without being precise, a summary of the ideas and the results of this paper. Roughly speaking, a subsystem S_1 of a system S is some part of S that may be considered as a system by itself. Note that S_1 may affect the "quotient system" S/S_1, but it is not affected by S/S_1 as depicted in Figure 1.1.

Remark 1.1. We stress that, in Figure 1.1, S/S_1 is not a subsystem.

Recall that a system is *flat* if and only if there exists a differentially independent set of functions $y = (y_1, \ldots, y_m)$, called the *flat output*, such that every variable of the system is a function of the flat output and its derivatives. A system S is said to be *relatively flat with respect to a given subsystem* S_1 if, after a convenient *endogenous feedback*, S is decomposed into two independent subsystems S_1 and S_2 such that S_2 is a flat system[3] (see Figure 1.2). We stress that the fact that the system is decomposed into two independent subsystems is not artificial since the same structure occurs for the algebraic counterpart of this definition (see Remark 5.1).

In this paper, a sufficient condition for relative flatness is given (see Theorem 8.2). One can easily conclude that a system S that is relatively flat with respect to a flat subsystem is also flat,[4] leading to a sufficient condition of flatness of system S.

Now, let y be the output (not necessarily a flat output) of system S. We will show that one can construct a subsystem Y of S such that Y contains only the "information" of time and of y and its derivatives $y^{(k)}, k \in \mathbb{N}$ (see Theorem 4.3). Subsystem Y will be called *output subsystem.*

[2] This class is more general than the one considered by [30, 26].

[3] See Definition 5.1 for a precise statement of *relative flatness.*

[4] See Proposition 5.2 for a precise statement of this idea.

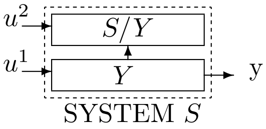


FIG. 1.3. *Structure of a system S with respect to the output subsystem Y.*

The structure of the implicit system obtained from S by setting y to be equal to zero (see Figure 1.3) is directly related to the properties of S with respect to the output subsystem Y. Under some regularity assumptions, if S is relatively flat with respect to Y, then the implicit system obtained from S by including the constraint $y = 0$ is also flat.[5]

The paper is organized as follows. In section 2 the notation and some mathematical background are presented. The infinite dimensional differential geometric approach of [19] is briefly summarized in section 3. The notion of *subsystem* is presented in section 4. The existence and some properties of local *output* subsystems are also discussed in section 4. The concept of *relative flatness* is discussed in section 5. In section 6 it is shown that, under regularity assumptions, an implicit system (1.1a)–(1.1b) may be considered as a system that is immersed in the explicit system (1.1a). In section 7, the results of the previous sections are used to derive a sufficient condition for flatness of implicit systems. A sufficient condition for relative flatness based on *relative derived flags* is developed in section 8. Some examples are discussed in section 9. Finally, some auxiliary results and proofs are presented in Appendices A and B.

2. Preliminaries and notation. The field of real numbers is denoted by $\mathbb{R}$ and $\mathbb{N}$ stands for the set natural numbers (including zero). The subset $\{1, \ldots, k\}$ of $\mathbb{N}$ is denoted by $\lfloor k \rceil$. Given a set W, then card W stands for the cardinality of W. We adopt the standard notations of differential geometry and exterior algebra in the finite and infinite dimensional case [55, 57]. Let us briefly recall the main definitions of the infinite dimension setting introduced in control systems theory [17, 40, 19]. This approach is mainly based on the differential geometry of jets and prolongations (see, for instance, [27, 57]), whereas the approach of [24] and [34] is based on finite dimensional differential geometry [55].

Let A be a countable set. Denote by $\mathbb{R}^A$ the set of functions from A to $\mathbb{R}$. One may define the coordinate function $x_i : \mathbb{R}^A \to \mathbb{R}$ by $x_i(\xi) = \xi(i), i \in A$. This set can be endowed with the Fréchet topology (i.e., an *inverse limit* topology [57]). A basis of this topology is given by the subsets of the form $\mathcal{B} = \{\xi \in \mathbb{R}^A \mid |x_i(\xi) - \delta_i| < \epsilon_i, i \in F\}$, where F is a finite subset of A, $\delta_i \in \mathbb{R}$, and ϵ_i is a positive real number for $i \in F$. A function $\phi : \mathbb{R}^A \to \mathbb{R}$ is smooth if $\phi = \psi(x_{i_1}, \ldots, x_{i_s})$, where $\psi : \mathbb{R}^s \to \mathbb{R}$ is a smooth function. Only the dependence on a finite number of coordinates is allowed.

From this notion of smoothness, one can easily state the notions of vector fields and differential forms[6] on $\mathbb{R}^A$ and smooth mappings from $\mathbb{R}^A$ to $\mathbb{R}^B$. The notion of an $\mathbb{R}^A$-manifold can be also established easily as in the finitely dimensional case [57].

[5] See Theorem 7.2 for a precise statement of this sufficient condition of flatness.

[6] We stress that the forms are finite combinations of the form $\sum_i a_{I_i} dx_{I_i}$, where I_i is the multi-index $(j_{i,1}, \ldots, j_{i,r_i})$, the a_{I_i} are smooth functions, $dx_{I_i} = dx_{j_{i,1}} \wedge \cdots \wedge dx_{j_{i,r_i}}$. On the other hand, the fields are (possibly) infinite sums of the form $\sum_{i \in A} a_i \frac{\partial}{\partial x_i}$.

Given an $\mathbb{R}^A$-manifold $\mathcal{P}$, $C^\infty(\mathcal{P})$ denotes the set of smooth maps from $\mathcal{P}$ to $\mathbb{R}$. Let $\mathcal{Q}$ be an $\mathbb{R}^B$-manifold and let $\phi : \mathcal{P} \to \mathcal{Q}$ be a smooth mapping. The corresponding tangent and cotangent mapping will be denoted, respectively, by $\phi_* : T_p\mathcal{P} \to T_{\phi(p)}\mathcal{Q}$ and $\phi^* : T^*_{\phi(p)}\mathcal{Q} \to T^*_p\mathcal{P}$.

The map $\phi : \mathcal{P} \to \mathcal{Q}$ is called an *immersion* if, around every $\xi \in \mathcal{P}$ and $\phi(\xi) \in \mathcal{Q}$, there exist local charts of $\mathcal{P}$ and $\mathcal{Q}$ such that, in these coordinates, $\phi(x) = (x, 0)$. The map ϕ is called a *submersion* if, around every $\xi \in \mathcal{P}$ and $\phi(\xi) \in \mathcal{Q}$, there exist local charts of $\mathcal{P}$ and $\mathcal{Q}$ such that, in these coordinates, $\phi(x, y) = x$.

In the finite dimensional case, immersion and submersions are locally characterized, respectively, by the injectivity and surjectivity of the tangent mappings. However, in the infinite dimensional case this is no longer true. Moreover, the inverse function theorem and the classical Frobenius theorem (for distributions) do not hold and a field does not admit a flow in general [57].

Given two forms η and ξ in $\Lambda(\mathcal{P})$, then $\eta \wedge \xi$ denotes their *wedge* multiplication. The *exterior derivative* of $\eta \in \Lambda(\mathcal{P})$ will be denoted by $d\eta$. Note that the graded algebra $\Lambda(\mathcal{P})$, as well as its homogeneous elements $\Lambda_k(\mathcal{P})$ of degree k, have a structure of $C^\infty(\mathcal{P})$-module. See [55, 3] for details. Given a family $\nu = (\nu_1, \ldots, \nu_k)$ of a $C^\infty(\mathcal{P})$-module, then span $\{\nu_1, \ldots, \nu_k\}$ stands for the span over $C^\infty(\mathcal{P})$.

Given a field f and a 1-form ω on $\mathcal{P}$, we denote $\omega(f)$ by $\langle \omega, f \rangle$. The set of smooth k-forms on $\mathcal{P}$ will be denoted by $\Lambda_k(\mathcal{P})$ and $\Lambda(\mathcal{P}) = \cup_{k\in\mathbb{N}}\Lambda_k(\mathcal{P})$.

The following useful result of finite dimensional differential geometry is known as the "Cartan lemma" [55, Ex. 16, p. 80]. Let $\{\omega_1, \ldots, \omega_r\} \subset \Lambda_1(\mathcal{P})$ be independent pointwise. Assume that there exist 1-forms $\eta_1, \ldots, \eta_r$ such that $\sum_{i=1}^r \eta_i \wedge \omega_i = 0$. Then there exist functions $a_{ij} \in C^\infty(\mathcal{P})$, with $a_{ij} = a_{ji}$, such that $\eta_i = \sum_{j=1}^r a_{ij}\omega_j \quad (i = 1, \ldots, r)$. The same result is also valid pointwise, i.e., $\sum_{i=1}^r \eta_i \wedge \omega_i|_p = 0$ implies that $\eta_i(p) = \sum_{j=1}^r a_{ij}\omega_j(p) \quad (i = 1, \ldots, r)$ for convenient $a_{ij} = a_{ji} \in \mathbb{R}$.

A smooth codistribution J is a $C^\infty(\mathcal{P})$-submodule $J \subset T^*\mathcal{P}$. Given a submodule S of $\Lambda(\mathcal{P})$ and $p \in \mathcal{P}$, then $S(p)$ denotes the $\mathbb{R}$-linear subspace of $\Lambda_1(\mathcal{P})|_p$ given by $\operatorname{span}_\mathbb{R} \{\zeta(p) |\ \zeta \in S\}$. In particular, if J is a codistribution, then $J(p)$ denotes the subspace of $T^*_p\mathcal{P}$ given by $\operatorname{span}_\mathbb{R} \{\omega(p) |\ \omega \in J\}$.[7]

Assume that a codistribution I is locally generated by $\eta_1, \ldots, \eta_k$ and that $\Psi = \{x_i |\ i \in A\}$ is a local coordinate system around some open set $U \subset \mathcal{P}$. Then one may apply to I the standard techniques of differential geometry, for instance, the Frobenius theorem, by "pulling back" the results that hold on the finite dimensional case (see [40] and [36, section 2]).

3. Diffieties and systems. In this section we recall the main concepts of the infinite dimensional geometric setting of [17, 40, 19]. We have chosen to present a simplified exposition. For a more complete and intrinsic presentation the reader may refer to the cited literature.

3.1. Diffieties. A *diffiety* M is an $\mathbb{R}^A$-manifold equipped with a distribution Δ of finite dimension r, called *Cartan distribution.* A section of the Cartan distribution is called a *Cartan field.* An *ordinary diffiety* is a diffiety for which $\dim \Delta = 1$ and a Cartan field ∂_M is distinguished and called *the Cartan field.* In this paper we will consider only ordinary diffieties, which will be called simply by *diffieties.*

A Lie–Bäcklund mapping $\phi : M \to N$ between diffieties is a smooth mapping that is compatible with the Cartan fields, i.e., $\phi_*\partial_M = \partial_N \circ \phi$. A *Lie–Bäcklund*

[7] One can also define a codistribution as a map $p \mapsto J(p)$, where $J(p)$ is a subspace of $T^*_p\mathcal{P}$.

immersion (resp., *submersion*) is a Lie–Bäcklund mapping that is an immersion (resp., submersion). A Lie–Bäcklund isomorphism between two diffieties is a diffeomorphism that is a Lie–Bäcklund mapping.

Context permitting, we will denote the Cartan field of an ordinary diffiety M simply by $\frac{d}{dt}$. Given a smooth object ϕ defined on M (a smooth function, field, or form), then $L_{\frac{d}{dt}}(\phi)$ will be denoted by $\dot{\phi}$ and $L^n_{\frac{d}{dt}}(\phi)$ by $\phi^{(n)}, n \in \mathbb{N}$. In particular, if ω is a 1-form given by $\omega = \sum_{\text{finite}} \alpha_i dx_i$, then $\dot{\omega} = \sum_{\text{finite}}(\dot{\alpha}_i dx_i + \alpha_i d\dot{x}_i)$.

3.2. Systems. The set of real numbers $\mathbb{R}$ has a trivial diffiety structure with the Cartan field defined by the operation of differentiation of smooth functions. A *system* is a triple $(S, \mathbb{R}, \tau)$ where S is a diffiety equipped with Cartan field $\frac{d}{dt}$, the mapping $\tau : S \to \mathbb{R}$ is a Lie–Bäcklund submersion, and $\frac{d}{dt}(\tau) = 1$. The function τ represents *time* that is chosen once and for all. Context permitting, the system $(S, \mathbb{R}, \tau)$ is denoted simply by S. A *Lie–Bäcklund mapping between two systems* $(S, \mathbb{R}, \tau)$ and $(S', \mathbb{R}, \tau')$ is a *time-respecting* Lie–Bäcklund mapping $\phi : S \to S'$, i.e., $\tau' = \tau \circ \phi$. The previous condition means that the notion of time of both systems coincide. This notion of system is *time-varying*, as will be explained below.

3.3. State representation. We present a simplified definition of state representation that introduces the state and the input and its derivatives as a local coordinate system (see [17, 19] for a more intrinsic presentation).

A local state representation of a system $(S, \mathbb{R}, \tau)$ is a local coordinate system $\psi = \{t, x, U\}$, where $x = \{x_i, i \in \lfloor n \rceil\}$, $U = \{u_j^{(k)} | \; j \in \lfloor m \rceil, k \in \mathbb{N}\}$, where $u_j^{(k)} = L^k_{\frac{d}{dt}} u_j$, and $\tau = t$. The set of functions $x = (x_1, \ldots, x_n)$ is called state and $u = (u_1, \ldots, u_m)$ is called input. In these coordinates the Cartan field is locally written by

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i} + \sum_{k \in \mathbb{N}} \sum_{j \in \lfloor m \rceil} u_j^{(k+1)} \frac{\partial}{\partial u_j^{(k)}}. \tag{3.1}$$

Note that f_i may depend on t, x, and a finite number of elements of U. In this sense, the state representation defined here is said to be generalized, since one accepts that f_i may depend on the derivatives of the input. If the functions f_i depend only on $\{t, x, u\}$ for $i \in \lfloor n \rceil$, then the state representation is said to be *classical*. A state representation of a system S is completely determined by the choice of the state x and the input u and will be denoted by (x, u). A state representation is said to be *analytic* if the f_i are all analytic.[8]

3.4. Output. An output y of a system S is a set $y = (y_1, \ldots, y_p)$ of smooth functions defined on S. If (x, u) is a state representation of S, then it is clear that

$$y_j = y_j(t, x, u, \ldots, u^{(\alpha_j)}), \; j \in \lfloor p \rceil. \tag{3.2}$$

If the y_j depend only on $\{t, x, u\}$ for $j \in \lfloor p \rceil$, then the output is said to be *classical* with respect to the state representation (x, u). A state representation (x, u) with output y is said to be *analytic* if the functions f_i and the y_j are all analytic with respect to its arguments x and $\{u^{(j)} \mid j \in \mathbb{N}\}$.

[8] This definition is coordinate dependent since only smooth atlases are considered on diffieties [57].

3.5. System associated to differential equations. Now assume that a control system is defined by a set of equations

$$\begin{aligned} \dot{t} &= 1, \\ \dot{x}_i &= f_i(t, x, u, \dots, u^{(\alpha_i)}),\ i \in \lfloor n \rceil, \\ y_j &= y_j(t, x, u, \dots, u^{(\beta_j)}),\ j \in \lfloor p \rceil. \end{aligned} \tag{3.3}$$

One can always associate to these equations a diffiety S of global coordinates $\psi = \{t, x, U\}$ and Cartan field given by (3.1).

3.6. Flatness. We present now a simple definition of flatness in terms of coordinates.[9] A system S equipped with Cartan field $\frac{d}{dt}$ and time function $t = \tau$ is locally flat around $\xi \in S$ if there exists a set of smooth functions $y = (y_1, \dots, y_m)$, called *flat output*, such that the set $\{t, y_i^{(j)} |\ i \in \lfloor m \rceil, j \in \mathbb{N}\}$ is a (local) coordinate system of S around $\xi \in S$, where $y_i^{(j)} = L^j_{\frac{d}{dt}} y_i$. Note that the Cartan field is locally given by

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \sum_{j \in \mathbb{N}} \sum_{i \in \lfloor m \rceil} y_i^{(j+1)} \frac{\partial}{\partial y_i^{(j)}}.$$

Let $\Psi : S \to T$ be a Lie–Bäcklund isomorphism between two systems. Then S is flat if and only if T is flat also. If $y = (y_1, \dots, y_m)$ is a flat output of T, then $\{y_1 \circ \Psi, \dots, y_m \circ \Psi\}$ is a flat output of S.

3.7. Endogenous feedback and coordinate changes. Since a local state representation (x, u) is by definition a local coordinate system, a new local state representation (z, v) induces a coordinate change from $\{t, x, (u^{(i)} : i \in \mathbb{N})\}$ to $\{t, z, (v^{(j)} : i \in \mathbb{N})\}$. The coordinate changes of this kind are called *endogenous feedbacks*.[10]

An example of endogenous feedback is static-state feedback. Two state representations (x, u) and (z, v) defined around $\xi \in S$ are said to be linked by (time-varying) static-state feedback if we locally have

$$\operatorname{span}\{dt, dx\} = \operatorname{span}\{dt, dz\}, \tag{3.4a}$$

$$\operatorname{span}\{dt, dx, du\} = \operatorname{span}\{dt, dz, dv\}. \tag{3.4b}$$

Let (x, u) be a classical state representation and let z and v be family of smooth functions such that card x = card z and card u = card v. Then it is easy to show that, if (3.4) locally holds, then (z, v) is a local state representation that is linked to (x, u) by static-state feedback [36, Prop. 3.2].

Another example of endogenous feedback is putting integrators in series with the first k inputs of the system (3.3). This procedure induces a local state representation (z, v) of the system S, where $z = (x_1, \dots, x_n, u_1, \dots, u_k)$ and $v = (\dot{u}_1, \dots, \dot{u}_k, u_{k+1}, \dots, u_m)$, called *dynamic extension* of the state.

4. Subsystems.

A (local) *subsystem* S_a of a given system S is a system S_a such that there exists a surjective[11] Lie–Bäcklund submersion $\pi : U \subset S \to S_a$, where U is an open subset of S. A (local) subsystem will be denoted by (S_a, π) or simply by S_a.

[9] For more intrinsic definitions and some variations, see [17, 19].

[10] See [17] for a definition of endogenous feedback that considers an equivalence relation between systems.

[11] Since submersions are open maps, one can always consider that $S_a = \pi(U)$ by restricting S_a to the image of π.

4.1. State equations adapted to subsystems. Assume that there exists a local classical state representation (x, u) of a system S of the form

$$\dot{x}_a = f_a(t, x_a, u_a), \tag{4.1a}$$

$$\dot{x}_b = f_b(t, x_a, x_b, u_a, u_b), \tag{4.1b}$$

where $x = (x_a, x_b)$ and $u = (u_a, u_b)$. Assume that (4.1a) represents the state equations of a subsystem S_a and $\pi : S \to S_a$ is such that $\pi(t, x, U) = (t, x_a, U_a)$, where U denotes the set $(u^{(j)}|\ j \in \mathbb{N})$ and U_a denotes the set $({u_a}^{(j)}|\ j \in \mathbb{N})$. A state representation of S the form (4.1a)–(4.1b) is said to be adapted to the subsystem S_a. In the end of this section we show that state equations adapted to a subsystem can be generically constructed (see Proposition 4.4).

4.2. Relative static-state feedback. We will consider now a special case of endogenous feedback that will be called by *relative static-state feedback*. Consider that $((x_a, x_b), (u_a, u_b))$ is a local state representation for system S such that the state equations are of the form (4.1a)–(4.1b). A relative state feedback is a new state representation $((x_a, z_b), (u_a, v_b))$ such that

$$\begin{aligned} z_b &= z_b(t, x_b,\, x_a, u_a, \dots, {u_a}^{(r)}),\\ v_b &= v_b(t, x_b, u_b, x_a, u_a, \dots, {u_a}^{(r+1)}), \end{aligned} \tag{4.2}$$

where r is a convenient integer and similar equations do exist for x_b, u_b as functions of $x_a, z_b,\, u_a, v_b$ and the derivatives of u_a. In other words, this is an invertible time-varying feedback. The next definition renders this notion more intrinsic.

DEFINITION 4.1. *Let S be a system and let (π, S_a) be a (local) subsystem of S. Let (x, u) and (z, v) be two (local) state representations of S. Let Σ be the codistribution defined by the pull-back*[12] $\Sigma = \pi^*(T^* S_a)$*. Then (x, u) and (z, v) are linked by a* relative static-state feedback *with respect to the subsystem S_a if* $\operatorname{span}\{dx\} + \Sigma = \operatorname{span}\{dz\} + \Sigma$ *and* $\operatorname{span}\{dx, du\} + \Sigma = \operatorname{span}\{dz, dv\} + \Sigma$.

PROPOSITION 4.2. *Let S be a system with local state representation (x, u) defined on $V_\xi \subset S$, where $x = (x_a, x_b)$, and $u = (u_a, u_b)$ are such that the state equations are of the form* (4.1a)*–*(4.1b)*. Let S_a be the (local) subsystem associated to equation* (4.1a)*. Consider the set of smooth functions $z = (x_a, z_b)$ and $v = (u_a, v_b)$ defined on V_ξ, where* $\operatorname{card} x = \operatorname{card} z = n$ *and* $\operatorname{card} u = \operatorname{card} v = m$*. Then the following statements are equivalent:*

(i) *(z, v) is a local state representation around ξ and (x, u) and (z, v) are linked by relative static-state feedback.*

(ii) $\operatorname{span}\{dx\} + \Sigma = \operatorname{span}\{dz\} + \Sigma$ *and* $\operatorname{span}\{dx, du\} + \Sigma = \operatorname{span}\{dz, dv\} + \Sigma$.

Proof. See [39]. □

Remark 4.1. The proof of the Proposition 4.2 shows that (i) implies that condition (4.2) is satisfied for a subsystem S_a defined by (4.1a). It will be shown (see Proposition 4.4) that all subsystems admit adapted state equations of the form (4.1a)–(4.1b), up to relative static-state feedbacks.

4.3. Output subsystem. Given a system S with output y, a (local) *output subsystem* is a (local) subsystem $\pi : U \subset S \to Y$ such that $\pi^*(T^*_{\pi(\xi)} Y) = \operatorname{span}\left\{dt, dy^{(k)} : k \in \mathbb{N}\right\}|_\xi, \xi \in U$.

[12] Note that $\operatorname{span}\{dt\} \subset \Sigma$.

4.4. Existence of local output subsystems. Without loss of generality, assume that (x, u) is a classic state representation with output y. If it is not the case, we can add integrators in series with the input until the required properties are fulfilled. The next theorem shows that local output subsystems can be constructed generically and they admit adapted state equations up to relative static-state feedback. Furthermore, they are unique up to Lie–Bäcklund isomorphisms.

Theorem 4.3 (existence and uniqueness of output subsystems). *Let S be a system and let (x, u) be a classical* analytical *state representation defined on an open neighborhood $W \subset S$. Let y be a classical output of S. Let $n =$ card x. Let $U \subset W$ be the set of regular points of the codistributions $Y_k, \mathcal{Y}_k, k \in \lfloor n \rceil$, where $Y_k = \operatorname{span}\left\{dt, dy, \dots, dy^{(k)}\right\}$ and $\mathcal{Y}_k = \operatorname{span}\left\{dt, dx, dy, \dots, dy^{(k)}\right\}$. Then, around any $\xi \in U$, there exists an open neighborhood V_ξ of ξ and a local* classical *state representation $(z, v) = ((z_a, z_b), (v_a, v_b))$ of the system S, defined on V_ξ, such that:*

(i) *The (local) state equations are*

$$\dot{z}_a = f_a(t, z_a, v_a), \tag{4.3a}$$

$$\dot{z}_b = f_b(t, z_a, z_b, v_a, v_b). \tag{4.3b}$$

(ii) *Let Y be the local subsystem associated to* (4.3a) *and let $\pi : V_\xi \to Y$ be the corresponding Lie–Bäcklund submersion. We have $\pi^*(T^*Y) = \operatorname{span}$ $\{dt,\ dz_a,\ (dv_a^{(k)} : k \in \mathbb{N})\} = \operatorname{span}\left\{dt, dy^{(k)} : k \in \mathbb{N}\right\}$. In particular, Y is an output subsystem of S. Let $\mathcal{Z} = \{z_a, (v_a^{(k)} : k \in \mathbb{N})\}$ and $\mathcal{Y} = \{y_j^{(k)} : j \in \lfloor p \rceil, k \in \mathbb{N}\}$. Then $\mathcal{Z} \subset \mathcal{Y}$.*

(iii) *The state representations (x, u) and (z, v) are linked by relative static-state feedback with respect to the subsystem Y associated to* (4.3a).

Furthermore, two local output subsystems around any $\xi \in S$ are (locally) Lie–Bäcklund isomorphic.

Proof. See Appendix A.1. □

We state now the result that assures that a subsystem can be generically represented by state equations of the form (4.1a)–(4.1b).

Proposition 4.4. *Assume that S_a is a subsystem of S and that there exist local state representations for S_a and S around every point of S_a and S. Then, generically, there exists local state representations of S of the form* (4.1a)–(4.1b) *in a way that* (4.1a) *is a state representation of S_a.*

Proof. Let $\pi : S \to S_a$ be the corresponding Lie–Bäcklund submersion. Take a local state representation (z_a, e_a) of S_a around $\pi(\xi) \in S_a$. We abuse notation and denote $z_a \circ \pi$ and $e_a \circ \pi$, respectively, by z_a and e_a. Now, consider system S with output $y = (z_a, e_a)$ and construct, possibly by extending the state with derivatives of the input, a classical state representation of S such that y is a classical output. The result follows easily from the application of Theorem 4.3 and the fact that $T^*S_a = \operatorname{span}\{dt, dz_a^{(k)}, de_a^{(k)}, k \in \mathbb{N}\}$. □

If the outputs are differentially independent, the next result shows that local output subsystems are generically flat.

Proposition 4.5. *Let U be the open and dense subset of Theorem* 4.3. *Assume that the (explicit) system* (1.1a) *with output $y = h(t, x, u)$ is right-invertible, i.e., the output rank ρ is equal to the number of output components.*[13] *Let $\pi : V_\xi \subset S \to Y$ be a local output subsystem with $V_\xi \subset U$. Then Y is (locally) flat with flat output y.*

Proof. See [39]. □

[13] See Appendix B for the definition of the output rank ρ.

5. Relative flatness.

5. Relative flatness. We now state the concept of *relative flatness.*

DEFINITION 5.1. *Let S be a system and (π_1, S_1) and (π_2, S_2) be two subsystems of S. The system S is said to be locally* decomposed by *S_1 and S_2 if, around $\xi \in S$, there exists local coordinates (t, x^1) for S_1, (t, x^2) for S_2, and (t, x^1, x^2) for*[14] *S such that $\pi_i(t, x^1, x^2) = (t, x^i), i = 1, 2$. A system S is said to be (locally)* relatively flat *with respect to a subsystem S_1 if there exists a flat subsystem S_2 such that S is (locally) decomposed by S_1 and S_2.*

The next proposition states a sufficient condition for flatness. It can be shown that it is not a necessary condition [39].

PROPOSITION 5.2. *Let S_1 be a (locally) flat subsystem of a system S. Assume that S is relatively flat with respect to S_1. Then S is (locally) flat.*

Proof. The union of flat outputs of S_1 and S_2 is a flat output of S. $\square$

Remark 5.1. In the differential algebraic approach of [14] (see also [21]) one can define a subsystem of a system K/k as a field extension L/k such that L is a subfield of K. Then a system K/k is relatively flat with respect to L if the system K/L is flat, considering L as the ground field (see [11] for a result similar to Proposition 5.2.). However, these algebraic notions are not suitable for our purposes because integrability conditions are not available in this algebraic context.

It can be shown that, if K/k is relatively flat with respect to L, then K/k can be decomposed into two independent subsystems L/k and F/k, where F/k is flat (see [56]). In this sense, the assumption that the system is decomposed into two independent subsystems in the definition of relative flatness is not restrictive with respect to the algebraic definition (see also [35] for similar facts that occur when L corresponds to the noncontrollable subsystem.)

The following proposition is a necessary and sufficient condition for completing a given output y into a flat output (see [42] for related results).

PROPOSITION 5.3. *Let S be a system and let S_1 be a flat subsystem of S. Let y be a (local) flat output for S_1. Then there exists a set z of smooth functions such that S is locally flat with flat output (y, z) if and only if S is relatively flat with respect to the subsystem S_1.*

Proof. The necessity is obvious. The sufficiency follows from the proof of Proposition 5.2. $\square$

6. Implicit systems regarded as Lie–Bäcklund immersions.

6. Implicit systems regarded as Lie–Bäcklund immersions. Let S be the nonconstrained system defined by (1.1a). We show that, under some regularity assumptions, (1.1a)–(1.1b) may be regarded as a system that is immersed in S. We construct a system Γ and a Lie–Bäcklund immersion $\iota : \Gamma \to S$ such that every integral curve $\sigma(t)$ of the Cartan field of S, respecting the constraints $y(t) \equiv 0$, is of the form $\sigma(t) = \iota \circ \gamma(t)$ for a suitable integral curve $\gamma(t)$ of the Cartan field of Γ.

Consider the explicit (nonconstrained) system S defined by (1.1a) with output $y = h(t, x, u)$, global coordinates $\{t, x, (u_i^{(j)} : i \in \lfloor m \rceil; j \in \mathbb{N})\}$, and Cartan field (3.1). Consider now the following assumptions.

A1. Existence and regularity assumption. *Let $\Gamma = \{\xi \in S \mid y^{(k)}(\xi) = 0$ for all $k \in \mathbb{N}\}$. Assume that $\Gamma \neq \emptyset$ and furthermore, $\Gamma \subset U$, where U is the open and dense subset of the system S such that the statement of Theorem* 4.3 *holds.*[15] *In other words, around every point $\xi \in \Gamma$, we can construct a local output subsystem.*

[14] We abuse notation and denote $x^i \circ \pi_i$ simply by x^i.

[15] Note that in this case the state representation (1.1a) is globally defined. According the proof of Theorem 4.3 we have that U is the open and dense set of regular points of the codistributions $Y_k = \text{span}\left\{dt, dy, \dots, dy^{(k)}\right\}$ and $\mathcal{Y}_k = \text{span}\left\{dt, dx, dy, \dots, dy^{(k)}\right\}$ for $k \in \{0, 1, \dots, n\}$.

A2. Time interval assumption. *For every $\xi \in \Gamma$ and every open neighborhood $U \subset S$ of ξ, there exists some $\epsilon \geq 0$ such that $\tau(\Gamma \cap U)$ contains an open interval $(\tau(\xi) - \epsilon, \tau(\xi) + \epsilon)$.*

Remark 6.1. Note that assumption A2 means that Γ "does exist" during an interval of time. If the system is time-invariant it is easy to verify that assumption A2 is not needed. Note also that the set Γ may be empty, and in this case the implicit system has no solution. For instance, let $y_1 = x_1 + 1$ and $y_2 = x_1^2 + 2$. Then $y_1 = 0$ implies that $y_2 \neq 0$. A problem of this nature may occur with output derivatives.

When the assumptions A1, A2 hold, the set $\Gamma \subset S$ may be endowed with the structure of an immersed Fréchet manifold by choosing the subset topology, as shown by the following proposition.

PROPOSITION 6.1. *Suppose that assumptions* A1 *and* A2 *are satisfied for system S. Then the subset $\Gamma \subset S$ has a structure of immersed manifold in S. Let $\iota : \Gamma \to S$ be the canonical insertion. We can define a Cartan field ∂_Γ on Γ by the equation $\iota_* \partial_\Gamma(\gamma) = \frac{d}{dt} \circ \iota(\gamma), \gamma \in \Gamma$. Equipped with this Cartan field, Γ is a system such that ι is a Lie–Bäcklund immersion. Furthermore, all the solutions $\xi(t)$ of* (1.1a) *obeying the restriction* (1.1b) *are of the form $\xi(t) = \iota \circ \nu(t)$, where $\nu(t)$ is a solution of* Γ.

Proof. We show first that Γ is an immersed manifold. For this, consider the topological subspace $\Gamma \subset S$ with the subset topology. For each point $\xi \in \Gamma$, Theorem 4.3 gives local charts $\phi : \hat{U} \to \tilde{U} \subset \mathbb{R}^A$, where $\phi = \{t, z_a, V_a, z_b, V_b\}$, $V_a = \{v_a^{(k)} : k \in \mathbb{N}\}$, $V_b = \{v_b^{(k)} : k \in \mathbb{N}\}$, and we have span $\{dt, dz_a, dV_a\}$ = span $\{dt, dy^{(k)} : k \in \mathbb{N}\}$. This local chart is adapted to a local output subsystem $\pi : \hat{U} \to Y$, and is such that $\pi(t, z_a, V_a, z_b, V_b) = (t, z_a, V_a)$. Furthermore, by part (ii) of Theorem 4.3, the functions of the set $\mathcal{Z} = \{z_a, V_a\}$ are such that $\mathcal{Z} \subset \mathcal{Y}$, where $\mathcal{Y} = \{y^{(k)} : k \in \mathbb{N}\}$. By construction, if $\nu \in \hat{U} \cap \Gamma$, then $y^{(k)}(\nu) = 0$ for all $k \in \mathbb{N}$. This implies that all the components of $\mathcal{Z}$ are also null in ν. If we show that the functions in $\mathcal{W} = \mathcal{Y} - \mathcal{Z}$ are also null in $\nu \in \Gamma \cap \hat{U}$, we will show that a point ν is in $\Gamma \cap \hat{U}$ if and only if $z_a = 0$ and $V_a = 0$ in ν. In fact, note first that, since span $\{dt, d\mathcal{Z}\}$ = span $\{dt, d\mathcal{Y}\}$, all the functions θ in $\mathcal{Y}$ can be locally written in the form $\theta = \theta(t, z_a, V_a)$. By assumption A2, if we restrict $\tilde{U}$ to a basic open set of the form $I_{\tau(\xi)} \times W$, where $I_{\tau(\xi)} = (\tau(\xi) - \epsilon, \tau(\xi) + \epsilon)$, we may assume that, for every $\bar{t} \in I_{\tau(\xi)}$, then $\hat{U} \cap \Gamma$ contains a point $\xi_{\bar{t}} = (\bar{t}, z_a, V_a, z_b, V_b) = (\bar{t}, 0, 0, z_b, V_b)$. For any fixed $\bar{t} \in I_{\tau(\xi)}$, since $\xi_{\bar{t}} \in \Gamma \cap \hat{U}$, we have that $\theta(\xi_{\bar{t}}) = \theta(\bar{t}, 0, 0) = 0$. Since this is true for all $\bar{t} \in I_{\tau(\xi)}$, we have shown our claim.

Now consider the map $\mu : \Gamma \cap \hat{U} \to \mu(\Gamma \cap \hat{U}) \subset \mathbb{R}^B$ such that $\mu(t, 0, 0, z_b, V_b) = (t, z_b, V_b)$. We shall show that these maps form a smooth atlas of Γ. By construction it is clear that these maps are homeomorphisms. Hence it suffices to show that these charts are C^∞ compatible. For convenience denote the functions of the chart ϕ by $\{t, X, Z\}$ and the functions of the chart μ by $\{t, Z\}$, where $X = \{z_a, V_a\}$ and $Z = \{z_b, V_b\}$.

Now let $\mu_i : \Gamma \cap U_i \to \tilde{V}_i$, $i = 1, 2$, be two local charts constructed in that way, based, respectively, on the local charts of S given by $\phi_i = \{t, X_i, Z_i\}$, $i = 1, 2$. In particular, it follows that $\mu_i \circ \phi_i(t, 0, Z_i) = (t, Z_i)$, $i = 1, 2$. Without loss of generality, assume that $U_1 = U_2$. Consider the local coordinate change $(t, X_1, Z_1) = \phi_1 \circ \phi_2^{-1}(t, X_2, Z_2)$. Note that the map $\theta : \tilde{V}_2 \to \tilde{V}_1$ such that $(t, Z_1) \mapsto (t, Z_2)$ defined by $(t, 0, Z_1) = \phi_1 \circ \phi_2^{-1}(t, 0, Z_2)$ is a local diffeomorphism with inverse defined by $(t, 0, Z_2) = \phi_2 \circ \phi_1^{-1}(t, 0, Z_1)$. Since $\theta = \mu_1 \circ \mu_2^{-1}$, we conclude that such charts are C^∞ compatible.

Now let $\iota : \Gamma \to S$ be the insertion map. In the coordinates ϕ and μ previously constructed, we have $\iota(t, Z) = (t, 0, Z)$. In particular, ι is an immersion between $\mathbb{R}^A$-manifolds and so $\iota_*(\zeta)$ is injective for all $\zeta \in \Gamma$. Remember that any function η of the set $X = \{z_a, V_a\} \subset \mathcal{Y}$ is such that $\dot{\eta}|_\nu = 0$ for every $\nu \in \Gamma \cap \hat{U}$. In particular, we have that the image of $\iota_*(\nu)$ contains $\frac{d}{dt}(\iota(\nu))$ for every $\nu \in \Gamma \cap \hat{U}$. So we can define ∂_Γ by the rule $\iota_*\partial_\Gamma = \frac{d}{dt} \circ \iota$. By definition, it follows that ι is a Lie–Bäcklund immersion.

The last affirmation of the statement is a consequence of the first one. □

Remark 6.2. Let $\phi = (t, x_a, V_a, x_b, V_b)$ and $\mu = (t, x_b, V_b)$ be, respectively, the coordinates of S and Γ constructed above. In this coordinates we have

$$\partial_\Gamma = \frac{\partial}{\partial t} + \sum_{i=1}^{n_b} f_{b_i}(t, 0, 0, x_b, V_b) \frac{\partial}{\partial x_{b_i}} + \sum_{i=1}^{m_b} \sum_{j \in \mathbb{N}} u_{b_i}^{(j+1)} \frac{\partial}{\partial u_{b_i}^{(j)}}, \tag{6.1}$$

where $f_{b_i} = \frac{d}{dt}(x_{b_i}) = f_{b_i}(t, x_a, V_a, x_b, V_b)$, $i \in \lfloor n_b \rceil$. In other words, (x_b, u_b) is a state representation of Γ.

It is easy to show that the pull-back (by ι) of a relative static-state feedback for S with respect to a local output subsystem Y induces a static-state feedback for Γ if one considers the state representation $((x_a, x_b), (u_a, u_b))$ for S and (x_b, u_b) for Γ.

7. Flatness of implicit systems. In this section we will derive a sufficient condition for flatness of implicit systems. Let us begin with an auxiliary result.

Proposition 7.1. *Let Γ, S, and Y be systems, where Γ is immersed in S and Y is a subsystem of S. Let $\iota : \Gamma \to S$ and $\pi : S \to Y$ be, respectively, the corresponding Lie–Bäcklund immersion and submersion. Assume that there exist local coordinates (t, γ) of Γ, (t, γ, y) of S, and (t, y) of Y such that $\iota(t, \gamma) = (t, \gamma, 0)$ and*[16] *$\pi(t, \gamma, y) = (t, y)$. Assume that S is relatively flat with respect to Y. Then Γ is (locally) flat.*

Proof. Let S_2 be a flat subsystem of S such that S_2 and Y decomposes S (see Definition 5.1). Let $\pi_2 : S \to S_2$ be the corresponding Lie–Bäcklund submersion. Recall that there exists coordinates $(t, z, \tilde{y})$ of S, $(t, \tilde{y})$ of Y, and (t, z) of S_2 such that $\pi_2 : (t, z, \tilde{y}) = (t, z)$ and $\pi : (t, z, \tilde{y}) = (t, \tilde{y})$. Since the coordinate change map $(t, y) \to (t, \tilde{y})$ is a local diffeomorphism, we may assume without loss of generality that $\tilde{y} = y$. With a possible restriction of domains, we can consider the coordinate change mapping $\phi(t, \gamma, y) = (t, z, y)$. Note that the map $\phi_0(t, \gamma) = (t, z)$ such that $\phi(t, \gamma, 0) = (\phi_0(t, \gamma), 0) = (t, z, 0)$ is a local diffeomorphism. Let $\Psi : \Gamma \to S_2$ be such that $\Psi = \pi_2 \circ \iota$. By definition, Ψ is a Lie–Bäcklund mapping since it is a composition of Lie–Bäcklund mappings. In the coordinates (t, z) for S_2 and (t, γ) for Γ we have $\Psi(t, \gamma) = \phi_0(t, \gamma)$. Hence Ψ is a local Lie–Bäcklund isomorphism and so Γ is flat. In particular if θ is a flat output of S_2, then $\theta \circ \Psi$ is a flat output of Γ. □

The following result is a sufficient condition for flatness of an implicit system. It can be shown that it is not a necessary condition [39].

Theorem 7.2. *Let S be the explicit system defined by* (1.1a). *Let $y = h(t, x, u)$ be an output for system S and let Y be the corresponding output subsystem of S. Suppose that assumptions* A1–A2 *of the previous section hold for the system* (1.1a) *with the constraints* (1.1b). *According to Proposition* 6.1, (1.1a)–(1.1b) *define a system Γ that is immersed in S. Assume that the explicit system* (1.1a) *is (locally) relatively flat with respect to the subsystem Y. Then the implicit system Γ is locally flat around all $\xi \in \Gamma$.*

Proof. Let $\iota : \Gamma \to S$ be the insertion map and let $\pi : U \subset S \to Y$ be the canonical submersion onto the local output subsystem Y. According to the proof of Proposition

[16] We assume that (t, y) is inside the domain of our local chart of Y for $y = 0$.

6.1, we can define local charts $\phi = (t, X, Z)$ of S, $\mu = (t, Z)$ of Γ and $\Psi = (t, X)$ of Y such that $\iota(t, X) = (t, 0, Z)$ and $\pi(t, X, Z) = (t, X)$. Hence, by Proposition 7.1 (for $\gamma = Z$ and $y = X$) the result follows. $\square$

Let (1.1a) be a flat (explicit) system and assume that the output y of (1.1b) is part of the flat output of the explicit system (1.1a). Then next result shows that the implicit system (1.1a)–(1.1b) is flat.

COROLLARY 7.3. *Assume that S is locally flat with flat output $y = (y_1, \ldots, y_m)$. Assume that the local coordinate system $\{t, y_i^{(j)} : i \in \lfloor m \rceil, j \in \mathbb{N}\}$ is defined on open set V whose image is a basic open set $\tilde{V}$.*[17] *Let $\Gamma \subset V$ defined by $\{\xi \in V \mid y_i^{(j)}(\xi) = 0, i \in \lfloor r \rceil, j \in \mathbb{N}\}$. Assume that Γ is nonempty. Then Γ is an immersed system in $V \subset S$. Furthermore, Γ is (locally) flat with flat output $y_{r+1}, \ldots, y_m$.*

Proof. Consider system S with output $\tilde{y} = (y_{r+1}, \ldots, y_m)$. Let $\tilde{x} = \emptyset$ and $\tilde{u} = (\tilde{y}_1, \ldots, \tilde{y}_m)$. Then $(\tilde{x}, \tilde{u})$ is a local state representation of S. Let $\tilde{Y}_r = \text{span}\ \{dt,\ dy,\ \ldots,\ dy^{(k)}\}$ and $\mathcal{Y}_r = \text{span}\ \{dt,\ d\tilde{x},\ dy,\ \ldots,\ dy^{(k)}\}$. Then $\tilde{Y}_r = \mathcal{Y}_r$ are nonsingular codistributions on S for $r \in \mathbb{N}$ and hence assumption A1 of section 6 holds. Since $\tilde{V}$ is a basic open set, it is also clear that assumption A2 holds. By Theorem 4.3, the output subsystem $\tilde{Y}$ is well defined, and by Proposition 4.5, it follows that $\tilde{Y}$ is locally flat. By Proposition 5.3, S is relatively flat with respect to $\tilde{Y}$. The desired result follows from Theorem 7.2. $\square$

8. A sufficient condition for relative flatness.

Consider a system S and a subsystem S_1 of S given by (4.1a)–(4.1b), where (4.1a) represents S_1. Let $\dim x_a = n_a$, $\dim x_b = n_b$, $\dim u_a = m_a$, and $\dim u_b = m_b$. For this system one can define the *relative derived flag* as follows.

DEFINITION 8.1. *The relative derived flag of the system* (4.1a)–(4.1b) *is the sequence of codistributions $I^{(k)}$ defined by $I^{(-1)} = \text{span}\{(dx_b - f_b dt), (du_b - \dot{u}_b dt)\}$, and $I^{(k)}(p) = \text{span}\{\omega(p) \mid \ \omega \in I^{(k-1)},\ d\omega(p) \bmod\ (I^{(k-1)} + J)|_p \equiv 0\}$, $k \in \mathbb{N}$, where*

$$J \;=\; \text{span}\left\{(dx_a - \dot{x}_a dt), (du_a{}^{(j)} - u_a{}^{(j+1)}dt)|\ \ j \in \mathbb{N}\right\}. \tag{8.1}$$

Remark 8.1. In the proof of Proposition 8.3 it is shown that if $I^{(k)}$ is nonsingular, then it is smooth (otherwise $I^{(k+1)}$ is not well defined). The 1-forms in span $\left\{\frac{d}{dt}\right\}^{\perp}$ are called *contact forms* [40]. Let $\pi : S \to S_a$ be the Lie–Bäcklund submersion of S onto subsystem S_a (see section 4.1). Then it is easy to show that J is the codistribution generated by the contact forms of S_a, i.e., $J = \pi^*(T^*S_a) \cap \text{span}\left\{\frac{d}{dt}\right\}^{\perp}$. It follows that J is invariant by coordinate changes, and in particular, it is invariant by endogenous feedback. In [39] it is shown that

$$I^{(0)} = \text{span}\left\{dx_b - \dot{x}_b dt\right\}. \tag{8.2}$$

By construction we have $\dim I^{(-1)} = n_b + m_b$ and $\dim I^{(0)} = n_b$. Note also that $I^{(k)} + J \subset I^{(-1)} + J \subset \text{span}\left\{\frac{d}{dt}\right\}^{\perp}, k \in \mathbb{N}$. We will show that the relative derived flag carries an intrinsic structural information, at least if one restricts the class of transformations to *relative static-state feedback* (see Corollary 8.4).

THEOREM 8.2. *Assume that the codistributions* span $\left\{I^{(k)}, dt, J\right\}$ *are involutive, that $I^{(k)}$ are nonsingular for all $k \in \mathbb{N}$, and that $I^{(N)} = 0$ for N big enough. Then the system S is (locally) relatively flat with respect to S_1.*

[17] Recall that a basic open set is of the form $\tilde{V} = \{\xi \in S \mid |y_i^{(j)}(\xi) - \bar{y}_i^{(j)}| < \epsilon_{ij}, (i,j) \in \Delta\}$, where Δ is a finite subset of $\lfloor m \rceil \times \mathbb{N}$, $\bar{y}_i^{(j)} \in \mathbb{R}$, and $\epsilon_{ij} \in \mathbb{R}^+$.

Remark 8.2. It is easy to verify that J is involutive, i.e., that $d\omega \bmod J \equiv 0$ for all 1-forms $\omega \in J$. Furthermore, the codistribution span $\left\{I^{(k)}, dt, J\right\}$ is involutive if and only if span $\left\{I^{(k)}, dt, J_{\rho_k}\right\}$ is involutive for ρ_k big enough, where

$$J_l = \operatorname{span}\left\{(dx_a - f_a dt), (d{u_a}^{(j)} - {u_a}^{(j+1)}dt) \mid j \in \lfloor l \rceil\right\}. \tag{8.3}$$

To prove Theorem 8.2 we need the following auxiliary result whose proof is deferred to Appendix A.2.

PROPOSITION 8.3. *Assume that the conditions of the Theorem* 8.2 *are satisfied on an open neighborhood V_ξ of ξ in S. Then, for every $p \in V_\xi$ and $k \in \mathbb{N}$ we have* $\dim(I^{(k)} + J)|_p / J(p) = \dim I^{(k)}(p)$. *Assume that $I^{(k-1)} + J$ has a local basis $B = \bar{B} \cup B_J$, where B_J is a local basis of J and $\bar{B}$ is of the form*

$$\bar{B} = \left\{\omega_i^{(j)} : i \in \lfloor s \rceil, j \in \{0, \ldots, r_i\}\right\}, \tag{8.4}$$

where $\omega_i = d\theta_i - \dot{\theta}_i dt$, $\theta_i \in C^\infty(S)$, $i \in \lfloor s \rceil$ (or $\bar{B} = \emptyset$). Assume that the subset $\{\omega_i^{(r_i)} : i \in \lfloor s \rceil\}$ is linearly independent mod $\{I^{(k)} + J\}$. *Let $\dot{B} = \{\omega_i^{(r_i+1)} : i \in \lfloor s \rceil\}$. Then we may complete the set $B \cup \dot{B}$ with a set $\hat{B} = \{\omega_i, i = s+1, \ldots, \sigma\}$, where $\omega_i = d\theta_i - \dot{\theta}_i dt$ in a way that $B \cup \dot{B} \cup \hat{B}$ is a basis of $I^{(k-2)} + J$ such that $\dot{B} \cup \hat{B}$ is linearly independent* mod $\{I^{(k-1)} + J\}$.

Proof (of Theorem 8.2*).*[18] Let $N \in \mathbb{N}$ be the smallest integer such that $I^{(k)} = I^{(k+1)} = 0$ for all $k \geq N$. Let $\mathcal{B}_N$ be a basis for $J = I^{(N)} + J$ given by $\mathcal{B}_N = \{\eta, \mu_l \mid\ l \in \mathbb{N}\}$, where $\eta = (dx_a - \dot{x}_a dt)$ and $\mu_l = (d{u_a}^{(l)} - {u_a}^{(l+1)}dt), l \in \mathbb{N}$. Since span $\left\{I^{(N-1)}, J, dt\right\}$ is involutive and $I^{(N-1)}$ is nonsingular, by Proposition 8.3 with $\bar{B} = \emptyset$, we can construct a local basis $\mathcal{B}_{N-1}$ of $I^{(N-1)} + J$ of the form $\mathcal{B}_{N-1} = A_{N-1} \cup \mathcal{B}_N$, where $A_{N-1} = \{(d\theta_{1,i_1} - \frac{d}{dt}\theta_{1,i_1}dt) \mid i_1 \in \lfloor s_{N-1} \rceil\}$. Let $\dot{A}_{N-1} = \{(d\theta_{1,i_1}^{(1)} - \theta_{1,i_1}^{(2)}dt), i_1 \in \lfloor s_{N-1} \rceil\}$. By Proposition 8.3, we may construct a set $\hat{A}_{N-1} = \{(d\theta_{2,i_2} - \frac{d}{dt}\theta_{2,i_2}dt) \mid i_2 \in \lfloor s_{N-2} \rceil\}$ in a way that $\mathcal{B}_{N-2} = A_{N-2} \cup \mathcal{B}_N$ is a basis of $I^{(N-2)} + J$, where $A_{N-2} = \hat{A}_{N-1} \cup \dot{A}_{N-1} \cup A_{N-1} = \{(d\theta_{k,i_k}^{(j-1)} - \theta_{k,i_k}^{(j)}dt) \mid k \in \lfloor 2 \rceil, i_k \in \lfloor s_{N-k} \rceil, j \in \lfloor 2-k+1 \rceil\}$. Note also that, by Proposition 8.3, it follows that the set $\hat{A}_{N-1} \cup \dot{A}_{N-1} = \{(d\theta_{k,i_k}^{(2-k)} - \theta_{k,i_k}^{(2-k+1)}dt) \mid k \in \lfloor 2 \rceil, i_k \in \lfloor s_{N-k} \rceil\}$ is linearly independent mod $I^{(N-1)} + J$.

Continuing in this way, using Proposition 8.3, we may construct in the rth step, a basis for $I^{(N-r)} + J$ of the form

$$\mathcal{B}_{N-r} = A_{N-r} \cup \mathcal{B}_N, \tag{8.5}$$

where $A_{N-r} = \hat{A}_{N-r+1} \cup \dot{A}_{N-r+1} \cup A_{N-r+1}$ and

$$\begin{aligned} A_{N-r+1} &= \{(d\theta_{k,i_k}^{(j-1)} - \theta_{k,i_k}^{(j)}dt) \mid k \in \lfloor r-1 \rceil, i_k \in \lfloor s_{N-k} \rceil, j \in \lfloor r-k \rceil\} \\ \hat{A}_{N-r+1} \cup \dot{A}_{N-r+1} &= \{(d\theta_{k,i_k}^{(r-k)} - \theta_{k,i_k}^{(r-k+1)}dt) \mid k \in \lfloor r \rceil, i_k \in \lfloor s_{N-k} \rceil\} \end{aligned} \tag{8.6}$$

and where $\hat{A}_{N-r+1} \cup \dot{A}_{N-r+1}$ is linearly independent mod $\{I^{(N-r+1)} + J\}$ for $r \in \lfloor N+1 \rceil$. From Proposition 8.3, note that $\dim(I^{(k)}(p) + J(p))/J(p) = \dim I^{(k)}(p)$, $k \in \mathbb{N}$.

[18] Most of the techniques that are necessary for the proof of our sufficient condition of relative flatness are very similar to the techniques of the proof of the main result of [35].

Taking $r = N+1$ in (8.6) we obtain a basis $\mathcal{B}_{-1} = A_{-1} \cup B_N$, where $A_{-1} = \hat{A}_0 \cup \dot{A}_0 \cup A_0$ and

$$\begin{aligned} A_0 &= \{(d\theta_{k,i_k}^{(j-1)} - \theta_{k,i_k}^{(j)} dt) \mid k \in \lfloor N \rceil, i_k \in \lfloor s_{N-k} \rceil, j \in \lfloor N-k+1 \rceil\}, \\ \hat{A}_0 \cup \dot{A}_0 &= \{(d\theta_{k,i_k}^{(N-k+1)} - \theta_{k,i_k}^{(N-k+2)} dt) \mid k \in \lfloor N+1 \rceil, i_k \in \lfloor s_{N-k} \rceil\}, \end{aligned} \tag{8.7}$$

where the set $\hat{A}_0 \cup \dot{A}_0$ is independent mod $I^{(0)} + J$. Since $\dim I^{(0)} = n_b$ and $\dim I^{(-1)} = n_b + m_b$ we have card $\hat{A}_0 \cup \dot{A}_0 = m_b$. Now define the set w of n_b (state) functions and the set v of m_b (input) functions given by

$$\begin{aligned} w &= \{w_{k,i_k}^l \mid w_{k,i_k}^l = \theta_{k,i_k}^{(l-1)} : k \in \lfloor N \rceil, i_k \in \lfloor s_{N-k} \rceil, l \in \lfloor N-k+1 \rceil\}, \\ v &= \{v_{k,i_k} \mid v_{k,i_k} = \theta_{k,i_k}^{(N-k+1)} : k \in \lfloor N+1 \rceil, i_k \in \lfloor s_{N-k} \rceil\}. \end{aligned}$$

By construction of $\mathcal{B}_0$ and $\mathcal{B}_{-1}$ it is clear that $I^{(0)} + J + \operatorname{span}\{dt\} = \operatorname{span}\{dt, dx_a, dw\} + J = \operatorname{span}\{dt, dx_a, dx_b\} + J$ and $I^{(-1)} + J + \operatorname{span}\{dt\} = \operatorname{span}\{dt, dx_a, dw, du_a, dv\} = \operatorname{span}\{dt, dx_a, dx_b, du_a, du_b\} + J$. Since card $x_b =$ card w and card $v_b =$ card v then, by Proposition 4.2 we conclude that $((x_a, w), (u_a, v))$ is a state representation that is linked to $((x_a, x_b), (u_a, u_b))$ by relative static-state feedback. Since $I^{(k)} \subset \operatorname{span}\left\{\frac{d}{dt}\right\}^{\perp}$, the equations $\langle (d\theta_{k,i_k}^{(j)} - \theta_{k,i_k}^{(j+1)} dt), \frac{d}{dt} \rangle = 0$, $k \in \lfloor N \rceil$, $i_k \in \lfloor s_{N-k} \rceil$, $j \in \lfloor N-k+1 \rceil$, imply the following closed loop state equations:

$$\begin{aligned} &\dot{t} = 1, \\ &\dot{x}_a = f_a(x_a, u_a) \\ &\left\{ \begin{aligned} \dot{w}_{k,i_k}^1 &= w_{k,i_k}^2, \\ \dot{w}_{k,i_k}^2 &= w_{k,i_k}^3, \\ &\vdots \\ \dot{w}_{k,i_k}^{N-k+1} &= v_{k,i_k}, \end{aligned} \right. \qquad k \in \lfloor N \rceil,\ i_k \in \lfloor s_{N-k} \rceil. \end{aligned} \tag{8.8}$$ □

Remark 8.3. Note that, if $s_{-1} > 0$, then the inputs $\{v_{k,i_k} \mid k = N+1, i_k \in \lfloor s_{-1} \rceil\}$ are completely decoupled from the state of system (8.8), i.e., (8.8) is not well formed in this case [46]. Note also that, if one restricts the coordinate transformations to the class of *relative static-state feedback* (see Definition 4.1), then the conditions of Theorem 8.2 are necessary and sufficient. This follows from the invariance of the relative derived flag with respect to relative static-state feedback (see Corollary 8.4) and after (tedious) calculations of the relative derived flag of a system of the form (8.8).

COROLLARY 8.4. *Consider the system S of equations* (4.1a)–(4.1b). *Let $x = (x_a, x_b)$ and J be defined by* (8.1). *Let $\widehat{I}^{(-1)} = \operatorname{span}\{dx - \dot{x}dt, du - \dot{u}dt\} + J$ and $\widehat{I}^{(k)}(p) = \operatorname{span}\{\omega(p) \mid \omega \in \widehat{I}^{(k-1)},\ d\omega(p) \bmod \widehat{I}^{(k-1)}|_p \equiv 0\}$ for $k \in \mathbb{N}$. Assume that the codistributions $\operatorname{span}\{\widehat{I}^{(k)}, dt\}$ are involutive, $\dim \widehat{I}^{(k)}(q)/J(q)$ is (locally) constant for $k \in \mathbb{N}$, and that $\widehat{I}^{(N)} = J$ for N big enough. Then the system S is (locally) relatively flat with respect to S_1. Furthermore, the codistributions $\widehat{I}^{(k)}$, $k \in \mathbb{N}$, are invariant by relative static-state feedback with respect to the subsystem defined by* (4.1a).

Proof. We show first that $\widehat{I}^{(k)} = I^{(k)} + J$ for $k \in \{-1\} \cup \mathbb{N}$. This is obviously true for $k = -1$. Assume that this is true for $k-1$ and let $\hat{\omega} \in \widehat{I}^{(k-1)}$. Then $\hat{\omega} = \omega + \mu$, where $\omega \in I^{(k-1)}$ and $\mu \in J$. As J is involutive, then $d\hat{\omega} \bmod \widehat{I}^{(k-1)} \equiv 0$ if and only if $d\omega \bmod (I^{(k-1)} + J) \equiv 0$. In particular, $\hat\omega \in \widehat{I}^{(k)}$ if and only if $\omega \in I^{(k)}$. It follows that $\widehat{I}^{(k)} = I^{(k)} + J$, showing our claim. Hence, the first affirmation follows easily from Theorem 8.2. To show the invariance of the flag $\widehat{I}^{(k)}$, let $(\tilde x, \tilde u)$ be a state representation of S that is linked to (x, u) by relative static-state feedback. Let $\widetilde{I}^{(-1)} = \text{span}\,\{d\tilde x - \dot{\tilde x}dt,\, d\tilde u - \dot{\tilde u}dt\} + J$. Since $J + \text{span}\,\{dt\} = \Sigma = \pi^*(T^*S_a)$, by Definition 4.1 it follows $\widetilde{I}^{(-1)} + \text{span}\,\{dt\} = \widehat{I}^{(-1)} + \text{span}\,\{dt\}$. Hence, $\tilde\omega \in \widetilde{I}^{(-1)}$ if and only if $\tilde\omega = \hat\omega + \beta dt$, where $\hat\omega \in \widehat{I}^{(-1)}$. Now note that $\widetilde{I}^{(-1)}$ and $\widehat{I}^{(-1)}$ are both contained in span $\left\{\frac{d}{dt}\right\}^{\perp}$. In particular, $\langle\tilde\omega, \frac{d}{dt}\rangle = \langle\hat\omega, \frac{d}{dt}\rangle = 0$ implies that $\beta = 0$. We conclude that $\widetilde{I}^{(-1)} = \widehat{I}^{(-1)}$. Since the computation of $\widetilde{I}^{(k)}$ follows the same rule as the computation $\widehat{I}^{(k)}$ and J is invariant by endogenous feedback (see Remark 8.1), we conclude that $\widetilde{I}^{(k)} = \widehat{I}^{(k)}, k \in \mathbb{N}$. □

Remark 8.4. Let $\mathcal{U} = \text{span}\,\{dx\}^{\perp}$ and $H = J^{\perp}$. Let $G_0 = \mathcal{U} \cap H$ and let $G_{k+1} = G_k + [\frac{d}{dt}, G_k]$. It can be shown [9] that the conditions of Theorem 8.2 for time-invariant systems are equivalent to the involutivity of the distributions G_i and the existence of k such that $G_i = H$ for all $i \geq k$.

8.1. Flatness and local output subsystems.

8.1. Flatness and local output subsystems. Theorem 8.5 is a sufficient condition for relative flatness with respect to a local output subsystem.

THEOREM 8.5. *Let S be the explicit system* (1.1a) *with state representation* (x, u) *and output $y = h(t, x, u)$. Let U be the open and dense set where Theorem* 4.3 *holds. Let* $\widehat{I}^{(0)} = \text{span}\,\{dx - \dot x dt\} + J$, *where* $J = \text{span}\{dy^{(k-1)} - y^{(k)}dt : k \in \mathbb{N}\}$. *Consider the relative derived flag* $\widehat{I}^{(i)}(p) = \text{span}\{\omega(p) \mid \omega \in \widehat{I}^{(i-1)},\, d\omega(p) \bmod \widehat{I}^{(i-1)}(p) \equiv 0\}$. *Assume that, in U, the codistributions* $\text{span}\{\widehat{I}^{(k)}, dt\}$ *are involutive, and that* $\dim \widehat{I}^{(k)}(q)/J(q)$ *is (locally) constant dimensional for $k \in \mathbb{N}$ and $\widehat{I}^{(N)} = J$ for N big enough. Then S is (locally) relatively flat* with respect to *the output subsystem Y around every $\xi \in U$.*

Proof. By Theorem 4.3 there exists a local output subsystem Y of S and new state representation $((z_a, z_b), (v_a, v_b))$ linked to (x, u) by a relative static-state feedback, such that the closed loop state equations are given by (4.3a)–(4.3b), where $\text{span}\{dt, dz_a, dv_a^{(k)} : k \in \mathbb{N}\} = \text{span}\,\left\{dt, dy^{(k)} : k \in \mathbb{N}\right\} = J + \text{span}\,\{dt\}$. Let $\tilde J = \text{span}\{(dz_a - \dot z_a), (dv_a^{(k)} - v_a^{(k+1)}) : k \in \mathbb{N}\}$. It follows easily that $\tilde J + \text{span}\,\{dt\} = J + \text{span}\,\{dt\}$. Using the fact that $J \subset \text{span}\,\left\{\frac{d}{dt}\right\}^{\perp}$ and $\tilde J \subset \text{span}\,\left\{\frac{d}{dt}\right\}^{\perp}$ (see the arguments of the proof of Corollary 8.4), it follows that $\tilde J = J$. Then the result follows from Corollary 8.4. □

9. Examples.

9.1. An academic example.

9.1. An academic example. Consider the implicit system

$$\dot x_1 = \frac{x_2^2}{(1 + x_3^2)^2} + e^{x_3}u_1, \quad \dot x_2 = (1 + x_3^2)u_1 + \frac{2x_2x_3}{(1 + x_3^2)}u_2, \quad \dot x_3 = u_2, \tag{9.1a}$$

$$y = x_1 = 0. \tag{9.1b}$$

Let S be the (explicit) system (9.1a) with output $y = x_1$. It is easy to verify that the codistributions $\mathcal{Y}_k = \text{span}\,\left\{dt, dy, \ldots, dy^{(k)}\right\}$ and $Y_k = \text{span}\{dt, dx, dy, \ldots, dy^{(k)}\}$ of Lemma B.1 are nonsingular everywhere for $k \in \mathbb{N}$, and $\sigma_k = 1, k \geq 1$. Note also that $\Gamma = \{\xi \in S \mid y^{(k)}(\xi) = 0\}$ is nonempty because Γ contains the point $\xi \in S$ defined by $x_1(\xi) = x_2(\xi) = x_3(\xi) = u_1^{(k)}(\xi) = u_2^{(k)}(\xi) = 0$, $k \in \mathbb{N}$ (for any t). Since the system is time-invariant then the assumptions A1 and A2 of section 6 are satisfied.

By Proposition 6.1, the implicit system is an immersed system in the nonconstrained system. Let $J = \text{span}\{dy^{(k)} - y^{(k+1)}dt : k \in \mathbb{N}\}$ and let $\widehat{I}^{(0)} = \text{span}\,\{dx - \dot{x}dt\} + J$. Using condition (A.3), some calculations show that[19] $\widehat{I}^{(1)} = \text{span}\{\eta - \langle \eta, \frac{d}{dt}\rangle dt\} + J$, where $\eta = dx_2 - \frac{2x_2x_3}{(1+x_3^2)}dx_3$ and $\widehat{I}^{(2)} = J$. Since $d\eta = \frac{2x_3}{(1+x_3^2)}(\eta \wedge dz_3)$, it follows from Theorem 8.5, for every local output subsystem Y, that the explicit system S is relatively flat with respect to Y. By Theorem 7.2, the implicit system Γ defined by (9.1a)–(9.1b) is locally flat around every point $\xi \in \Gamma$. By the proof of Theorem 8.2 and the construction of Γ in section 6, a flat output of the implicit system can be constructed by finding a function ψ such that $d\phi = \alpha\eta$. A possible solution is $\psi = \frac{x_2}{(1+x_3^2)}$.

By Propositions 4.5 and 5.3, one may complete the output y into a flat output (y, z) for system S. In this case one may take $z = \psi$.

9.2. Constrained robots. Constrained robots are robots whose movement is restricted by some physical contact surfaces. Such restrictions can be represented by adding r holonomic constraints $\phi_i(q) = 0 \ \ (i = 1, \ldots, r)$ to its original equations.

The following model can be obtained by taking into account the contact forces [28]:

$$\text{(9.2a)} \qquad \frac{d}{dt}\begin{pmatrix} q \\ \dot{q} \end{pmatrix} = \begin{pmatrix} \dot{q} \\ -M^{-1}H \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ M^{-1}(J\phi)^T & M^{-1} \end{pmatrix}\begin{pmatrix} \lambda \\ \tau \end{pmatrix},$$

$$\text{(9.2b)} \qquad 0 = \phi_i(q),\ i = 1, \ldots, r,$$

where $q \in \mathbb{R}^n$, $J\phi(q) = \partial\phi/\partial q$, $\lambda = (\lambda_1, \ldots, \lambda_r)^T$ is a vector corresponding to the contact forces, $M(q)$ is the symmetric positive definite mass matrix, and $H(q, \dot{q})$ corresponds to Coriolis and gravity forces. We will assume that $\partial\phi/\partial q$ has rank r for all q in the operation region of the robot. Note that system (9.2a)–(9.2b) is in the form (1.1a)–(1.1b).

Let $\psi = (\psi_1, \ldots, \psi_{n-r})$ be chosen in a way that map $q \mapsto (\phi, \psi)$ is a local diffeomorphism. Considering only the explicit system S defined by (9.2a), it is easy to show that (q, λ) is a flat output for S. In particular, (ϕ, ψ, λ) is also a flat output for S. From Corollary 7.3, it follows that (ψ, λ) is a flat output for the constrained robot. Note now that ψ are local coordinates of the constraint surface. In particular, the simultaneous tracking of the position along the constraint surface and the contact forces are possible. The reader may refer to [38] for details and the presentation of the design of a flatness based control, including the underactuated case. Another approach for the solution of this problem is considered, for instance, in [28].

10. Conclusions. In this paper we show that the concept of relative flatness, introduced here, is directly related to the flatness of implicit systems. Sufficient conditions of relative flatness are provided (see Theorem 8.5). This result can be combined with Theorem 7.2 in order to study flatness of implicit systems (1.1a)–(1.1b), as illustrated in the example of section 9.1.

We show, under regularity assumptions, that an implicit system (1.1a)–(1.1b) defines a system Γ (in the sense of section 3.2) that admits state space representations and is immersed in the (explicit) system S defined by (1.1a) (Proposition 6.1). This immersion is in fact an embedding since the topology of the immersed system is the

[19] The application of part (ii) of Lemma A.2 is the easiest way for computing relative derived flags, and lead to linear equations with coefficients that are functions defined on S as shown in the proof of Lemma A.2.

subset topology. This result may be regarded as a generalization of the fact that equation $f(x) = 0$, where $f : \mathbb{R}^n \to \mathbb{R}^p$, defines implicitly an embedded submanifold of $\mathbb{R}^n$ when the Jacobian matrix $Jf(x)$ has constant rank in the solutions of $f(x) = 0$.

Although it is assumed that system (1.1a)–(1.1b) is analytic, this hypothesis is only needed to assure that the output rank ρ of the explicit system (1.1a) with output $y = h(x, u, t)$ is a global invariant, at least in the subset $U \subset S$ of nonsingular points of the codistributions Y_k and $\mathcal{Y}_k$ for $k = 0, \dots, n$ (see Lemma B.1). Note that the differential dimension[20] of the implicit system Γ defined by (1.1a)–(1.1b) is $\tilde{m} = m - \rho$, where $m = \operatorname{card} u$. Hence the assumption of analyticity implies that $\tilde{m}$ is an invariant. All the results of this paper could be rewritten in the smooth case (see [36, Lemma 6.2] for a smooth version of Lemma B.1), but in this case the differential dimension of Γ may depend on the working point. In the same way, it is easy to restrict our results to the time-invariant case (see [36, Lemma 8.1]).

All the definitions and results of this paper are local (note that the time-varying notions are also local in time). The only exception is the construction of the system Γ in section 6 (see Proposition 6.1), that is, a "global" construction.

Appendix A. Proof of auxiliary results.

A.1. Proof of Theorem 4.3. In this proof we use the results and the notations of Lemma B.1. Let $n = \dim x$. By that lemma, around $\xi \in U$, there exists a local state representation (x_n, u_n) defined in V_ξ such that

$$\operatorname{span}\{dt, dx_n\} = \operatorname{span}\left\{dt, dx, dy, \dots, dy^{(n)}\right\}, \tag{A.1a}$$

$$\operatorname{span}\{dt, dx_n, du_n\} = \operatorname{span}\left\{dt, dx, du, dy, \dots, dy^{(n+1)}\right\}, \tag{A.1b}$$

and where $u_n = (\bar{y}_n^{(n+1)}, \hat{u}_n)$. Now choose a subset z_a of $\{y, \dots, y^{(n)}\}$ in a way that $\{dt, dz_a\}$ is a local basis of $\operatorname{span}\left\{dt, dy, \dots, dy^{(n)}\right\}$ and choose z_b in a way that $\{dt, dz_a, dz_b\}$ is a local basis of $\operatorname{span}\left\{dt, dx, dy, \dots, dy^{(n)}\right\}$ around ξ. Let $u_a = \bar{y}_n^{(n+1)}$ and $u_b = \hat{u}_n$. By construction, $((z_a, z_b), (u_a, u_b))$ is a local state-representation of S around ξ, since it is linked to (x_n, u_n) by local static-state feedback (see (3.4)).

By Lemma B.1 part 8, it follows that $\operatorname{span}\{d\dot{z}_a\} \subset \operatorname{span}\{t, z_a, u_a\}$ and that (i) and (ii) hold. Now note that (iii) follows easily from Definition 4.1 and conditions (A.1).

To show that two output subsystems are Lie–Bäcklund isomorphic, let $\pi_i : V_\xi^i \to Y_i$ be local output subsystems for $i = 1, 2$. Assume that $V_\xi^1 \cap V_\xi^2 \neq \emptyset$. We will show that there exist a local Lie–Bäcklund isomorphism $\delta : W_1 \to W_2$, where H is some open neighborhood of ξ for which $H \subset V_\xi^1 \cap V_\xi^2$ and $W_i = \pi_i(H), i = 1, 2$.

Since the π_i are Lie–Bäcklund submersions for $i = 1, 2$, there exists local charts of $\phi_i = (t, X_i, Z_i)$, $i = 1, 2$, defined in some $H \subset S$ and local charts $\psi_i = (t, X_i)$, of Y_i, $i = 1, 2$, defined on $W_i = \pi_i(H)$ such that, in these coordinates $\phi_i \circ \pi_i^{-1} \circ \psi_i(t, X_i, Z_i) = (t, X_i)$, $i = 1, 2$. Since Y_1 and Y_2 are both local subsystems we have $\operatorname{span}\{dt, dX_i\} = \operatorname{span}\left\{dt, dy^{(k)} : k \in \lfloor \mathbb{N} \rceil\right\}$, for $i = 1, 2$. In particular, it follows that the local coordinate change $(t, X_1, Z_1) = \phi_1 \circ \phi_2^{-1}(t, X_2, Z_2)$ is such that $X_1 = \theta(t, X_2)$ and $X_2 = \tilde{\theta}(t, X_1)$. So the map μ defined by $(t, X_2) \mapsto (t, \theta(t, X_2))$ is a local

[20] The local differential dimension is the cardinal of the input of a local state representation. Note that a differential dimension $\tilde{m}$ of a connected smooth system that admits a local state representation around every point is a global invariant [17], [36, Corollary 7.2].

diffeomorphism.[21] Let $\delta : W_2 \subset Y_2 \to W_1 \subset Y_1$ be the local diffeomorphism defined by $\delta = \psi_1^{-1} \circ \mu \circ \psi_2$. To complete the proof it suffices to show that δ is Lie–Bäcklund. For this, we show first that $\delta \circ \pi_2|_H = \pi_1|_H$. In fact, note that

$$\begin{aligned} &\psi_1 \circ (\delta \circ \pi_2) \circ \phi_1^{-1}(t, X_1, Z_1) = \psi_1 \circ (\delta \circ \pi_2 \circ \phi_2^{-1}) \circ (\phi_2 \circ \phi_1^{-1})(t, X_1, Z_1) \\ = \;& (\psi_1 \circ \delta) \circ \pi_2 \circ \phi_2^{-1}(t, X_2, Z_2) = (\mu \circ \psi_2) \circ \pi_2 \circ \phi_2^{-1}(t, X_2, Z_2) \\ = \;& \mu \circ (\psi_2 \circ \pi_2 \circ \phi_2^{-1})(t, X_2, Z_2) = \mu(t, X_2) \\ & \qquad = \; (t, X_1) = \psi_1 \circ \pi_1 \circ \phi_1^{-1}(t, X_1, Z_1). \end{aligned}$$

From the first and the last terms above, we have that $\delta \circ \pi_2|_H = \pi_1|_H$. Denote by ∂_i the Cartan fields, respectively, of Y_i for $i = 1, 2$. By definition, $\pi_i^* \frac{d}{dt} = \partial_i \circ \pi_i$. In particular, $\partial_1 \circ \delta \circ \pi_2 = \partial_1 \circ \pi_1 = (\pi_1) * \frac{d}{dt} = (\delta \circ \pi_2)_* \frac{d}{dt} = \delta_* (\pi_2)_* \frac{d}{dt} = \delta_* \partial_2 \circ \pi_2$. As π_2 is surjective it follows that $\partial_1 \circ \delta = \delta_* \partial_2$, showing that δ is Lie–Bäcklund. □

A.2. Proof of Proposition 8.3. In order to prove Proposition 8.3 we need the following lemmas.

LEMMA A.1. *For all integers $k \geq 0$, $r \geq 0$ and for every point $p \in S$, we have*

(i) $\left(I^{(k)} + J_r + \operatorname{span}\{dt\}\right)\big|_p \cap \operatorname{span}\left\{\frac{d}{dt}\right\}^{\perp} \subset I^{(k)}(p) + J_r$. *The same result also holds when replacing J_r by J;*

(ii) $\operatorname{span}\left\{I^{(k)}, J, dt\right\}|_p = I^{(k)}(p) \oplus J(p) \oplus \operatorname{span}\{dt\}(p)$.

Proof. See [39]. □

LEMMA A.2. *Assume that the conditions of Theorem* 8.2 *are satisfied on an open neighborhood V_ξ of ξ in S. Then $I^{(k)}, k \in \mathbb{N}$, is a smooth codistribution and for every $p \in V_\xi$ and $k \in \mathbb{N}$ we have the following:*

(i) *For all $k \in \mathbb{N}$ there exists a set of covector fields $\Omega = \{\omega_1, \ldots, \omega_{r_k}\} \subset I^{(k)} + J$, where $r_k = \dim I^{(k)}$, $\omega_i = (d\theta_i - \dot{\theta}_i dt)$, with $\theta_i \in C^\infty(S)$, and an open neighborhood V of ξ such that the canonical projections of the elements of $\Omega(\nu)$ form a basis for $(I^{(k)}(\nu) + J(\nu))$ mod $J(\nu)$ for all ν in V.*

(ii) *If ω is of the form $(d\theta - \dot{\theta}dt)$ for a function $\theta \in C^\infty(S)$, then $\omega \in I^{(k+1)} + J$ if and only if $\dot{\omega} \in I^{(k)} + J$. In particular, $I^{(k)} + J \supset I^{(k+1)} + \frac{d}{dt} I^{(k+1)} + J$.*

(iii) *Let $\{\omega_1, \ldots, \omega_r\} \subset I^{(k-1)} + J$ be a set of 1-forms such that $\omega_i = (d\theta_i - \dot{\theta}_i dt)$, where $\theta_i \in C^\infty(S)$. Assume that the set $\{\omega_1(p), \ldots, \omega_r(p)\}$ is linearly independent*[22] mod $I^{(k)}(p) + J(p)$. *Then $\{\dot{\omega}_1(p), \ldots, \dot{\omega}_r(p)\} \subset (I^{(k-2)}(p) + J(p))$ is linearly independent* mod $(I^{(k-1)} + J + \operatorname{span}\{dt\})|_p$.

Proof. Assume by induction that $I^{(j)}, j = -1, 0, \ldots, k$ is smooth. We will show first that (i) and (ii) holds.

(i) We now show that, for an integer l_k big enough, $\operatorname{span}\left\{I^{(k)}, J_{l_k}, dt\right\}$ is involutive (see (8.3)). In fact, since $I^{(k)}$ is nonsingular and finite dimensional, there exist a local basis $\{\tilde{\omega}_i : i \in \lfloor r_k \rceil\}$ of $I^{(k)}$. By part (ii) of Lemma A.1, it follows that the set $\{(\tilde{\omega}_i : i \in \lfloor r_k \rceil), dx_a, du_a, \ldots, du_a^{(l_k)}, dt\}$ is a local basis of $I^{(k)} + J_{l_k} + \operatorname{span}\{dt\}$. Since the codistribution $\operatorname{span}\left\{I^{(k)}, J, dt\right\}$ is involutive, then $d\tilde{\omega}_i = \sum_{j=1}^{r_k} \eta_{ij} \wedge \nu_{ij}$ for convenient 1-forms η_{ij}, ν_{ij} with $\nu_{ij} \in \operatorname{span}\left\{I^{(k)}, J, dt\right\}$. Hence $\nu_{ij} \in \operatorname{span}\{(\tilde{\omega}_i : i \in \lfloor r_k \rceil), dx_a, du_a, \ldots, du_a^{(s_{ij})}, dt\}$. Let $l_k^* = \max_{i,j}\{s_{ij}\}$. Then $\operatorname{span}\left\{I^{(k)}, J_{l_k}, dt\right\}$ is involutive for every $l_k \geq l_k^*$. By the Frobenius theorem and part

[21] We stress that we are not using the inverse function theorem, but only the existence of the inverse of the coordinate change map.

[22] The linear independence of the set $\{\omega_1(p), \ldots, \omega_r(p)\}$ mod $(I^{(k)}(p) + J)$ for some $p \in S$ means that $\left(\sum_{i=1}^{r} \alpha_i \omega_i(p) + \omega(p)\right)\big|_p = 0$ for $\omega \in I^{(k)} + J$ and $\alpha_i \in \mathbb{R}$ implies that $\omega(p) = 0$ and $\alpha_i = 0$.

(ii) of Lemma A.1, we see that span $\left\{I^{(k)}, J_{l_k}, dt\right\}$ is spanned by linearly independent 1-forms $\{d\theta_1, \dots d\theta_{r_k}, dx_a, du_a, \dots, du_a{}^{(l_k)}, dt\}$, where $\dim I^{(k)} = r_k$. Now note that $\omega_i = (d\theta_i - \dot{\theta}_i dt) \in \text{span}\left\{\frac{d}{dt}\right\}^{\perp}$. Since $\omega_i \in \text{span}\left\{I^{(k)}, J_{l_k}, dt\right\}$, by Lemma A.1 part (i) it follows that $\omega_i \in I^{(k)} + J_{l_k}$. Let $K = \text{span}\,\{J_{l_k}, dt\}\,|_p$ and $L = \text{span}\,\{J, dt\}_p$. By construction the canonical projection of the set $\{\omega_i, i \in \lfloor r_k \rceil\}$ on $(I^{(k)}(p) + K)/K$ forms a basis of $(I^{(k)}(p) + K)/K$. By part (ii) of Lemma A.1, it is easy to see that the map $\Psi : (I^{(k)}(p) + K)/K \to (I^{(k)}(p) + L)/L$ such that $\omega(p) \bmod K \mapsto \omega(p) \bmod L$ is an isomorphism. In particular, the canonical projections of the ω_i on $(I^{(k)}(p) + L)/L$ also form a basis.

(ii) It is easy to verify by direct computation that (see [39])

$$d\omega(p) \bmod (I^{(k)}(p) + J(p)) \equiv - \left.\dot{\omega} \wedge dt\right|_p \bmod (I^{(k)}(p) + J(p)) \tag{A.2}$$

for all $p \in S$.

Now we will show that, for all $p \in S$ and $\omega \in I^{(k)}$, we have

$$\omega(p) \in I^{(k+1)}(p) \;\Leftrightarrow\; \dot{\omega}(p) \in \text{span}\left\{I^{(k)}, J, dt\right\}(p). \tag{A.3}$$

Let $\{dt, (\omega_i : i \in \lfloor r_k \rceil), \eta, (\mu_j : j \in \lfloor l_k \rceil)\}$ be a basis for span $\left\{I^{(k)}, J_{l_k}, dt\right\}$. Notice that $\left.\dot{\omega} \wedge dt\right|_p \bmod (I^{(k)}(p) + J(p)) \equiv 0$ means that $\left.\dot{\omega} \wedge dt\right|_p + \sum_{i=1}^{r_k} \left.\zeta_i \wedge \omega_i\right|_p + \left.\xi \wedge \eta\right|_p + \sum_{j=i}^{l_k} \rho_j \wedge \mu_j|_p = 0$ for convenient 1-forms ζ_i, ξ, ρ_j. From the Cartan lemma (see section 2), we conclude that $\dot{\omega}(p) \in \text{span}\left\{I^{(k)}, J_{l_k}, dt\right\}(p)$. Then, (A.3) follows from (A.2) and Definition 8.1. It is easy to show that the same arguments and the fact that J is involutive imply that

$$\omega(p) \in I^{(k+1)}(p) + J(p) \;\Leftrightarrow\; \dot{\omega}(p) \in \text{span}\left\{I^{(k)}, J, dt\right\}(p). \tag{A.4}$$

If $\omega = d\theta - \dot{\theta}dt$ then $\dot{\omega} \in \text{span}\left\{\frac{d}{dt}\right\}^{\perp}$. By (A.4) and from Lemma A.1 part (i), it follows that $\dot{\omega} \in I^{(k)} + J$. Now note that, by (i), $I^{(k+1)} + J$ has a basis for this particular form. This completes the proof of (ii). We show now that our induction hypothesis (i.e., that $I^{(j)}$ is smooth for $j = -1, 0, \dots, k$) implies that $I^{(k+1)}$ is smooth. In fact, by the proof of (i), given a local basis $\{\tilde{\omega}_i : i \in \lfloor r_k \rceil\}$ of $I^{(k)}$, there exists a local basis $\{(\tilde{\omega}_i : i \in \lfloor r_k \rceil), dx_a, (du_a^{(k)} : k \in \mathbb{N}), dt\}$ of $W_k = \text{span}\left\{I^{(k)}, J, dt\right\}$. Note that $W_k \subset W_0 = \text{span}\,\{dx_b, J, dt\}$. In particular, we have $\tilde{\omega}_i = \hat{\omega}_i + \gamma_i$, where $\hat{\omega}_i \in \text{span}\,\{dx_b\}$ and $\gamma_i \in \text{span}\,\{J, dt\} = T^*S_a$. Note that $\mu_i = \dot{\gamma}_i \in \text{span}\,\{J, dt\}$ and we may replace $\tilde{\omega}_i$ by $\hat{\omega}_i$ in the basis of W_k , obtaining another basis of W_k. Note also that there exists a subset $\hat{x}_b$ of x_b such that $\{d\hat{x}_b, du_b, (\tilde{\omega}_i : i \in \lfloor r_k \rceil), dx_a, (du_a^{(k)} : k \in \mathbb{N}), dt\}$ is a basis of W_{-1}. Let $z = (\hat{x}_b, u_b)$. Let $\dot{\tilde{\omega}}_i = \dot{\hat{\omega}}_i + \dot{\gamma}_i = \sum_j a_{ij} dz_j + \mu_i$, where $\mu_i \in \text{span}\,\{J, dt\}$. Denote the matrix formed by the functions a_{ij} by A. By (A.3), $\omega(p) = \sum_j \alpha_i \tilde{\omega}_i \in I^{(k+1)}(p)$ if and only if $\sum_j (\dot{\alpha}_i \tilde{\omega}_i + \alpha_i \dot{\tilde{\omega}}_i)|_p \in W_k(p)$. Denoting by α the column vector with components α_i, then $\omega(p) \in I^{(k+1)}(p)$ if and only if $A(p)\alpha(p) = 0$. Then $I^{(k+1)}$ is nonsingular if and only if $A(p)$ has (locally) constant rank and in this case it is clear that $I^{(k)}$ is smooth.[23]

(iii) To prove (iii), assume that there exists ω in $I^{(k-1)} + J$ and functions $\alpha_i \in C^\infty(S)$ such that for $p \in S$, then $\left.(\omega + \alpha_0 dt + \sum_{i=1}^{r} \alpha_i \dot{\omega}_i)\right|_p = 0$. Hence, $\left.\{[\omega - \sum_{i=1}^{r} (\dot{\alpha}_i)\omega_i] + \alpha_0 dt + \frac{d}{dt}(\sum_{i=1}^{r} \alpha_i \omega_i)\}\right|_p = 0$. Since $[\omega - \sum_{i=1}^{r} (\dot{\alpha}_i)\omega_i](p) \in I^{(k-1)}(p) + J(p)$, it follows that $\left.\frac{d}{dt}(\sum_{i=1}^{r} \alpha_i \omega_i)\right|_p \in \text{span}\left\{I^{(k-1)}, J, dt\right\}(p)$. It follows

[23] This proof also shows that one may compute the relative derived flag by solving linear equations.

from (A.4) that $(\sum_{i=1}^{r}\alpha_i\omega_i)(p) \in I^{(k)}(p) + J(p)$ and hence the set $\{\omega_1, \dots, \omega_r\}$ is not linearly independent mod $I^{(k)} + J(p)$ in $p \in S$. □

Proof (of Proposition 8.3*).* Since $\omega_i = d\theta_i - \dot{\theta}_i dt$, $i \in \lfloor s\rceil$, by part (ii) of Lemma A.1, the set $\mathcal{B} = \{(d\theta_i^{(j)}, j \in \{0, \dots, r_i\}, i \in \lfloor s\rceil), dx_a, du_a, \dots, du_a^{(l_{k-1})}, dt\}$ is a local basis of $I^{(k-1)} + J_{l_{k-1}} + \operatorname{span}\{dt\}$ for any $l_{k-1} > l^*_{k-1}$ for some l^*_{k-1}.

By part (iii) of Lemma A.2, $\dot{B}$ is linearly independent mod $\{I^{(k-1)} + J + \operatorname{span}\{dt\}\}$. Hence $\mathcal{B} \cup \dot{\mathcal{B}} = \{(d\theta_i^{(j)}, j \in \{0, \dots, r_i + 1\}, i \in \lfloor s\rceil), dx_a, du_a, \dots, du_a^{(l_{k-1})}, dt\}$ is linearly independent for all l_{k-1}.

From the proof of part (i) of Lemma A.2, we also have that there exists a local basis $\{(\tilde{\theta}_i : i \in \lfloor r\rceil), dx_a, du_a, \dots, du_a^{(l_{k-2})}, dt\}$ of $I^{(k-2)} + J_{l_{k-2}} + \operatorname{span}\{dt\}$ for every $l_{k-2} \geq l^*_{k-2}$. Let $l_{k-1} = l_{k-2} = \max\{l^*_{k-1}, l^*_{k-2}\}$.

As $I^{(k-1)} \subset I^{(k-2)}$, we may complete $\mathcal{B} \cup \dot{\mathcal{B}}$ with a subset $\hat{\mathcal{B}} = \{\theta_i, i = s+1, \dots, \sigma\}$ of $\{\tilde{\theta}_i : i \in \lfloor r\rceil\}$ in order to form a basis of $I^{(k-2)} + J_{l_{k-2}} + \operatorname{span}\{dt\}$. By the same reasoning of the end of the proof of part (i) of Lemma A.2, it follows that $B \cup \dot{B} \cup \hat{B}$ is a basis of $I^{(k-2)} + J$. The fact that $\dot{\mathcal{B}} \cup \hat{\mathcal{B}}$ is linearly independent mod $(I^{(k-1)} + J + \operatorname{span}\{dt\})$ implies that $\dot{B} \cup \hat{B}$ is also linearly independent mod $(I^{(k-1)} + J)$. □

Appendix B. Geometric interpretation of the dynamic extension algorithm.

In [13] it was shown, using an algebraic approach, that the output rank (the number of differentially independent outputs [14]) can be computed by the *structure algorithm* [49] and the *dynamic extension algorithm* [12, 33]. This interpretation was developed further in [10] in order to study control synthesis problems by quasi-static state feedback. In [36], the algebraic results of [13, 10] are translated to the differential geometric approach of [19], giving the following lemma.

Lemma B.1 (see [36, Lemma 8.2]). *Consider the analytic (explicit) system S defined by* (1.1a) *with analytic output $y = h(t, x, u)$. Let S_k be the open and dense set of regular points of the codistributions $Y_i = \operatorname{span}\{dt, dy, \dots, dy^{(i)}\}$ and $\mathcal{Y}_i =$ span $\{dt, dx, dy, \dots, dy^{(i)}\}$. In the kth step of the dynamic extension algorithm, one may construct a partition*[24] *$y = (\bar{y}_k, \hat{y}_k)$ and a new local classical state representation (x_k, u_k) of the system S with state $x_k = (x, \bar{y}_0^{(0)}, \dots, \bar{y}_k^{(k)})$ and input $u_k = (\bar{y}_k^{(k+1)}, \hat{u}_k)$, defined in an open neighbourhood V_ξ of $\xi \in S_k$, such that*

1. $\operatorname{span}\{dt, dx_k\} = \operatorname{span}\{dt, dx, dy, \dots, dy^{(k)}\}$.
2. $\operatorname{span}\{dt, dx_k, du_k\} = \operatorname{span}\{dt, dx, dy, \dots, dy^{(k+1)}, du\}$.
3. *It is always possible to choose $\bar{y}_{k+1}^{(k+1)}$ in a way that $\bar{y}_k^{(k+1)} \subset \bar{y}_{k+1}^{(k+1)}$.*
4. *It is always possible to choose $\hat{u}_{k+1} \subset \hat{u}_k$.*
5. *Let $\mathcal{D}(\mathcal{C})$ denote the generic dimension of a codistribution $\mathcal{C}$ generated by the differentials of a finite set of analytic functions. The sequence $\sigma_k = \mathcal{D}(\mathcal{Y}_k) - \mathcal{D}(\mathcal{Y}_{k-1})$ is nondecreasing, the sequence $\rho_k = \mathcal{D}(Y_k) - \mathcal{D}(Y_{k-1})$ is nonincreasing, and both sequences converge to the same integer ρ, called the* output rank, *for some $k^* \leq n =$* $\dim x$.
6. *$S_k = S_{k^*}$ for $k \geq k^*$.*
7. *$Y_k \cap \operatorname{span}\{dx\}|_\nu = Y_{k^*-1} \cap \operatorname{span}\{dx\}|_\nu$ for every $\nu \in S_{k^*}$ and $k \geq k^*$.*
8. *Around $\xi \in U_k$, one may choose, $\bar{y}_k = \bar{y}_{k^*}$ for $k \geq k^*$. Furthermore, $Y_{k+1} =$* $Y_k + \operatorname{span}\{\bar{y}_k^{(k+1)}\}$ *for $k \geq k^*$.*

[24] Including a possible reordering of the outputs.

Proof. A complete proof of this result can be found in [36] (see [13, Theorem 2.5] and [10, Lemma 4.1.6] for similar results in algebraic contexts). □

Acknowledgments. The authors are indebted to Michel Fliess, Joachim Rudolph, Felipe M. Pait, and Emmanuel Delaleau for their suggestions concerning this paper.

REFERENCES

[1] E. Aranda-Bricaire, C. H. Moog, and J. B. Pomet, *A linear algebraic framework for dynamic feedback linearization*, IEEE Trans. Automat. Control, 40 (1995), pp. 127–132.

[2] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Springer-Verlag, New York, 1995.

[3] R. Briant, S. Chern, R. Gardner, H. Goldschmidt, and P. Griffiths, *Exterior Differential Systems*, Springer-Verlag, New York, 1991.

[4] S. L. Campbell, *Singular Systems of Differential Equations*, Pitman, London, 1982.

[5] B. Charlet, J. Lévine, and R. Marino, *On dynamic feedback linearization*, Systems Control Lett., 13 (1989), pp. 143–151.

[6] B. Charlet, J. Lévine, and R. Marino, *Sufficient conditions for dynamic state feedback linearization*, SIAM J. Control Optim., 29 (1991), pp. 38–57.

[7] X. Chen and M. A. Shayman, *Dynamics and control of constrained nonlinear systems with application to robotics*, in Proceedings of the American Control Conference, Chicago, IL, 1992, pp. 2962–2966.

[8] M. A. Christodoulou and C. Işik, *Feedback control for nonlinear singular systems*, Internat. J. Control, 51 (1990), pp. 487–494.

[9] C. Corrêa Filho and P. S. Pereira da Silva, *A Sufficient Condition of Relative Flatness*, 2000, manuscript.

[10] E. Delaleau and P. S. Pereira da Silva, *Filtrations in feedback synthesis: Part* I*—systems and feedbacks*, Forum Math., 10 (1998), pp. 147–174.

[11] E. Delaleau and J. Rudolph, *Control of flat systems by quasi-static feedback of generalized states*, Internat. J. Control, 71 (1998), pp. 745–765.

[12] J. Descusse and C. H. Moog, *Decoupling with dynamic compensation for strong invertible affine non-linear systems*, Internat. J. Control, 42 (1985), pp. 1387–1398.

[13] M. D. Di Benedetto, J. W. Grizzle, and C. H. Moog, *Rank invariants of nonlinear systems*, SIAM J. Control Optim., 27 (1989), pp. 658–672.

[14] M. Fliess, *Automatique et corps différentiels*, Forum Math., 1 (1989), pp. 227–238.

[15] M. Fliess, *Some basic structural properties of generalized linear systems*, Systems Control Lett., 15 (1990), pp. 391–396.

[16] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Sur les systèmes non linéaires différentiellement plats*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 619–624.

[17] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Linéarisation par bouclage dynamique et transformations de Lie-Bäcklund*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 981–986.

[18] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Flatness and defect of non-linear systems: Introductory theory and examples*, Internat. J. Control, 61 (1995), pp. 1327–1361.

[19] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 922–937.

[20] R. B. Gardner and W. F. Shadwick, *The GS algorithm for exact linearization to Brunovsky normal form*, IEEE Trans. Automat. Control, 37 (1992), pp. 224–230.

[21] S. T. Glad, *Differential algebraic modelling of nonlinear systems*, in Realization and Modelling in System Theory (Amsterdam, 1989), Progr. Systems Control Theory 3, Birkhäuser, Boston, MA, 1990, pp. 97–105.

[22] M. Guay, P. J. McLellan, and D. W. Bacon, *A condition for dynamic feedback linearization of control-affine nonlinear systems*, Internat. J. Control, 68 (1997), pp. 87–106.

[23] L. R. Hunt, R. Su, and G. Meyer, *Design for multi-input nonlinear systems*, in Differential Geometric Methods in Nonlinear Control Theory, R. Brocket, R. Millmann, and H. J. Sussmann, eds., Birkhäuser, Boston, MA, 1983, pp. 268–298.

[24] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, Berlin, 1995.

[25] B. Jakubczyk and W. Respondek, *On linearization of control systems*, Bull. Acad. Polon. Sci., Sér. Sci. Math., 28 (1980), pp. 517–522.

[26] S. Kawaji and E. Z. Taha, *Feedback linearization of a class of nonlinear descriptor systems*, in Proceedings of the 33rd IEEE Conference on Decision Control, Vol. 4, IEEE, New York, 1994, pp. 4035–4037.

[27] I. S. Krasil'shchik, V. V. Lychagin, and A. M. Vinogradov, *Geometry of Jet Spaces and Nonlinear Partial Differential Equations*, Gordon and Breach, New York, 1986.

[28] H. Krishnan and N. H. McClamroch, *Tracking in nonlinear differential-algebraic control systems with applications to constrained robot systems*, Automatica J. IFAC, 30 (1994), pp. 1885–1897.

[29] J. Y. Lin and N. U. Ahmed, *Approach to controllability problems for singular systems*, Internat. J. Control, 22 (1991), pp. 675–690.

[30] X. P. Liu, *On linearization of nonlinear singular control systems*, in Proceedings of the American Control Conference, San Francisco, CA, 1993, pp. 2284–2287.

[31] D. Liyi, *Singular Control Systems*, Springer-Verlag, New York, 1989.

[32] N. H. McClamroch, *Feedback stabilization of control systems described by a class of nonlinear differential-algebraic equations*, Systems Control Lett., 15 (1990), pp. 53–60.

[33] H. Nijmeijer and W. Respondek, *Dynamic input-output decoupling of nonlinear control systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 1065–1070.

[34] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[35] P. S. Pereira da Silva, *On decompositions for noncontrollable nonlinear systems*, Revista Controle & Automação, 6 (1997), pp. 134–144. Also available online at http://www.lac.usp.br/~paulo/.

[36] P. S. Pereira da Silva, *Some Geometric Properties of the Dynamic Extension Algorithm*, Tech. report BT / PTC / 0008, Escola Politécnica da Universidade de São Paulo, São Paulo, Brazil, 2000. Also available online at http://www.lac.usp.br/~paulo/.

[37] P. S. Pereira da Silva and C. Corrêa Filho, *Relative flatness and flatness of implicit systems*, in Proceedings of the 4th IFAC Nonlinear Control Systems Design Symposium, Enschede, The Netherlands, 1998, Vol. 2, Pergamon, New York, 1998, pp. 516–522.

[38] P. S. Pereira da Silva and C. Corrêa Filho, *Flatness Based Position/Force Control of Constrained Under-Actuated Robots*, in CDROM Anais Do XIII Congresso Brasileiro de Automática, Florianópolis, Brazil, 2000. Also available online at http://www.lac.usp.br/~paulo/.

[39] P. S. Pereira da Silva and C. Corrêa Filho, *Relative Flatness and Flatness of Implicit Systems*, Tech. report BT/PTC/0019, Escola Politécnica da Universidade de São Paulo, São Paulo, Brazil, 2000.

[40] J.-B. Pomet, *A differential geometric setting for dynamic equivalence and dynamic linearization*, in Geometry in Nonlinear Control and Differential Inclusions, B. Jakubczyk, W. Respondek, and T. Rzezuchowski, eds., Banach Center Publ., Polish Academy of Science, Warsaw, 1995, pp. 319–339.

[41] J.-B. Pomet, *On dynamic feedback linearization of four-dimensional affine control systems with two inputs*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 151–230.

[42] M. Rathinam and W. M. Sluis, *A test for differential flatness by reduction to single input systems*, in CDROM Proceedings of the 13th IFAC World Congress, paper 2b–08 3, 1996, pp. 257–262.

[43] W. Rheinboldt, *On the existence and uniqueness of solutions of nonlinear semi-implicit differential-algebraic equations*, Nonlinear Anal., 16 (1991), pp. 642–661.

[44] P. Rouchon, *Necessary condition and genericity of dynamic feedback linearization*, J. Math. Systems Estim. Control, 5 (1995), pp. 345–358.

[45] P. Rouchon, M. Fliess, and J. Levine, *Kronecker's canonical forms for nonlinear implicit differential systems*, in Proceedings of the 2nd IFAC Conference on System Struc. Control, Nantes, France, 1995, pp. 248–251.

[46] J. Rudolph, *Well-formed dynamics under quasi-static state feedback*, in Geometry in Nonlinear Control and Differential Inclusions, B. Jakubczyk, W. Respondek, and T. Rzezuchowski, eds., Banach Center Publ., Polish Academy of Science, Warsaw, 1995, pp. 349–360.

[47] K. Schlacher, A. Kugi, and W. Haas, *Geometric control of a class of nonlinear descriptor systems*, in Proceedings of the 4th IFAC Nonlinear Control Systems Design Symposium, Enschede, The Netherlands, Vol. 2, 1998, pp. 387–392.

[48] W. F. Shadwick, *Absolute equivalence and dynamic feedback linearization*, Systems Control Lett., 15 (1990), pp. 35–39.

[49] S. N. Singh, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, AC–26 (1981), pp. 595–598.

[50] W. M. Sluis, *Absolute Equivalence and Its Application to Control Theory*, Ph.d. thesis, University of Waterloo, Canada, 1992.

[51] W. M. Sluis, *A necessary condition for dynamic feedback linearization*, Systems Control Lett., 21 (1993), pp. 277–283.

[52] W. M. Sluis and D. M. Tilbury, *A bound on the number of integrators needed to linearize a control system*, Systems Control Lett., 29 (1996), pp. 43–50.

[53] D. Tilbury, R. M. Murray, and S. R. Sastry, *Trajectory generation for the n-trailer problem using Goursat normal form*, IEEE Trans. Automat. Control, 40 (1995), pp. 802–819.

[54] M. van Nieuwstadt, M. Rathinam, and R. M. Murray, *Differential flatness and absolute equivalence of nonlinear control systems*, SIAM J. Control Optim., 36 (1998), pp. 1225–1239.

[55] F. W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Scott Foresman and Company, Glenview, IL, 1971.

[56] C. J. Watanabe, P. S. Pereira da Silva, and P. A. Tonelli, *Algebra diferencial em teoria de controle*, Tech. report, Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, Brazil, 2000. Also available online at http://www.lac.usp.br/∼paulo/.

[57] V. V. Zharinov, *Geometrical Aspects of Partial Differential Equations*, World Scientific, Singapore, 1992.

SIAM J. CONTROL OPTIM.
Vol. 39, No. 6, pp. 1952–1973

FEEDBACK STABILIZATION OVER COMMUTATIVE RINGS: FURTHER STUDY OF THE COORDINATE-FREE APPROACH[*]

KAZUYOSHI MORI[†][‡] AND KENICHI ABE[†]

Abstract. This paper is concerned with the coordinate-free approach to control systems. The coordinate-free approach is a factorization approach but does not require the coprime factorizations of the plant. We present two criteria for feedback stabilizability for multi-input multi-output (MIMO) systems in which transfer functions belong to the total rings of fractions of commutative rings. Both of them are generalizations of Sule's results in [*SIAM J. Control Optim.*, 32 (1994), pp. 1675–1695]. The first criterion is expressed in terms of modules generated from a causal plant and does not require the plant to be strictly causal. It shows that if the plant is stabilizable, the modules are projective. The other criterion is expressed in terms of ideals called generalized elementary factors. This gives the stabilizability of a causal plant in terms of the coprimeness of the generalized elementary factors. As an example, a discrete finite-time delay system is considered.

Key words. linear systems, feedback stabilization, coprime factorization over commutative rings

AMS subject classifications. 93C05, 93D15, 93B50, 93B25

PII. S0363012998336625

1. Introduction.

In this paper we are concerned with the coordinate-free approach to control systems. This approach is a factorization approach but does not require the coprime factorizations of the plant.

The factorization approach to control systems has the advantage that it embraces, within a single framework, numerous linear systems such as continuous-time as well as discrete-time systems, lumped as well as distributed systems, one-dimensional as well as n-dimensional systems, etc. [14]. This factorization approach was patterned after Desoer et al. [3] and Vidyasagar, Schneider, and Francis [14]. In this approach, when problems such as feedback stabilization are studied, one can focus on the key aspects of the problem under study rather than be distracted by the special features of a particular class of linear systems. A transfer function of this approach is given as the ratio of two stable causal transfer functions, and the set of stable causal transfer functions is a commutative ring. For a long time, the theory of the factorization approach had been founded on the coprime factorizability of transfer matrices, which is satisfied in the case where the set of stable causal transfer functions is such a commutative ring as a Euclidean domain, a principal ideal, or a Bézout domain. However, Anantharam in [1] showed that there exist models in which some stabilizable plants do not have right-/left-coprime factorizations.

Recently, Shankar and Sule in [10] have presented a theory of feedback stabilization for single-input single-output (SISO) transfer functions having fractions over general integral domains. Moreover, Sule in [11, 12] has presented a theory of the

[*] Received by the editors April 1, 1998; accepted for publication (in revised form) October 10, 2000; published electronically April 24, 2001. This research was supported by the Ministry of Education, Japan, under grant 12780301.

http://www.siam.org/journals/sicon/39-6/33662.html

[†] Department of Electrical Engineering, Faculty of Engineering, Tohoku University, Sendai 980-8579, Japan (abe@abe.ecei.tohoku.ac.jp).

[‡] Current address: School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu 965-8580, Japan (Kazuyoshi.MORI@IEEE.ORG). This paper was partially written while the first author was visiting the Institut de Recherche en Cybernétique de Nantes, UMR 6597, France (September 1998–June 1999).

feedback stabilization of strictly causal plants for multi-input multi-output (MIMO) transfer matrices, in which transfer functions belong to the total rings of fractions of commutative rings, with some restrictions. Their approach to the control systems is called a "coordinate-free approach" [10, p. 15] in the sense that they do not require the coprime factorizability of transfer matrices. Our objective in this paper is to decrease the restrictions in order to make further comprehensive theory of the coordinate-free approach, so that the theory can be applied to more and more linear control models including ones not yet well understood.

The main contribution of this paper consists of providing two criteria for feedback stabilizability for MIMO systems in which transfer functions belong to the total rings of fractions of commutative rings: the first criterion is expressed in terms of modules ((ii) of Theorem 3.3) and the other in terms of ideals called generalized elementary factors ((iii) of Theorem 3.3). They are more general than Sule's results in the following sense: (i) our results do not require that plants are strictly causal; (ii) we do not employ the restriction of commutative rings. Further, we will not use the theory of algebraic geometry.

The paper is organized as follows. In section 2, we give mathematical preliminaries, set up the feedback stabilization problem, present the previous results, and define the causality of the transfer functions. In section 3, we state our main results. As a preface to our main results, we also introduce there the notion of the generalized elementary factor of a plant. In section 4, we give intermediate results which we will utilize in the proof of the main theorem. In section 5, we prove our main theorem. In section 6, we discuss the causality of the stabilizing controllers. Also, in order to make the contents clear, we present examples concerning a discrete finite-time delay system in sections 3, 4, and 5 consecutively.

2. Preliminaries.

In the following we begin by introducing the notations of commutative rings, matrices, and modules commonly used in this paper. Then we give the formulation of the feedback stabilization problem and the previous results.

2.1. Notations.

Commutative rings. In this paper, we consider that any commutative ring has the identity 1 different from zero. Let $\mathcal{R}$ denote a commutative ring. A *zerodivisor* in $\mathcal{R}$ is an element x for which there exists a nonzero y such that $xy = 0$. In particular, a zerodivisor x is said to be *nilpotent* if $x^n = 0$ for some positive integer n. The set of all nilpotent elements in $\mathcal{R}$, which is an ideal, is called the *nilradical* of $\mathcal{R}$. A *nonzerodivisor* in $\mathcal{R}$ is an element which is not a zerodivisor. The total ring of fractions of $\mathcal{R}$ is denoted by $\mathcal{F}(\mathcal{R})$.

The set of all prime ideals of $\mathcal{R}$ is called the *prime spectrum* of $\mathcal{R}$ and is denoted by $\operatorname{Spec}\mathcal{R}$. The prime spectrum of $\mathcal{R}$ is said to be *irreducible* as a topological space if every nonempty open set is dense in $\operatorname{Spec}\mathcal{R}$.

We will consider that the set of stable causal transfer functions is a commutative ring, denoted by $\mathcal{A}$. From the sense of the transfer functions we consider that the commutative ring $\mathcal{A}$ satisfies the invariant basis property (cf. [6]). In addition to $\mathcal{A}$, we will use the following three kinds of rings of fractions:

(i) The first one appears as the total ring of fractions of $\mathcal{A}$, which is denoted by $\mathcal{F}(\mathcal{A})$ or simply by $\mathcal{F}$; that is, $\mathcal{F} = \{n/d \mid n, d \in \mathcal{A},\ d \text{ is a nonzerodivisor}\}$. This will be considered to be the set of all possible transfer functions. If the commutative ring $\mathcal{A}$ is an integral domain, $\mathcal{F}$ becomes a field of fractions of $\mathcal{A}$. However, if $\mathcal{A}$ is not an integral domain, then $\mathcal{F}$ is not a field, because any zerodivisor of $\mathcal{F}$ is not a unit.

(ii) The second one is associated with the set of powers of a nonzero element of $\mathcal{A}$. Suppose that f denotes a nonzero element of $\mathcal{A}$. Given a set $S_f = \{1, f, f^2, \ldots\}$, which is a multiplicative subset of $\mathcal{A}$, we denote by $\mathcal{A}_f$ the ring of fractions of $\mathcal{A}$ with respect to the multiplicative subset S_f; that is, $\mathcal{A}_f = \{n/d \,|\, n \in \mathcal{A},\ d \in S_f\}$. We point out two facts here: (a) In the case where f is nilpotent, $\mathcal{A}_f$ becomes isomorphic to $\{0\}$. (b) In the case where f is a zerodivisor, even if the equality $a/1 = b/1$ holds over $\mathcal{A}_f$ with $a, b \in \mathcal{A}$, we cannot say in general that $a = b$ over $\mathcal{A}$; alternatively, $a = b + z$ over $\mathcal{A}$ holds with some zerodivisor z of $\mathcal{A}$ such that $zf^\omega = 0$ with a sufficiently large integer ω.

(iii) The last one is the total ring of fractions of $\mathcal{A}_f$, which is denoted by $\mathcal{F}(\mathcal{A}_f)$; that is, $\mathcal{F}(\mathcal{A}_f) = \{n/d \,|\, n, d \in \mathcal{A}_f,\ d \text{ is a nonzerodivisor of } \mathcal{A}_f\}$. If f is a nonzerodivisor of $\mathcal{A}$, $\mathcal{F}(\mathcal{A}_f)$ coincides with the total ring of fractions of $\mathcal{A}$. Otherwise, they may not coincide.

The reader is referred to Chapter 3 of [2] for the ring of fractions and to Chapter 1 of [2] for the prime spectrum.

In the rest of the paper, we will use $\mathcal{R}$ as an unspecified commutative ring and mainly suppose that $\mathcal{R}$ denotes either $\mathcal{A}$ or $\mathcal{A}_f$.

Matrices. Suppose that x and y denote sizes of matrices.

The set of matrices over $\mathcal{R}$ of size $x \times y$ is denoted by $\mathcal{R}^{x\times y}$. In particular, the set of square matrices over $\mathcal{R}$ of size x is denoted by $(\mathcal{R})_x$. A square matrix is called *singular* over $\mathcal{R}$ if its determinant is a zerodivisor of $\mathcal{R}$ and *nonsingular* otherwise. The identity and the zero matrices are denoted by E_x and $O_{x\times y}$, respectively, if the sizes are required, otherwise they are denoted simply by E and O. For a matrix A over $\mathcal{R}$, the inverse matrix of A is denoted by A^{-1} provided that $\det(A)$ is a unit of $\mathcal{F}(\mathcal{R})$. The ideal generated by $\mathcal{R}$-linear combination of all minors of size m of a matrix A is denoted by $I_{m\mathcal{R}}(A)$.

Matrices A and B over $\mathcal{R}$ are *right-coprime over* $\mathcal{R}$ if there exist matrices $\widetilde{X}$ and $\widetilde{Y}$ over $\mathcal{R}$ such that $\widetilde{X}A + \widetilde{Y}B = E$. Analogously, matrices $\widetilde{A}$ and $\widetilde{B}$ over $\mathcal{R}$ are *left-coprime over* $\mathcal{R}$ if there exist matrices X and Y over $\mathcal{R}$ such that $\widetilde{A}X + \widetilde{B}Y = E$. Note that, in the sense of the above definition, even if two matrices have no common right-(left-)divisors except invertible matrices, they may not be right-(left-)coprime over $\mathcal{R}$. (For example, two matrices $[\,z_1\,]$ and $[\,z_2\,]$ of size 1×1 over the bivariate polynomial ring $\mathbb{R}[z_1, z_2]$ over the real field $\mathbb{R}$ are neither right- nor left-coprime over $\mathbb{R}[z_1, z_2]$ in our setting.) Further, a pair (N, D) of matrices N and D is said to be a *right-coprime factorization of P over $\mathcal{R}$* if (i) the matrix D is nonsingular over $\mathcal{R}$, (ii) $P = ND^{-1}$ over $\mathcal{F}(\mathcal{R})$, and (iii) N and D are right-coprime over $\mathcal{R}$. Also, a pair $(\widetilde{N}, \widetilde{D})$ of matrices $\widetilde{N}$ and $\widetilde{D}$ is said to be a *left-coprime factorization of P over $\mathcal{R}$* if (i) $\widetilde{D}$ is nonsingular over $\mathcal{R}$, (ii) $P = \widetilde{D}^{-1}\widetilde{N}$ over $\mathcal{F}(\mathcal{R})$, and (iii) $\widetilde{N}$ and $\widetilde{D}$ are left-coprime over $\mathcal{R}$. As we have seen, in the case where a matrix is potentially used to express *left* fractional form and/or *left* coprimeness, we usually attach a tilde "$\widetilde{\ }$" to a symbol; for example $\widetilde{N}$, $\widetilde{D}$ for $P = \widetilde{D}^{-1}\widetilde{N}$ and $\widetilde{Y}$, $\widetilde{X}$ for $\widetilde{Y}N + \widetilde{X}D = E$.

Modules. For a matrix A over $\mathcal{R}$, we denote by $M_r(A)$ $(M_c(A))$ the $\mathcal{R}$-module generated by rows (columns) of A.

Suppose that A, B, $\widetilde{A}$, $\widetilde{B}$ are matrices over $\mathcal{R}$ and X is a matrix over $\mathcal{F}(\mathcal{R})$ such that $X = AB^{-1} = \widetilde{B}^{-1}\widetilde{A}$ with B and $\widetilde{B}$ being nonsingular. Then the $\mathcal{R}$-module $M_r([\,A^t \quad B^t\,]^t)$ $(M_c([\,\widetilde{A} \quad \widetilde{B}\,]))$ is uniquely determined up to isomorphism with respect to any choice of fractions AB^{-1} $(\widetilde{B}^{-1}\widetilde{A})$ of X as shown in Lemma 2.1 below. Thus for a matrix X over $\mathcal{F}(\mathcal{R})$, we denote by $\mathcal{T}_{X,\mathcal{R}}$ and $\mathcal{W}_{X,\mathcal{R}}$ the mod-

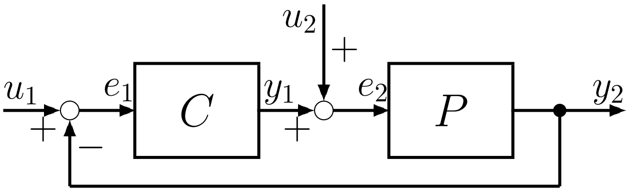


FIG. 2.1. *Feedback system* Σ.

ules $M_r([\,A^t \quad B^t\,]^t)$ and $M_c([\,\widetilde{A} \quad \widetilde{B}\,])$, respectively. If $\mathcal{R} = \mathcal{A}$, we write simply $\mathcal{T}_X$ and $\mathcal{W}_X$ for $\mathcal{T}_{X,\mathcal{A}}$ and $\mathcal{W}_{X,\mathcal{A}}$, respectively. We will use, for example, the notations $\mathcal{T}_P$, $\mathcal{W}_P$, $\mathcal{T}_C$, and $\mathcal{W}_C$ for the matrices P and C over $\mathcal{F}$.

An $\mathcal{R}$-module M is called *free* if it has a basis, that is, a linearly independent system of generators. The *rank* of a free $\mathcal{R}$-module M is equal to the cardinality of a basis of M, which is independent of the basis chosen. An $\mathcal{R}$-module M is called *projective* if it is a direct summand of a free $\mathcal{R}$-module, that is, there is a module N such that $M \oplus N$ is free. The reader is referred to Chapter 2 of [2] for the module theory.

LEMMA 2.1. *Suppose that X is a matrix over $\mathcal{F}(\mathcal{R})$ and is expressed in the form of a fraction $X = AB^{-1} = \widetilde{B}^{-1}\widetilde{A}$ with some matrices A, B, $\widetilde{A}$, $\widetilde{B}$ over $\mathcal{R}$. Then the $\mathcal{R}$-module $M_r([\,A^t \quad B^t\,]^t)$ $(M_c([\,\widetilde{A} \quad \widetilde{B}\,]))$ is uniquely determined up to isomorphism with respect to any choice of fractions AB^{-1} $(\widetilde{B}^{-1}\widetilde{A})$ of X.*

Proof. Without loss of generality, it is sufficient to show that $M_r([\,A_1^t \quad b_1E\,]^t) \simeq M_r([\,A_2^t \quad B_2^t\,]^t)$, where $b_1 \in \mathcal{R}$ and $A_1(b_1E)^{-1} = A_2B_2^{-1}$. Since b_1 is a nonzerodivisor and B_2 is nonsingular, we have $M_r([\,A_1^t \quad b_1E\,]^t) \simeq M_r([\,A_1^t \quad b_1E\,]^t B_2) \simeq M_r([\,A_2^t \quad B_2^t\,]^t b_1E) \simeq M_r([\,A_2^t \quad B_2^t\,]^t)$. The other isomorphism can be proved analogously. □

2.2. Feedback stabilization problem. The stabilization problem considered in this paper follows that of Sule in [11], who considers the feedback system Σ [13, Chapter 5, Figure 5.1] as in Figure 2.1. For further details, see [13]. Let a commutative ring $\mathcal{A}$ represent the set of *stable causal* transfer functions. The total ring of fractions of $\mathcal{A}$, denoted by $\mathcal{F}$, consists of *all* possible transfer functions. The set of matrices of size $x \times y$ over $\mathcal{A}$, denoted by $\mathcal{A}^{x\times y}$, coincides with the set of stable causal transfer matrices of size $x \times y$. Also the set of matrices of size $x \times y$ over $\mathcal{F}$, denoted by $\mathcal{F}^{x\times y}$, coincides with all possible transfer matrices of size $x \times y$. Throughout the paper, the plant we consider has m inputs and n outputs, and its transfer matrix, which itself is also called simply a *plant*, is denoted by P and belongs to $\mathcal{F}^{n\times m}$. We will occasionally consider matrices over $\mathcal{A}$ $(\mathcal{F})$ as ones over $\mathcal{A}_f$ or $\mathcal{F}$ $(\mathcal{F}(\mathcal{A}_f))$ by natural mapping.

DEFINITION 2.2. *Define $\widehat{F}_{\mathrm{ad}}$ by*

$$\widehat{F}_{\mathrm{ad}} = \{(X,Y) \in \mathcal{F}^{x\times y} \times \mathcal{F}^{y\times x} \mid \ \det(E_x + XY) \text{ is a unit of } \mathcal{F}, \\ x \text{ and } y \text{ are positive integers}\}.$$

For $P \in \mathcal{F}^{n\times m}$ and $C \in \mathcal{F}^{m\times n}$, the matrix $H(P,C) \in (\mathcal{F})_{m+n}$ is defined by

$$H(P,C) = \begin{bmatrix} (E_n + PC)^{-1} & -P(E_m + CP)^{-1} \\ C(E_n + PC)^{-1} & (E_m + CP)^{-1} \end{bmatrix} \tag{2.1}$$

provided $(P, C) \in \widehat{F}_{\mathrm{ad}}$. *This* $H(P, C)$ *is the transfer matrix from* $[\, u_1^t \quad u_2^t \,]^t$ *to* $[\, e_1^t \quad e_2^t \,]^t$ *of the feedback system* Σ. *If* (i) $(P, C) \in \widehat{F}_{\mathrm{ad}}$ *and* (ii) $H(P, C) \in (\mathcal{A})_{m+n}$, *then we say that the plant* P *is* stabilizable, P *is* stabilized *by* C, *and* C *is a* stabilizing controller *of* P.

Here we define the causality of transfer functions, which is an important physical constraint, used in this paper. We employ the definition of causality from Vidyasagar, Schneider, and Francis [14, Definition 3.1] and introduce two terminologies later used frequently.

Definition 2.3. *Let* $\mathcal{Z}$ *be a prime ideal of* $\mathcal{A}$, *with* $\mathcal{Z} \neq \mathcal{A}$, *including all zerodivisors. Define the subsets* $\mathcal{P}$ *and* $\mathcal{P}_S$ *of* $\mathcal{F}$ *as follows:*

$$\mathcal{P} = \{n/d \in \mathcal{F} \,|\, n \in \mathcal{A},\ d \in \mathcal{A}\backslash\mathcal{Z}\}, \quad \mathcal{P}_S = \{n/d \in \mathcal{F} \,|\, n \in \mathcal{Z},\ d \in \mathcal{A}\backslash\mathcal{Z}\}.$$

A transfer function in $\mathcal{P}$ $(\mathcal{P}_S)$ *is called* causal (strictly causal). *Similarly, if every entry of a transfer matrix over* $\mathcal{F}$ *is in* $\mathcal{P}$ $(\mathcal{P}_S)$, *the transfer matrix is called* causal (strictly causal). *A transfer matrix is said to be* $\mathcal{Z}$-nonsingular *if the determinant is in* $\mathcal{A}\backslash\mathcal{Z}$ *and to be* $\mathcal{Z}$-singular *otherwise.*

In [14], the ideal $\mathcal{Z}$ is not restricted to a prime ideal in general. On the other hand, in [11], the set of the denominators of causal transfer functions is a multiplicatively closed subset of $\mathcal{A}$. This property is natural since the multiplication of two causal transfer functions should be considered as causal one. Note that this multiplicativity is equivalent to $\mathcal{Z}$ being prime provided that $\mathcal{Z}$ is an ideal. By following the multiplicativity, we consider in this paper that $\mathcal{Z}$ is prime.

In this paper, we do not assume that the prime spectrum of $\mathcal{A}$ is irreducible and the plant P is strictly causal as in [11]. Alternatively, in the rest of the paper we assume only the following:

Assumption 2.4. The given plant is causal in the sense of Definition 2.3.

One can represent a causal plant P in the form of fractions $P = ND^{-1} = \widetilde{D}^{-1}\widetilde{N}$, where the matrices N, D, $\widetilde{N}$, $\widetilde{D}$ are over $\mathcal{A}$, and the matrices D, $\widetilde{D}$ are $\mathcal{Z}$-nonsingular.

It should be noted that when using "a stabilizing controller," we do not guarantee the causality. However, in the classical case of the factorization approach, once we restrict ourselves to strictly proper plants, it is known that any stabilizing controller of strictly causal plant is causal (cf. Corollary 5.2.20 of [13], Theorem 4.1 of [14]). One can see, in fact, that many practical systems are strictly causal. On the other hand, including noncausal stabilizing controllers seems to make the theory easy and simple from the mathematical viewpoint. From these observations, we have accepted the possibility of the noncausality of stabilizing controllers.

In our case, the fact "any stabilizing controller of strictly causal plant is causal" still holds (Proposition 6.2). Further we will show that for any causal plant there exists a causal stabilizing controller (Proposition 6.1).

2.3. Previous results. In [11] Sule gave the results of the feedback stabilizability. We show them after introducing the notion of the elementary factor which is used to state his results.

Definition 2.5 (elementary factors [11, p. 1689]). *Assume that* $\mathcal{A}$ *is a unique factorization domain. Denote by* T *the matrix* $[\, N^t \quad dE_m \,]^t$ *and by* W *the matrix* $[\, N \quad dE_n \,]$ *over* $\mathcal{A}$, *where* $P = Nd^{-1}$ *with* $N \in \mathcal{A}^{n \times m}$, $d \in \mathcal{A}$. *Let* $\{T_1, T_2, \ldots, T_r\}$ *be the family of all nonsingular* $m \times m$ *submatrices of the matrix* T, *and for each index* i, *let* f_i *be the radical of the least common multiple of all the denominators of* TT_i^{-1}. *The family* $F = \{f_1, f_2, \ldots, f_r\}$ *is called the family of* elementary factors of the matrix T. *Analogously let* $\{W_1, W_2, \ldots, W_r\}$ *be the family of all nonsingular*

$n \times n$ submatrices of the matrix W, and for each index j let g_j be the radical of the least common multiple of all the denominators of $W_j^{-1}W$. Let $G = \{g_1, g_2, \ldots, g_l\}$ denote the family of elementary factors of the transposed matrix W^t. *Now let $H = \{f_i g_j \,|\, i = 1, \ldots, r,\ j = 1, \ldots, l\}$. This family H is called the* elementary factor of the transfer matrix P.

Then, Sule's two elegant results can be rewritten as follows. The first result assumes that the prime spectrum of $\mathcal{A}$ is irreducible. The second one assumes that $\mathcal{A}$ is a unique factorization domain.

Theorem 2.6 (see [11, Theorem 1]). *Suppose that the prime spectrum of $\mathcal{A}$ is irreducible. Further suppose that a plant P of size $n \times m$ is strictly causal, where the notion of the strictly causal is defined as in* [12] *(rather than* [11]*). Then the plant P is stabilizable if and only if the following conditions are satisfied:*

(i) *The module $\mathcal{T}_P$ is projective of rank m.*

(ii) *The module $\mathcal{W}_P$ is projective of rank n.* □

Recall that for a matrix X over $\mathcal{F}$ we use the notations $\mathcal{T}_X$ and $\mathcal{W}_X$ to denote $\mathcal{A}$-modules generated by using the matrix X. Further it should be noted that the definitions of $\mathcal{T}_P$ and $\mathcal{W}_P$ in [11] are slightly different from those of this paper. Nevertheless this is not a problem by virtue of Lemma 2.1.

Theorem 2.7 (see [11, Theorem 4]). *Suppose that $\mathcal{A}$ is a unique factorization domain. The plant P is stabilizable if and only if the elementary factors of P are coprime, that is, $\sum_{h \in H}(h) = \mathcal{A}$.* □

3. Main results.

To state our results precisely we define the notion of generalized elementary factors, which is a generalization of the elementary factors in Definition 2.5. Then the main theorem will be presented.

Generalized elementary factors. Originally, the elementary factors have been defined over unique factorization domains as in Definition 2.5. We enlarge this concept in the case of commutative rings.

Before stating the definition, we introduce several symbols used in the definition and widely in the rest of this paper. The symbol $\mathcal{I}$ denotes the set of all sets of m distinct integers between 1 and $m+n$ (recall that m and n are the numbers of the inputs and the outputs, respectively). Normally, an element of $\mathcal{I}$ will be denoted by I, possibly with suffixes such as integers. We will use an element of $\mathcal{I}$ as a suffix as well as a set. For $I \in \mathcal{I}$, if $i_1, \ldots, i_m$ are elements of I in ascending order, that is, $i_a < i_b$ if $a < b$, then the symbol Δ_I denotes the $m \times (m+n)$ matrix whose (k, i_k)-entry is 1 for $i_k \in I$ and zero otherwise.

Definition 3.1 (generalized elementary factors). *Let $P \in \mathcal{F}^{n \times m}$, and N and D are matrices over $\mathcal{A}$ with $P = ND^{-1}$. Denote by T the matrix $[\,N^t \quad D^t\,]^t$. For each $I \in \mathcal{I}$, define the ideal Λ_{PI} of $\mathcal{A}$ by*

$$\Lambda_{PI} = \{\lambda \in \mathcal{A} \,|\, \exists K \in \mathcal{A}^{(m+n)\times m} \ \lambda T = K \Delta_I T\}.$$

We call the ideal Λ_{PI} the generalized elementary factor *of the plant P with respect to I.*

Whenever we use the symbol Λ with some suffix, it will denote a generalized elementary factor. We will also frequently use the symbols λ and λ_I with $I \in \mathcal{I}$ as particular elements of Λ_{PI}. Note that in Definitions 2.5 and 3.1, the matrices represented by T are different in general. However this difference is not a problem since the generalized elementary factors are independent of the choice of the fractions ND^{-1} as shown below.

Lemma 3.2. *For any $I \in \mathcal{I}$, the generalized elementary factor of the plant P with respect to I is independent of the choice of matrices N and D over $\mathcal{A}$ satisfying $P = ND^{-1}$.*

Proof. Let N, N', D be matrices over $\mathcal{A}$ and d' be a scalar of $\mathcal{A}$ such that $P = ND^{-1} = N'{d'}^{-1}$ hold. Further, let

$$\begin{aligned}\Lambda_{PI1} &= \{\lambda \in \mathcal{A} \,|\, \exists K \in \mathcal{A}^{(m+n)\times m}\ \lambda\,[\,N^t \quad D^t\,]^t = K\Delta_I\,[\,N^t \quad D^t\,]^t\},\\ \Lambda_{PI2} &= \{\lambda \in \mathcal{A} \,|\, \exists K \in \mathcal{A}^{(m+n)\times m}\ \lambda\,[\,{N'}^t \quad d'E_m\,]^t = K\Delta_I\,[\,{N'}^t \quad d'E_m\,]^t\}.\end{aligned}$$

In order to prove this lemma it is sufficient to show that the ideals Λ_{PI1} and Λ_{PI2} are equal. Suppose that λ is an element of Λ_{PI1}. Then there exists a matrix K such that $\lambda\,[\,N^t \quad D^t\,]^t = K\Delta_I\,[\,N^t \quad D^t\,]^t$. Multiplying $d'E_m$ on the right of both sides, we have $\lambda\,[\,{N'}^t \quad d'E_m\,]^t\,D = K\Delta_I\,[\,{N'}^t \quad d'E_m\,]^t\,D$. Since the matrix D is nonsingular, we have $\lambda\,[\,{N'}^t \quad d'E_m\,]^t = K\Delta_I\,[\,N^t \quad d'E_m\,]^t$, so that $\lambda \in \Lambda_{PI2}$, which means that $\Lambda_{PI1} \subset \Lambda_{PI2}$. The opposite inclusion relation $\Lambda_{PI1} \supset \Lambda_{PI2}$ can be proved analogously. □

Note also that for every I in $\mathcal{I}$, the generalized elementary factor of the plant with respect to I is not empty since it contains at least zero. In the case where the set of stable causal transfer functions is a unique factorization domain, the generalized elementary factor of the plant with the matrix $\Delta_I T$ being nonsingular becomes a principal ideal and the generator of its radical an elementary factor of the matrix T (in Definition 2.5) up to a unit multiple. Analogously, the elementary factor of the matrix W (in Definition 2.5) corresponds to the generalized elementary factor of the transposed plant P^t.

Main results. We are now in position to state our main results.

Theorem 3.3. *Consider a causal plant P. Then the following statements are equivalent:*

(i) *The plant P is stabilizable.*

(ii) *The $\mathcal{A}$-modules $\mathcal{T}_P$ and $\mathcal{W}_P$ are projective.*

(iii) *The set of all generalized elementary factors of P generates $\mathcal{A}$; that is,*

$$\sum_{I\in\mathcal{I}} \Lambda_{PI} = \mathcal{A}. \tag{3.1}$$

In the theorem, (ii) and (iii) are criteria for feedback stabilizability. Comparing the theorem above with Theorems 2.6 and 2.7, we observe the following: (ii) and (iii) can be considered as generalizations of Theorems 2.6 and 2.7, respectively. For (ii), we do not assume as mentioned earlier that the prime spectrum of $\mathcal{A}$ is irreducible and the plant P is strictly causal. The rank conditions of $\mathcal{T}_P$ and $\mathcal{W}_P$ are deleted. For (iii), the commutative ring $\mathcal{A}$ is not restricted to a unique factorization domain. The elementary factors are replaced by the generalized elementary factors. Although two matrices T and W in Definition 2.5 are used to state Theorem 2.7, only one matrix T in Definition 3.1 is used in (iii).

We will present the proof of the theorem in section 5.

To make the notion of the generalized elementary factors familiar, we present here an example of the generalized elementary factors.

Example 3.4. Some synchronous high-speed electronic circuits such as computer memory devices often cannot have nonzero small delays (see [5], for example). We suppose here that the system cannot have the unit delay as a nonzero small delay. Further we suppose that the impulse response of a transfer function being stable is finitely terminated. Thus the set $\mathcal{A}$ becomes the set of polynomials generated by z^2 and z^3, that is, $\mathcal{A} = \mathbb{R}[z^2, z^3]$, where z denotes the unit delay operator. Then $\mathcal{A}$ is not a unique factorization domain but a Noetherian domain. The total field $\mathcal{F}$ of fractions of $\mathcal{A}$ is $\mathbb{R}(z^2, z^3)$, which is equal to $\mathbb{R}(z)$. The ideal $\mathcal{Z}$ used to define the causality is given as the set of polynomials in $\mathbb{R}[z^2, z^3]$ whose constant terms are zero; that is, $\mathcal{Z} = z^2\mathcal{A} + z^3\mathcal{A} = \{az^2 + bz^3 \mid a, b \in \mathcal{A}\}$. Thus the set of causal transfer functions $\mathcal{P}$ is given as n/d, where n, d are in $\mathcal{A}$ and the constant term of d is nonzero; that is, $\mathcal{P} = \{n/(a + bz^2 + cz^3) \mid n \in \mathcal{A},\ \ a \in \mathbb{R}\backslash\{0\},\ \ b, c \in \mathcal{A}\}$. Further the set of strictly causal transfer functions $\mathcal{P}_{\rm S}$ is given as $\mathcal{P}_{\rm S} = \{(a_1z^2 + b_1z^3)/(a_2 + b_2z^2 + c_2z^3) \mid a_2 \in \mathbb{R}\backslash\{0\},\ \ a_1, b_1, b_2, c_2 \in \mathcal{A}\}$.

Since some factorized polynomials are sometimes expressed more compactly and are thus easier to understand than the expanded ones, we here introduce the following notation: a polynomial in $\mathbb{R}[z]$ surrounded by "$\langle$" and "$\rangle$" indicates that it is in $\mathcal{A}$ as well as in $\mathbb{R}[z]$ even though some factors between "$\langle$" and "$\rangle$" may not be in $\mathcal{A}$.

Let us consider the plant below:

$$P := \begin{bmatrix} (1-z^3)/(1-z^2) \\ (1-8z^3)/(1-4z^2) \end{bmatrix} \in \mathcal{P}^{2\times 1}. \tag{3.2}$$

The representation of the plant is not unique. For example, the (1,1)-entry of the plant has an alternative form $(1 + z^2 + z^4)/(1 + z^3)$ different from the expression in (3.2). Consider parameterizing the representation of the plant. To do so we consider the plant P over $\mathbb{R}(z)$ rather than over $\mathcal{F}$. Thus P can be expressed as

$$P = \begin{bmatrix} (1+z+z^2)/(1+z) \\ (1+2z+4z^2)/(1+2z) \end{bmatrix} \quad \text{over } \mathbb{R}(z). \tag{3.3}$$

However, the coefficients of all numerators and denominators in (3.3) of z with degree 1 are not zero. To make them zero, we should multiply them by $(a_1(1-z)+b_1z^2+c_1z^3)$ or $(a_2(1-2z)+b_2z^2+c_2z^3)$ with $a_1, b_1, c_1, a_2, b_2, c_2 \in \mathcal{A}$ as follows:

$$P = \begin{bmatrix} \frac{\langle(1+z+z^2)(a_1(1-z)+b_1z^2+c_1z^3)\rangle}{\langle(1+z)(a_1(1-z)+b_1z^2+c_1z^3)\rangle} \\ \frac{\langle(1+2z+4z^2)(a_2(1-2z)+b_2z^2+c_2z^3)\rangle}{\langle(1+2z)(a_2(1-2z)+b_2z^2+c_2z^3)\rangle} \end{bmatrix}. \tag{3.4}$$

Every expression of the plant is given in the form of (3.4) with $a_1, b_1, c_1, a_2, b_2, c_2$ in $\mathcal{A}$ provided that the denominators are not zero. From this, we have two observations. One is that the plant P does not have its right- and left-coprime factorizations over $\mathcal{A}$ (even so, it will be shown later that the plant is stabilizable). The other is that the elementary factor of this plant cannot be consistently defined over $\mathcal{A}$. Thus we employ the notion of the generalized elementary factor.

In the following, we calculate the generalized elementary factors of the plant. We choose the following matrices as N, D, and T used in Definition 3.1:

$$\begin{bmatrix} n_1 \\ n_2 \end{bmatrix} := N := \begin{bmatrix} (1-z^3)(1-4z^2) \\ (1-8z^3)(1-z^2) \end{bmatrix},$$
$$[\,d\,] := D := [\,(1-z^2)(1-4z^2)\,], \quad T := [\,N^t \quad D^t\,]^t.$$

Since $m = 1$ (the number of inputs) and $n = 2$ (the number of outputs), the set $\mathcal{I}$ is given as $\mathcal{I} = \{\{1\}, \{2\}, \{3\}\}$ and we let $I_1 = \{1\}$, $I_2 = \{2\}$, $I_3 = \{3\}$.

Let us calculate the generalized elementary factor Λ_{PI_1}. Let $i_1 = 1$ so that $I_1 = \{i_1\}$. Then the $(1, i_1)$-entry of the matrix Δ_{I_1} is 1 and the other entries are zero.

Thus we have $\Delta_{I_1} = [\,1 \quad 0 \quad 0\,]$. The generalized elementary factor Λ_{PI_1} is originally given as follows:

$$\begin{aligned} \Lambda_{PI_1} &= \{\lambda \in \mathcal{A} \,|\, \exists K \in \mathcal{A}^{(m+n)\times m} \;\; \lambda T = K \Delta_{I_1} T\} \\ &= \{\lambda \in \mathcal{A} \,|\, \exists k_1, k_2 \in \mathcal{A} \;\; \lambda\,[\,n_2 \quad d\,]^t = n_1\,[\,k_1 \quad k_2\,]^t\}. \end{aligned} \tag{3.5}$$

Consider (3.5) over $\mathbb{R}[z]$ instead of $\mathcal{A}$. Then the matrix equation in the set of (3.5) can be expressed as

$$\begin{aligned} \lambda \begin{bmatrix} (1-z)(1+z)(1-2z)(1+2z+4z^2) \\ (1-z)(1+z)(1-2z)(1+2z) \end{bmatrix} &= \\ (1-z)(1-2z)(1+2z)(1+z+z^2) \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}&. \end{aligned} \tag{3.6}$$

The set of λ's such that there exist $k_1, k_2 \in \mathbb{R}[z]$ satisfying (3.6) is given as $\{(1+2z)(1+z+z^2)a \mid a \in \mathbb{R}[z]\}$, denoted by L_1. Then the intersection of L_1 and $\mathcal{A}$ is given as follows:

$$L_1 \cap \mathcal{A} = \{\langle (1+2z)(1+z+z^2)(a(1-3z)+bz^2+cz^3)\rangle \in \mathcal{A} \mid a,b,c \in \mathcal{A}\}. \tag{3.7}$$

This is equal to Λ_{PI_1} as shown below. First it is obvious that $L_1 \cap \mathcal{A} \supset \Lambda_{PI_1}$. For each $(1+2z)(1+z+z^2)(a(1-3z)+bz^2+cz^3)$ with $a,b,c \in \mathcal{A}$, we have k_1 and k_2 as follows from (3.6):

$$\begin{aligned} k_1 &= (1+z)(1+2z+4z^2)(a(1-3z)+bz^2+cz^3), \\ k_2 &= (1+z)(1+2z)(a(1-3z)+bz^2+cz^3). \end{aligned}$$

Both k_1 and k_2 are in $\mathcal{A}$. Hence $L_1 \cap \mathcal{A} \subset \Lambda_{PI_1}$ and so $L_1 \cap \mathcal{A} = \Lambda_{PI_1}$. By (3.7), we can also consider that Λ_{PI_1} is generated by $\langle (1+2z)(1+z+z^2)(1-3z)\rangle$, $\langle (1+2z)(1+z+z^2)z^2\rangle$, and $\langle (1+2z)(1+z+z^2)z^3\rangle$.

Analogously, we can calculate the generalized elementary factors Λ_{PI_2} and Λ_{PI_3} of the plant with respect to I_2 and I_3 as follows:

$$\begin{aligned} \Lambda_{PI_2} &= \{\langle (1+z)(1+2z+4z^2)(a(1-3z)+bz^2+cz^3)\rangle \in \mathcal{A} \mid a,b,c \in \mathcal{A}\}, \\ \Lambda_{PI_3} &= \{\langle (1+z)(1+2z)(a(1-3z)+bz^2+cz^3)\rangle \in \mathcal{A} \mid a,b,c \in \mathcal{A}\}. \end{aligned}$$

Observe now that

$$\begin{aligned} \Lambda_{PI_1} &\ni \langle (1+2z)(1-3z)(1+z+z^2)\rangle =: \lambda_{0I_1}, \\ \Lambda_{PI_2} &\ni \langle (1+z)(1+2z+4z^2)(1-3z+z^2)\rangle =: \lambda_{0I_2} \end{aligned}$$

and further

$$\alpha_{I_1}\lambda_{0I_1} + \alpha_{I_2}\lambda_{0I_2} = 1,$$

where

$$\begin{aligned} \alpha_{I_1} &= \tfrac{-4233-23646z^2-39836z^3-201780z^4-113016z^5+75344z^6}{5852} \in \mathcal{A}, \\ \alpha_{I_2} &= \tfrac{10085+18418z^2+121140z^3+131852z^4+113016z^5}{5852} \in \mathcal{A}. \end{aligned}$$

Now let

$$\lambda_{I_1} := \alpha_{I_1}\lambda_{0I_1} \in \Lambda_{PI_1}, \quad \lambda_{I_2} := \alpha_{I_2}\lambda_{0I_2} \in \Lambda_{PI_2}. \tag{3.8}$$

Thus $\Lambda_{PI_1} + \Lambda_{PI_2} = \mathcal{A}$ and $\lambda_{I_1} + \lambda_{I_2} = 1$. Hence by Theorem 3.3, the plant P is stabilizable. $\square$

4. Intermediate results. In this section we provide intermediate results which will be used in the proof of our main theorem stated in the preceding section. This section consists of three parts. We first show that a number of modules generated from plants, controllers, and feedback systems are isomorphic to one another. Next we develop the results which will help to show the existence of a well-defined stabilizing controller. We then give the coprime factorizability of the plant over $\mathcal{A}_f$, where f is an element of the generalized elementary factor of the plant.

Relationship in terms of modules between transfer matrices P, C, and $H(P,C)$. The first intermediate result is the relations, expressed in terms of modules, among the matrices P, C, and $H(P,C)$ as well as their transposed matrices. A number of modules are isomorphic to one another as follows.

PROPOSITION 4.1. *Suppose that P and C are matrices over $\mathcal{F}(\mathcal{R})$. Suppose further that $\det(E_n + PC)$ is a unit of $\mathcal{F}(\mathcal{R})$.*

(i) *The following $\mathcal{R}$-modules are isomorphic to one another:* $\mathcal{T}_{P,\mathcal{R}} \oplus \mathcal{T}_{C,\mathcal{R}}$, $\mathcal{T}_{H(P,C),\mathcal{R}}$, $\mathcal{T}_{H(P^t,C^t)^t,\mathcal{R}}$, $\mathcal{W}_{H(P^t,C^t),\mathcal{R}}$, $\mathcal{T}_{H(C,P),\mathcal{R}}$.

(ii) *The following $\mathcal{R}$-modules are isomorphic to one another:* $\mathcal{W}_{P,\mathcal{R}} \oplus \mathcal{W}_{C,\mathcal{R}}$, $\mathcal{W}_{H(P,C),\mathcal{R}}$, $\mathcal{W}_{H(P^t,C^t)^t,\mathcal{R}}$, $\mathcal{T}_{H(P^t,C^t),\mathcal{R}}$, $\mathcal{W}_{H(C,P),\mathcal{R}}$.

Further for a matrix X over $\mathcal{F}(\mathcal{R})$,

(iii) $\mathcal{T}_{X,\mathcal{R}} \simeq \mathcal{W}_{X^t,\mathcal{R}}$ *and* $\mathcal{W}_{X,\mathcal{R}} \simeq \mathcal{T}_{X^t,\mathcal{R}}$.

Note here that in the proposition above, the controller C need not be a stabilizing controller. For the case where C is a stabilizing controller, see Lemma 2 of [11].

We can consider that the proposition above, especially the relations $\mathcal{T}_{P,\mathcal{R}} \oplus \mathcal{T}_{C,\mathcal{R}} \simeq \mathcal{T}_{H(P,C),\mathcal{R}} \simeq \mathcal{T}_{H(C,P),\mathcal{R}}$, gives an interpretation of the structure of the feedback system in the sense that the module generated by the feedback system is given as the direct sum of the modules generated by the plant and the controller. In the proof ((i)→(ii)) of Theorem 3.3, this proposition will play a key role.

Proof. We first prove (iii). Let A and B be matrices over $\mathcal{R}$ with $X = AB^{-1}$. Then we have $\mathcal{T}_{X,\mathcal{R}} \simeq M_r([\,A^t \quad B^t\,]^t) \simeq M_c([\,A^t \quad B^t\,]) \simeq \mathcal{W}_{(B^{-1})^t A^t,\mathcal{R}} \simeq \mathcal{W}_{X^t,\mathcal{R}}$. The other relation $\mathcal{W}_{X,\mathcal{R}} \simeq \mathcal{T}_{X^t,\mathcal{R}}$ can be proved in a similar way.

Next we prove (i). Suppose that $\det(E_n + PC)$ is a unit of $\mathcal{F}(\mathcal{R})$. We prove the following relations in order: (a) $\mathcal{T}_{P,\mathcal{R}} \oplus \mathcal{T}_{C,\mathcal{R}} \simeq \mathcal{T}_{H(P,C),\mathcal{R}}$, (b) $\mathcal{T}_{H(P,C),\mathcal{R}} \simeq \mathcal{T}_{H(P^t,C^t)^t,\mathcal{R}}$, (c) $\mathcal{T}_{H(P^t,C^t)^t,\mathcal{R}} \simeq \mathcal{W}_{H(P^t,C^t),\mathcal{R}}$, (d) $\mathcal{T}_{H(P,C),\mathcal{R}} \simeq \mathcal{T}_{H(C,P),\mathcal{R}}$.

(a) of (i). The proof of (a) follows mainly the proof of Lemma 2 in [11]. By virtue of Lemma 2.1, it is sufficient to show the relation $\mathcal{T}_{P,\mathcal{R}} \oplus \mathcal{T}_{C,\mathcal{R}} \simeq M_r([\,N_H^t \quad d_H E_{m+n}\,]^t)$ with $N_H \in (\mathcal{R})_{m+n}$, $d_H \in \mathcal{R}$, where $H(P,C) = N_H d_H^{-1}$. Let N, N_C be matrices over $\mathcal{R}$ and d, d_C be in $\mathcal{R}$ with $P = Nd^{-1}$ and $C = N_C d_C^{-1}$. Further, let

$$Q = \begin{bmatrix} d_C E_n & N \\ -N_C & dE_m \end{bmatrix}, \quad S = \begin{bmatrix} d_C E_n & O \\ O & dE_m \end{bmatrix}.$$

From these we have $\mathcal{T}_{P,\mathcal{R}} \oplus \mathcal{T}_{C,\mathcal{R}} \simeq M_r([\,Q^t \quad S^t\,]^t)$. In addition, since $\det(E_n + PC)$ is a unit of $\mathcal{F}(\mathcal{R})$, the matrix N_H is nonsingular so that $M_r([\,Q^t \quad S^t\,]^t) \simeq M_r([\,Q^t \quad S^t\,]^t (\det(N_H)E_{m+n}))$ holds. A simple calculation shows that

$$\begin{bmatrix} Q \\ S \end{bmatrix} (\det(N_H)E_{m+n}) = \begin{bmatrix} d_H E_{m+n} \\ N_H \end{bmatrix} \operatorname{adj}(N_H) S.$$

Because both matrices S and $\operatorname{adj}(N_H)$ are nonsingular, we finally have that

$$\begin{aligned}\mathcal{T}_{P,\mathcal{R}} \oplus \mathcal{T}_{C,\mathcal{R}} &\simeq M_r\left(\begin{bmatrix} Q \\ S \end{bmatrix}\right) \simeq M_r\left(\begin{bmatrix} Q \\ S \end{bmatrix}(\det(N_H)E_{m+n})\right) \\ &\simeq M_r\left(\begin{bmatrix} d_H E_{m+n} \\ N_H \end{bmatrix}\right) \simeq \mathcal{T}_{H(P,C),\mathcal{R}}.\end{aligned}$$

(b) of (i). Observe that the following relation holds:

$$H(P^t, C^t)^t = \begin{bmatrix} O & E_m \\ E_n & O \end{bmatrix} H(P,C) \begin{bmatrix} O & E_n \\ E_m & O \end{bmatrix}. \tag{4.1}$$

Let N_H and d_H be a matrix over $\mathcal{R}$ and a scalar of $\mathcal{R}$, respectively, with $H(P,C) = N_H d_H^{-1}$. Then (4.1) can be rewritten as follows:

$$H(P^t, C^t)^t = \begin{bmatrix} O & E_m \\ E_n & O \end{bmatrix} N_H \left(\begin{bmatrix} O & E_m \\ E_n & O \end{bmatrix} d_H\right)^{-1}.$$

Hence, we have matrices A and B over $\mathcal{R}$ with $H(P^t, C^t)^t = AB^{-1}$ such that

$$A = \begin{bmatrix} O & E_m \\ E_n & O \end{bmatrix} N_H, \quad B = \begin{bmatrix} O & E_m \\ E_n & O \end{bmatrix} d_H.$$

This gives the relation $\mathcal{T}_{H(P,C),\mathcal{R}} \simeq \mathcal{T}_{H(P^t,C^t)^t,\mathcal{R}}$.

(c) of (i). This is directly obtained by applying (iii) to the matrix $H(P^t, C^t)^t$.

(d) of (i). Between the matrices $H(P,C)$ and $H(C,P)$, the following relation holds:

$$H(C,P) = \begin{bmatrix} O & -E_m \\ E_n & O \end{bmatrix} H(P,C) \begin{bmatrix} O & E_n \\ -E_m & O \end{bmatrix}.$$

Letting N_H and d_H be a matrix over $\mathcal{R}$ and a scalar of $\mathcal{R}$ with $H(P,C) = N_H d_H^{-1}$ as in (b), we have matrices N_H' and D_H' over $\mathcal{R}$ such that

$$\begin{bmatrix} N_H' \\ D_H' \end{bmatrix} = \begin{bmatrix} \begin{matrix} O & -E_m \\ E_n & O \end{matrix} & O \\ O & \begin{matrix} O & -E_m \\ E_n & O \end{matrix} \end{bmatrix} \begin{bmatrix} N_H \\ d_H E_{m+n} \end{bmatrix}$$

holds. Since $H(C,P) = N_H' {D_H'}^{-1}$ holds and the first matrix of the right-hand side of the equation above is invertible, the relation $\mathcal{T}_{H(P,C),\mathcal{R}} \simeq \mathcal{T}_{H(C,P),\mathcal{R}}$ holds.

Finally, arguments similar to (i) prove (ii). □

Before moving to the next intermediate result, we prove an easy lemma useful to give results for the transposed plants.

LEMMA 4.2. *A plant P is stabilizable if and only if its transposed plant P^t is stabilizable. Moreover, in the case where the plant P is stabilizable, C is a stabilizing controller of P if and only if C^t is a stabilizing controller of the transposed plant P^t.*

Proof. (Only If) Suppose that a plant P is stabilizable. Let C be a stabilizing controller of P. First, (P^t, C^t) is in $\widehat{F}_{\mathrm{ad}}$, since $(P,C) \in \widehat{F}_{\mathrm{ad}}$ and $\det(E_n + PC) = \det(E_m + P^t C^t)$. From (4.1) in the proof of Proposition 4.1, if $H(P,C) \in (\mathcal{A})_{m+n}$, then $H(P^t, C^t) \in (\mathcal{A})_{m+n}$.

(If) Because $(P^t)^t = P$, the "If" part can be proved analogously. □

$\mathcal{Z}$-nonsingularity of transfer matrices. In order to prove the stabilizability of the given causal plant, which will be necessary in the proof of the main theorem (Theorem 3.3), we should show the existence of the stabilizing controller. To do so, we will need to show that the denominator matrix of the stabilizing controller is nonsingular. The following result will help.

LEMMA 4.3. *Suppose that there exist matrices A, B, C_1, C_2 over $\mathcal{A}$ such that the following square matrix is $\mathcal{Z}$-nonsingular:*

$$\begin{bmatrix} A & C_1 \\ B & C_2 \end{bmatrix}, \tag{4.2}$$

where the matrix A is square and the matrices A and B have same number of columns. Then there exists a matrix R over $\mathcal{A}$ such that the matrix $A + RB$ is $\mathcal{Z}$-nonsingular.

Before starting the proof, it is worth reviewing some easy facts about the prime ideal $\mathcal{Z}$.

Remark 4.4. (i) If a is in $\mathcal{A}\backslash\mathcal{Z}$ and expressed as $a = b + c$ with $b, c \in \mathcal{A}$, then at least one of b and c is in $\mathcal{A}\backslash\mathcal{Z}$. (ii) If a is in $\mathcal{A}\backslash\mathcal{Z}$ and b is in $\mathcal{Z}$, then the sum $a + b$ is in $\mathcal{A}\backslash\mathcal{Z}$. (iii) Every factor in $\mathcal{A}$ of an element of $\mathcal{A}\backslash\mathcal{Z}$ belongs to $\mathcal{A}\backslash\mathcal{Z}$ (that is, if $a, b \in \mathcal{A}$ and $ab \in \mathcal{A}\backslash\mathcal{Z}$, then $a, b \in \mathcal{A}\backslash\mathcal{Z}$).

They will be used in the proofs of Lemma 4.3 and Theorem 3.3.

Proof of Lemma 4.3. This proof mainly follows that of Lemma 4.4.21 of [13].

If the matrix A itself is $\mathcal{Z}$-nonsingular, then we can select the zero matrix as R. Hence we assume in the following that A is $\mathcal{Z}$-singular.

Since (4.2) is $\mathcal{Z}$-nonsingular, there exists a full-size minor of $[\,A^t \quad B^t\,]^t$ in $\mathcal{A}\backslash\mathcal{Z}$ by Laplace's expansion of (4.2) and Remark 4.4(i), (iii). Let a be such a $\mathcal{Z}$-nonsingular full-size minor of $[\,A^t \quad B^t\,]^t$ having as few rows from B as possible.

We here construct a matrix R such that $\det(A + RB) = \pm a + z$ with $z \in \mathcal{Z}$. Since A is $\mathcal{Z}$-singular, the full-size minor a must contain at least one row of B from the matrix $[\,A^t \quad B^t\,]^t$. Suppose that a is obtained by excluding the rows $i_1, \dots, i_k$ of A and including the rows $j_1, \dots, j_k$ of B, where both $i_1, \dots, i_k$ and $j_1, \dots, j_k$ are in ascending order. Now define $R = (r_{ij})$ by $r_{i_1 j_1} = \cdots = r_{i_k j_k} = 1$ and $r_{ij} = 0$ for all other i, j. Observe that $\det(A + RB)$ is expanded in terms of full-size minors of the matrices $[\,E \quad R\,]$ and $[\,A^t \quad B^t\,]^t$ from the factorization $A + RB = [\,E \quad R\,][\,A^t \quad B^t\,]^t$ by the Binet–Cauchy formula. Every minor of $[\,E \quad R\,]$ containing more than k columns of R is zero. By the method of choosing the rows from $[\,A^t \quad B^t\,]^t$ for the full-size minor a, every full-size minor of $[\,A^t \quad B^t\,]^t$ having less than k rows of B is in $\mathcal{Z}$. There is only one nonzero minor of $[\,E \quad R\,]$ containing exactly k columns of R, which is obtained by excluding the columns $i_1, \dots, i_k$ of the identity matrix E and including the columns $j_1, \dots, j_k$ of R; it is equal to ± 1. From the Binet–Cauchy formula, the corresponding minor of $[\,A^t \quad B^t\,]^t$ is a. As a result, $\det(A + RB)$ is given as a sum of $\pm a$ and elements in $\mathcal{Z}$. By Remark 4.4(ii), the sum is in $\mathcal{A}\backslash\mathcal{Z}$ and so is $\det(A + RB)$. The matrix $A + RB$ is now $\mathcal{Z}$-nonsingular. □

Coprimeness and generalized elementary factors. We present here that for each nonnilpotent element λ of the generalized elementary factors, the plant has a right-coprime factorization over $\mathcal{A}_\lambda$. This will be independent of the stabilizability of the plant.

LEMMA 4.5 (cf. Proposition 2.2 of [9]). *Let Λ_{PI} be the generalized elementary factor of the plant P with respect to $I \in \mathcal{I}$ and further let $\sqrt{\Lambda_{PI}}$ denote the radical of Λ_{PI} (as an ideal). Suppose that λ is in $\sqrt{\Lambda_{PI}}$ but not nilpotent. Then, the $\mathcal{A}_\lambda$-module $\mathcal{T}_{P,\mathcal{A}_\lambda}$ is free of rank m.*

Proof. Let T, N, D be the matrices over $\mathcal{A}$ as in Definition 3.1. Recall that $\mathcal{T}_{P,\mathcal{A}_\lambda}$ denotes the $\mathcal{A}_\lambda$-module generated by rows of the matrix T. By Definition 3.1, there exists a matrix K over $\mathcal{A}$ such that $\lambda^r T = K\Delta_I T$ holds for some positive integer r. Then we have a factorization of the matrix T over $\mathcal{A}_\lambda$ as $T = (\lambda^{-r}K)(\Delta_I T)$, where all entries of the matrix $\lambda^{-r}K$ are in $\mathcal{A}_\lambda$. In order to show that the $\mathcal{A}_\lambda$-module $\mathcal{T}_{P,\mathcal{A}_\lambda}$ is free of rank m, provided that λ is not nilpotent, it is sufficient to prove the following two facts: (i) The matrix $\Delta_I T$ is nonsingular over $\mathcal{A}_\lambda$. (ii) There is a matrix X such that the matrix $[\,\lambda^{-r}K \quad X\,]$ is invertible over $\mathcal{A}_\lambda$ and the matrix equation $T = [\,\lambda^{-r}K \quad X\,]\,[\,(\Delta_I T)^t \quad O\,]^t$ holds.

(i). Observe that the matrix D is nonsingular over $\mathcal{A}_\lambda$ as well as over $\mathcal{A}$. Since $D = \Delta_{\{n+1,\ldots,m+n\}}T = (\lambda^{-r}\Delta_{\{n+1,\ldots,m+n\}}K)(\Delta_I T)$ holds (note that the suffix of the symbol Δ is an ordered set of m distinct integers between 1 and $m+n$ as before Definition 3.1), the matrix $\Delta_I T$ is also nonsingular over $\mathcal{A}_\lambda$ provided that λ is not nilpotent.

(ii). Let $\overline{i_1}, \ldots, \overline{i_n}$ be n distinct integers in ascending order between 1 and $m+n$ excluding the integers in I. Then let X be the matrix whose $(\overline{i_k}, k)$-entry is 1 for each $\overline{i_k}$ and zero otherwise. Then the determinant of $[\,\lambda^{-r}K \quad X\,]$ becomes ±1 since the matrix $\lambda^{-r}\Delta_I K$ is the identity matrix of $(\mathcal{A}_\lambda)_m$. □

The lemma below will be used in the proof ((ii)→(iii)) of the main theorem by letting $\mathcal{R} = \mathcal{A}_f$, where f is an element of the generalized elementary factor of the plant but not nilpotent.

LEMMA 4.6. *If $\mathcal{R}$-module $\mathcal{T}_{P,\mathcal{R}}$ $(\mathcal{W}_{P,\mathcal{R}})$ is free of rank m (n), there exist matrices A and B ($\widetilde{A}$ and $\widetilde{B}$) over $\mathcal{R}$ such that (A, B) is a right-coprime factorization ($(\widetilde{A}, \widetilde{B})$ is a left-coprime factorization) of the plant P $(\in \mathcal{F}(\mathcal{R})^{n\times m})$ over $\mathcal{R}$.*

Proof. This lemma is an analogy of the result given in the proof of Lemma 3 of [11]. See this proof. □

Example 4.7. Here we continue Example 3.4. Let us follow Lemmas 4.5 and 4.6 with the plant of (3.2). Let the notation be as in Example 3.4.

First we proceed along the proof of Lemma 4.5. As an example, we pick $I_1 \in \mathcal{I}$ as I and $\lambda_{I_1} \in \Lambda_{PI_1}$ as λ in the proof of Lemma 4.5. Recall that for each $\lambda \in \Lambda_{PI}$, there exists a matrix K such that $\lambda T = K\Delta_I T$ holds. In the case of $\lambda_{I_1} \in \Lambda_{PI_1}$, the matrix K is given as

$$K = \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} = \begin{bmatrix} \lambda_{I_1} \\ \alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle \\ \alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle \end{bmatrix}. \tag{4.3}$$

Thus we have the factorization $T = (\lambda^{-r}K)(\Delta_{I_1}T)$:

$$\begin{bmatrix} (1-z^3)(1-4z^2) \\ (1-8z^3)(1-z^2) \\ (1-z^2)(1-4z^2) \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle \\ \lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle \end{bmatrix} [\,(1-z^3)(1-4z^2)\,],$$

where $r = 1$ and $\Delta_{I_1}T = [\,(1-z^3)(1-4z^2)\,]$. As shown in part (i) of the proof of Lemma 4.5, $\Delta_{I_1}T = [\,(1-z^3)(1-4z^2)\,]$ is nonsingular.

The matrix X in part (ii) of the proof of Lemma 4.5 is given as $X = \left[\begin{smallmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right]^t$ by letting $\overline{i}_1 = 2$ and $\overline{i}_2 = 3$ according to $I_1 = \{1\}$. We can see that the matrix

$$[\,\lambda^{-1}K \quad X\,] = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle & 1 & 0 \\ \lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle & 0 & 1 \end{bmatrix} \tag{4.4}$$

is invertible. Therefore the $\mathcal{A}_{\lambda_{I_1}}$-module $\mathcal{T}_{P,\mathcal{A}_{\lambda_{I_1}}}$ is free and its rank is 1. (However, we will see that the $\mathcal{A}$-module $\mathcal{T}_P$ is not free. See Example 5.2.)

From (4.4) and the matrix equation $T = [\,\lambda^{-r}K \quad X\,]\,[\,(\Delta_{I_1}T)^t \quad O\,]^t$, we let

$$\begin{bmatrix} N_{I_1} \\ D_{I_1} \end{bmatrix} := \begin{bmatrix} 1 \\ \lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle \\ \lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle \end{bmatrix} = \lambda_{I_1}^{-1}K, \tag{4.5}$$

$$\begin{aligned} \begin{bmatrix} \widetilde{Y}_{I_1} & \widetilde{X}_{I_1} \\ \times & \times \end{bmatrix} &:= \begin{bmatrix} 1 & 0 & 0 \\ -\lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle & 1 & 0 \\ -\lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle & 0 & 1 \end{bmatrix}, \\ &= [\,\lambda_{I_1}^{-1}K \quad X\,]^{-1}, \end{aligned} \tag{4.6}$$

where $N_{I_1} \in \mathcal{A}_{\lambda_{I_1}}^{2\times 1}$, $\widetilde{Y}_{I_1} \in \mathcal{A}_{\lambda_{I_1}}^{1\times 2}$, $D_{I_1}, \widetilde{X}_{I_1} \in (\mathcal{A}_{\lambda_{I_1}})_1$, and × denotes some matrix. Then (N_{I_1}, D_{I_1}) is a right-coprime factorization of the plant over $\mathcal{A}_{\lambda_{I_1}}$ with $\widetilde{Y}_{I_1}N_{I_1} + \widetilde{X}_{I_1}D_{I_1} = E_1$, which is consistent with Lemma 4.6.

5. Proof of main results.

Now we give the proof of the main theorem.

Proof of Theorem 3.3. We prove the following relations in order: (a) "(i)→(ii)," (b) "(ii)→(iii)," and (c) "(iii)→(i)."

(a) "(i)→(ii)": Suppose that C is a stabilizing controller of the plant P. Then, the $\mathcal{A}$-module $\mathcal{T}_{H(P,C)}$ is obviously free. By the relation $\mathcal{T}_{P,\mathcal{R}} \oplus \mathcal{T}_{C,\mathcal{R}} \simeq \mathcal{T}_{H(P,C),\mathcal{R}}$ in Proposition 4.1(i), we have that the $\mathcal{A}$-module $\mathcal{T}_P$ is projective. By using Proposition 4.1(iii) and Lemma 4.2, the projectivity of the $\mathcal{A}$-module $\mathcal{W}_P$ can be proved analogously.

(b) "(ii)→(iii)": Suppose that (ii) holds, that is, the modules $\mathcal{T}_P$ and $\mathcal{W}_P$ are projective. We let T, N, D be the matrices over $\mathcal{A}$ as in Definition 3.1. According to Theorem IV.32 of [7, p. 295], there exist finite sets F_1 and F_2 such that (1) each of them generates $\mathcal{A}$, and (2) for any $f \in F_1$ ($f \in F_2$) the $\mathcal{A}_f$-module $\mathcal{T}_{P,\mathcal{A}_f}$ ($\mathcal{W}_{P,\mathcal{A}_f}$) is free. Let F be the set of all $f_1 f_2$'s with $f_1 \in F_1$ and $f_2 \in F_2$. Then F generates $\mathcal{A}$ again, and the $\mathcal{A}_f$-modules $\mathcal{T}_{P,\mathcal{A}_f}$ and $\mathcal{W}_{P,\mathcal{A}_f}$ are free for every $f \in F$. We suppose without loss of generality that the sets F_1, F_2, and F do not contain any nilpotent element because $1 + x$ is a unit of $\mathcal{A}$ for any nilpotent x (cf. [2, p. 10]). (However, we note that other zerodivisors cannot be excluded from the set F.) The rank of the free $\mathcal{A}_f$-module $\mathcal{T}_{P,\mathcal{A}_f}$ is m, since m rows of the denominator matrix D are independent over $\mathcal{A}_f$ as well as over $\mathcal{A}$. Analogously the rank of $\mathcal{W}_{P,\mathcal{A}_f}$ is n.

In order to show that (iii) holds, it suffices to show that the relation $\sum_{f\in F}(f^\xi) \subset \sum_{I\in\mathcal{I}} \Lambda_{PI}$ holds for a sufficiently large integer ξ. Once this relation is obtained, since $\sum_{f\in F}(f^\xi) = \mathcal{A}$ holds, we have $\sum_{I\in\mathcal{I}} \Lambda_{PI} = \mathcal{A}$.

Let f be an arbitrary but fixed element of F. Let V_f be a square matrix of size m whose rows are m distinct generators of the $\mathcal{A}_f$-module $M_r([\,N^t \quad D^t\,]^t)$ ($\simeq \mathcal{T}_{P,\mathcal{A}_f}$). We assume without loss of generality that V_f is over $\mathcal{A}$, that is, the denominators of all entries of V_f are 1. Otherwise if V_f is over $\mathcal{A}_f$ but not over $\mathcal{A}$, V_f multiplied by f^x, with a sufficiently large integer x, will be over $\mathcal{A}$, so that we can consider such $V_f f^x$ as "V_f." Thus the following matrix equation holds over $\mathcal{A}$:

$$f^\nu T = K_f V_f \tag{5.1}$$

with a nonnegative integer ν and a matrix $K_f \in \mathcal{A}^{(m+n)\times m}$.

In order to prove the relation $\sum_{f\in F}(f^{\xi}) \subset \sum_{I\in\mathcal{I}} \Lambda_{PI}$, we will first show the relation

$$I_{m\mathcal{A}}(f^{\nu}K_f) \subset \sum_{I\in\mathcal{I}} \Lambda_{PI} \tag{5.2}$$

and then

$$(f^{\xi}) \subset I_{m\mathcal{A}}(f^{\nu}K_f). \tag{5.3}$$

Observe first that $\det(f^{\nu}\Delta_I K_f) \in \Lambda_{PI}$ because

$$\det(f^{\nu}\Delta_I K_f)T = f^{m\nu}K_f \operatorname{adj}(\Delta_I K_f)\Delta_I T.$$

Since every element of $I_{m\mathcal{A}}(f^{\nu}K_f)$ is an $\mathcal{A}$-linear combination of $\det(f^{\nu}\Delta_I K_f)$'s for all $I \in \mathcal{I}$, we have (5.2).

We next present (5.3). Let N_0 and D_0 be matrices with $K_f = [\,N_0^t \quad D_0^t\,]^t$. Since each row of V_f is generated by rows of $[\,N^t \quad D^t\,]^t$ over $\mathcal{A}_f$, there exist matrices $\widetilde{Y}_0$ and $\widetilde{X}_0$ over $\mathcal{A}_f$ such that $V_f = \left[\,\widetilde{Y}_0 f^{\nu} \quad \widetilde{X}_0 f^{\nu}\,\right][\,N^t \quad D^t\,]^t$. Thus, since V_f is nonsingular over $\mathcal{A}_f$, we have $\left[\,\widetilde{Y}_0 \quad \widetilde{X}_0\,\right][\,N_0^t \quad D_0^t\,]^t = E_m$, which implies that (N_0, D_0) is a right-coprime factorization of the plant P over $\mathcal{A}_f$. Recall here that $\mathcal{W}_{P,\mathcal{A}_f}$ is free of rank n. Thus by Lemma 4.6 there exist matrices $\widetilde{N}_0$ and $\widetilde{D}_0$ such that $(\widetilde{N}_0, \widetilde{D}_0)$ is a left-coprime factorization of the plant P over $\mathcal{A}_f$. Let Y_0 and X_0 be matrices over $\mathcal{A}_f$ such that $\widetilde{N}_0 Y_0 + \widetilde{D}_0 X_0 = E_n$ holds. Then we have the following matrix equation:

$$\begin{bmatrix} \widetilde{Y}_0 & \widetilde{X}_0 \\ -\widetilde{D}_0 & \widetilde{N}_0 \end{bmatrix} \begin{bmatrix} N_0 & -X_0 \\ D_0 & Y_0 \end{bmatrix} = \begin{bmatrix} E_m & -\widetilde{Y}_0 X_0 + \widetilde{X}_0 Y_0 \\ O & E_n \end{bmatrix}. \tag{5.4}$$

Denote by R the matrix $[\,-X_0^t \quad Y_0^t\,]^t$. Then the matrix $[\,K_f \quad R\,]$ is invertible over $\mathcal{A}_f$ since the right-hand side of (5.4) is invertible. For each $I \in \mathcal{I}$, let $\overline{I}$ be the ordered set of n distinct integers between 1 and $m+n$ excluding m integers in I and let $\overline{i_1}, \ldots, \overline{i_n}$ be elements of $\overline{I}$ in ascending order. Let $\Delta_{\overline{I}} \in \mathcal{A}^{m\times(m+n)}$ denote the matrix whose $(k, \overline{i_k})$-entry is 1 for $\overline{i_k} \in \overline{I}$ and zero otherwise. Then, by using Laplace's expansion, the following holds:

$$\det([\,K_f \quad R\,]) = \sum_{I\in\mathcal{I}} (\pm \det(\Delta_I K_f)\det(\Delta_{\overline{I}} R)),$$

which is a unit of $\mathcal{A}_f$. From this and since the ideal $I_{m\mathcal{A}_f}(K_f)$ is generated by $\det(\Delta_I K_f)$'s for all $I \in \mathcal{I}$, we have $I_{m\mathcal{A}_f}(K_f) = \mathcal{A}_f$. This equality over $\mathcal{A}_f$ gives (5.3) for a sufficiently large integer ξ.

From (5.2) and (5.3), the relation $\sum_{f\in F}(f^{\xi}) \subset \sum_{I\in\mathcal{I}} \Lambda_{PI}$ holds. Therefore we conclude that the relation $\sum_{I\in\mathcal{I}} \Lambda_{PI} = \mathcal{A}$ holds.

(c) "(iii)→(i)": To prove the stabilizability, we will construct a stabilizing controller of the causal plant P from right-coprime factorizations over $\mathcal{A}_f$ with some f's in $\mathcal{A}$. Let N and D be matrices over $\mathcal{A}$ such that $P = ND^{-1}$ and D is $\mathcal{Z}$-nonsingular. From (3.1), there exist λ_I's such that $\sum \lambda_I = 1$, where λ_I is an element of generalized elementary factor Λ_{PI} of the plant P with respect to I in $\mathcal{I}$; that is, $\lambda_I \in \Lambda_{PI}$. Now let these λ_I's be fixed. Further, let $\mathcal{I}^{\sharp}$ be the set of I's of these nonzero λ_I's; that is, $\sum_{I\in\mathcal{I}^{\sharp}} \lambda_I = 1$. As in (b), we can consider without loss of generality that for every

$I \in \mathcal{I}^\sharp$, λ_I is not a nilpotent element of $\mathcal{A}$. For each $I \in \mathcal{I}^\sharp$, the $\mathcal{A}_{\lambda_I}$-module $\mathcal{T}_{P,\mathcal{A}_{\lambda_I}}$ is free of rank m by Lemma 4.5. As in (b) again, let V_{λ_I} be a square matrix of size m whose rows are m distinct generators of the $\mathcal{A}_{\lambda_I}$-module $M_r([\,N^t \quad D^t\,]^t)$ ($\simeq \mathcal{T}_{P,\mathcal{A}_{\lambda_I}}$) and assume without loss of generality that V_{λ_I} is over $\mathcal{A}$. Then there exist matrices $\widetilde{X}_I$, $\widetilde{Y}_I$, N_I, D_I over $\mathcal{A}_{\lambda_I}$ such that

$$\begin{gathered}[\,N^t \quad D^t\,]^t = [\,N_I^t \quad D_I^t\,]^t V_{\lambda_I}, \quad \left[\widetilde{Y}_I \quad \widetilde{X}_I\right][\,N^t \quad D^t\,]^t = V_{\lambda_I}, \\ \widetilde{Y}_I N_I + \widetilde{X}_I D_I = E_m \end{gathered} \tag{5.5}$$

with $P = N_I D_I^{-1}$ over $\mathcal{F}(\mathcal{A}_{\lambda_I})$.

We here present a formula of a stabilizing controller which is constructed from the matrices $\widetilde{X}_I$ and $\widetilde{Y}_I$. For any positive integer ω, there are coefficients a_I's in $\mathcal{A}$ with $\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega = 1$. Let ω be a sufficiently large integer. Thus the matrices $\lambda_I^\omega D_I \widetilde{X}_I$ and $\lambda_I^\omega D_I \widetilde{Y}_I$ are over $\mathcal{A}$ for every $I \in \mathcal{I}^\sharp$. The stabilizing controller we will construct has the form

$$C = \left(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I\right)^{-1} \left(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{Y}_I\right). \tag{5.6}$$

In the following we first consider that the matrix $(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I)$ is $\mathcal{Z}$-nonsingular and show that the plant is stabilized by the matrix C of (5.6). After showing it, we will be concerned with the case where the matrix $(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I)$ is $\mathcal{Z}$-singular.

Suppose that the matrix $(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I)$ is $\mathcal{Z}$-nonsingular. To prove that C is a stabilizing controller of P, it is sufficient to show that $(P, C) \in \widehat{F}_{\mathrm{ad}}$ and that four blocks of (2.1) are over $\mathcal{A}$.

We first show that $(P, C) \in \widehat{F}_{\mathrm{ad}}$. The following matrix equation holds:

$$\begin{aligned} E_m + CP &= E_m + \left(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I\right)^{-1} \left(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{Y}_I\right) N D^{-1} \\ &= \left(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I\right)^{-1} \\ &\qquad \left(\left(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I\right) D + \left(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{Y}_I\right) N\right) D^{-1}. \end{aligned}$$

By the (1,1)-block of (5.8), we have

$$E_m + CP = \left(\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I\right)^{-1}. \tag{5.7}$$

This shows that $\det(E_m + CP)$ is a unit of $\mathcal{F}$ so that $(P, C) \in \widehat{F}_{\mathrm{ad}}$.

Next we show that four blocks of (2.1) are over $\mathcal{A}$. The $(2, 2)$-block is the inverse of (5.7):

$$(E_m + CP)^{-1} = \sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I.$$

Similarly, simple calculations show that other blocks are also over $\mathcal{A}$ as follows:
$(2,1)$-block:

$$C(E_n+PC)^{-1}=\sum_{I\in\mathcal{I}^\sharp}a_I\lambda_I^\omega D_I\widetilde{Y}_I,$$

$(1,1)$-block:

$$(E_n+PC)^{-1}=E_n-\sum_{I\in\mathcal{I}^\sharp}a_I\lambda_I^\omega N_I\widetilde{Y}_I,$$

$(1,2)$-block:

$$-P(E_m+CP)^{-1}=-\sum_{I\in\mathcal{I}^\sharp}a_I\lambda_I^\omega N_I\widetilde{X}_I.$$

To finish the proof, we proceed to deal with the case where the matrix $(\sum_{I\in\mathcal{I}^\sharp}a_I\lambda_I^\omega D_I\widetilde{X}_I)$ is $\mathcal{Z}$-singular. To make the matrix $\mathcal{Z}$-nonsingular, we reconstruct the matrices $\widetilde{X}_I$ and $\widetilde{Y}_I$ with an $I\in\mathcal{I}^\sharp$.

Since the sum of $a_I\lambda_I^\omega$'s for $I\in\mathcal{I}^\sharp$ is equal to 1, by Remark 4.4(i, iii) there exists at least one summand, say, $a_{I_0}\lambda_{I_0}^\omega$ with an $I_0\in\mathcal{I}^\sharp$, such that both a_{I_0} and λ_{I_0} belong to $\mathcal{A}\backslash\mathcal{Z}$. Let R_{I_0} be a parameter matrix of $\mathcal{A}_{\lambda_{I_0}}^{m\times n}$. Then the following matrix equation holds over $\mathcal{A}_{\lambda_{I_0}}$:

$$(\widetilde{Y}_{I_0}+R_{I_0}\widetilde{D}_{I_0})N_{I_0}+(\widetilde{X}_{I_0}-R_{I_0}\widetilde{N}_{I_0})D_{I_0}=E_m,$$

where $\widetilde{D}_{I_0}=\det(\lambda_{I_0}^\omega D_{I_0})E_n$ and $\widetilde{N}_{I_0}=\lambda_{I_0}^\omega N_{I_0}\operatorname{adj}(\lambda_{I_0}^\omega D_{I_0})$. Since ω is a sufficiently large integer, the following matrix equation is over $\mathcal{A}$:

$$\begin{aligned}&\big(\lambda_{I_0}^\omega(\widetilde{Y}_{I_0}+R_{I_0}\widetilde{D}_{I_0})\big)\big(\lambda_{I_0}^\omega N_{I_0}\big)\\&\qquad+\big(\lambda_{I_0}^\omega(\widetilde{X}_{I_0}-R_{I_0}\widetilde{N}_{I_0})\big)\big(\lambda_{I_0}^\omega D_{I_0}\big)=\lambda_{I_0}^{2\omega}E_m,\end{aligned}$$

where the matrices surrounded by "$\big($" and "$\big)$" in the left-hand side are over $\mathcal{A}$. From the first matrix equation of (5.5), $\det(D)=\det(D_I)\det(V_{\lambda_I})$ over $\mathcal{A}_{\lambda_I}$ for every $I\in\mathcal{I}^\sharp$. Thus by Remark 4.4(iii) the matrix $\lambda_{I_0}^\omega D_{I_0}$ is $\mathcal{Z}$-nonsingular and so is the matrix $\widetilde{D}_{I_0}$ $(=\det(\lambda_{I_0}^\omega D_{I_0})E_n)$.

Consider now the following matrix equation over $\mathcal{A}$:

$$\begin{aligned}&\begin{bmatrix}\sum_{I\in\mathcal{I}^\sharp}a_I\lambda_I^\omega D_I\widetilde{X}_I & \sum_{I\in\mathcal{I}^\sharp}a_I\lambda_I^\omega D_I\widetilde{Y}_I\\ -a_{I_0}\lambda_{I_0}^\omega\det(\lambda_{I_0}^\omega D_{I_0})\widetilde{N}_{I_0} & a_{I_0}\lambda_{I_0}^\omega\det(\lambda_{I_0}^\omega D_{I_0})\widetilde{D}_{I_0}\end{bmatrix}\begin{bmatrix}D & O\\ N & E_n\end{bmatrix}\\&\qquad=\begin{bmatrix}D & \sum_{I\in\mathcal{I}^\sharp}a_I\lambda_I^\omega D_I\widetilde{Y}_I\\ O & a_{I_0}\lambda_{I_0}^\omega\det(\lambda_{I_0}^\omega D_{I_0})\widetilde{D}_{I_0}\end{bmatrix}.\end{aligned}\tag{5.8}$$

The $(1,1)$-block of (5.8) can be understood in the following way. From the last matrix equation in (5.5) we have the following matrix equation over $\mathcal{A}_{\lambda_I}$:

$$D_I\widetilde{Y}_IN+D_I\widetilde{X}_ID=D.$$

Considering the above equation multiplied by $a_I\lambda_I^\omega$ over $\mathcal{A}$, we have the following equation over $\mathcal{A}$:

$$a_I\lambda_I^\omega D_I\widetilde{Y}_IN+a_I\lambda_I^\omega D_I\widetilde{X}_ID=a_I\lambda_I^\omega D+a_I\lambda_I^\omega Z,\tag{5.9}$$

where Z is a matrix over $\mathcal{A}$ such that $\lambda_I^x Z$ is the zero matrix for some positive integer x. Since ω is a large positive integer, we can consider that the matrix $a_I \lambda_I^\omega Z$ in (5.9) becomes the zero matrix. Therefore, the $(1,1)$-block of (5.8) holds. Then the matrix of the right-hand side of (5.8) is $\mathcal{Z}$-nonsingular since both of the matrices D and $a_{I_0}\lambda_{I_0}^\omega \det(\lambda_{I_0}^\omega D_{I_0})\widetilde{D}_{I_0}$ in the right-hand side of (5.8) are $\mathcal{Z}$-nonsingular. Hence the first matrix of (5.8) is also $\mathcal{Z}$-nonsingular by Remark 4.4(iii). By Lemma 4.3 and (5.8), there exists a matrix R'_{I_0} of $\mathcal{A}^{m\times n}$ such that the following matrix is $\mathcal{Z}$-nonsingular:

$$\sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I - a_{I_0}\lambda_{I_0}^{2\omega} D_{I_0} \operatorname{adj}(\lambda_{I_0}^\omega D_{I_0}) R'_{I_0} \widetilde{N}_{I_0}. \tag{5.10}$$

Now let $R_{I_0} := \lambda_{I_0}^\omega \operatorname{adj}(\lambda_{I_0}^\omega D_{I_0}) R'_{I_0}$, $\widetilde{X}_{I_0} := \widetilde{X}_{I_0} - R_{I_0}\widetilde{N}_{I_0}$, and $\widetilde{Y}_{I_0} := \widetilde{Y}_{I_0} + R_{I_0}\widetilde{D}_{I_0}$. Then the matrix $\sum_{I\in\mathcal{I}^\sharp} a_I\lambda_I^\omega D_I \widetilde{X}_I$ becomes equal to (5.10) and $\mathcal{Z}$-nonsingular. □

Remark 5.1. From the proof above, if (i) we can check (3.1) and if (ii) we can construct the right-coprime factorizations of the given causal plant over $\mathcal{A}_{\lambda_I}$ for every $I \in \mathcal{I}^\sharp$, then we can construct stabilizing controllers of the plant, where λ_I is an element of the generalized elementary factor of the plant. For (i), if we can compute, for example, the Gröbner basis [4] over $\mathcal{A}$ and if the generalized elementary factors of the plant are finitely generated, (3.1) can be checked. For (ii), it is already known by Lemmas 4.5 and 4.6 that there exist the right-coprime factorizations of the plant over $\mathcal{A}_{\lambda_I}$.

Let us give an example concerning the Gröbner basis. Consider the generalized elementary factors of Example 3.4. They are expressed as

$$\begin{aligned}
\Lambda_{PI_1} &= (\langle(1+2z)(1+z+z^2)(1-3z)\rangle) + (\langle(1+2z)(1+z+z^2)z^2\rangle)\\
&\qquad +(\langle(1+2z)(1+z+z^2)z^3\rangle),\\
\Lambda_{PI_2} &= (\langle(1+z)(1+2z+4z^2)(1-3z)\rangle) + (\langle(1+z)(1+2z+4z^2)z^2\rangle)\\
&\qquad +(\langle(1+z)(1+2z+4z^2)z^3\rangle),\\
\Lambda_{PI_3} &= (\langle(1+z)(1+2z)(1-3z)\rangle) + (\langle(1+z)(1+2z)z^2\rangle)\\
&\qquad +(\langle(1+z)(1+2z)z^3\rangle).
\end{aligned}$$

Hence each of them has three generators and so is finitely generated. Suppose here that we can calculate the Gröbner basis over $\mathcal{A}$ (of Example 3.4). Then as above the plant is stabilizable if and only if the Gröbner basis of the nine generators contains 1.

In the following two examples we follow the proof of Theorem 3.3. In the first one, we construct a stabilizing controller with part (c). In the other example, we follow part (b). On the other hand we do not follow part (a) since it can be followed easily with part (a) of (i) in the proof of Proposition 4.1.

Example 5.2. We continue Example 3.4 (and 4.7) and construct a stabilizing controller of the plant as in (iii)→(i) of the proof above. Let the notation be as in Examples 3.4 and 4.7.

Since, in this example, $\Lambda_{PI_1} + \Lambda_{PI_2} = \mathcal{A}$ holds, $\mathcal{I}^\sharp = \{I_1, I_2\}$. For $I_1 \in \mathcal{I}^\sharp$, the matrices N_{I_1}, D_{I_1}, $\widetilde{X}_{I_1}$, and $\widetilde{Y}_{I_1}$ of (5.5) over $\mathcal{A}_{\lambda_{I_1}}$ have been calculated as (4.5) and (4.6). For $I_2 \in \mathcal{I}^\sharp$, the matrices N_{I_2}, D_{I_2}, $\widetilde{X}_{I_2}$, and $\widetilde{Y}_{I_2}$ of (5.5) over $\mathcal{A}_{\lambda_{I_2}}$ can be calculated analogously as follows:

$$\begin{aligned}
N_{I_2} &= \begin{bmatrix} \lambda_{I_2}^{-1}\alpha_{I_2}\langle(1+2z)(1-3z+z^2)(1+z+z^2)\rangle \\ 1 \end{bmatrix},\\
D_{I_2} &= [\,\lambda_{I_2}^{-1}\alpha_{I_2}\langle(1+z)(1+2z)(1-3z+z^2)\rangle\,],
\end{aligned}$$

$$\widetilde{Y}_{I_2} = [\,0 \quad 1\,]\,, \widetilde{X}_{I_2} = [\,0\,]\,.$$

Then the following matrices are over $\mathcal{A}$:

$$\begin{aligned}
&\lambda_{I_1} D_{I_1} \widetilde{X}_{I_1} = [\,0\,]\,, \quad \lambda_{I_1} D_{I_1} \widetilde{Y}_{I_1} = [\,\alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle \quad 0\,]\,,\\
&\lambda_{I_2} D_{I_2} \widetilde{X}_{I_2} = [\,0\,]\,, \quad \lambda_{I_2} D_{I_2} \widetilde{Y}_{I_2} = [\,0 \quad \alpha_{I_2}\langle(1+z)(1+2z)(1-3z+z^2)\rangle\,]\,.
\end{aligned}$$

Hence in this example, we can let $\omega = 1$ as a sufficiently large integer and $a_I = 1$ for all $I \in \mathcal{I}^\sharp$ (since $\sum_{I\in\mathcal{I}^\sharp} \lambda_I^\omega = 1$).

Note here that the matrix $\lambda_{I_1} D_{I_1} \widetilde{X}_{I_1} + \lambda_{I_2} D_{I_2} \widetilde{X}_{I_2}$ is $\mathcal{Z}$-singular. Hence we should reconstruct the matrices $\widetilde{Y}_{I_i}$ and $\widetilde{X}_{I_i}$ with i being either 1 or 2 as in the proof of Theorem 3.3. Since, in this example, both λ_{I_1} and λ_{I_2} are nonzerodivisors, we can choose each of 1 and 2. This example proceeds by reconstructing the matrices $\widetilde{Y}_{I_1}$ and $\widetilde{X}_{I_1}$, which means that I_1 is used as I_0 in the proof of Theorem 3.3. The actual reconstruction is done by following the proof of Lemma 4.3.

Consider the first matrix of (5.8). Recall that $\widetilde{N}_{I_0} = \lambda_{I_0}^\omega N_{I_0} \operatorname{adj}(\lambda_{I_0}^\omega D_{I_0})$ and $\widetilde{D}_{I_0} = \det(\lambda_{I_0}^\omega D_{I_0}) E_n$. In this example, they are given as

$$\begin{aligned}
\widetilde{N}_{I_1} &= (\widetilde{N}_{I_0} =) \begin{bmatrix} \lambda_{I_1} \\ \alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle \end{bmatrix},\\
\widetilde{D}_{I_1} &= (\widetilde{D}_{I_0} =) \alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle E_2.
\end{aligned}$$

One can check that the first matrix of (5.8) is $\mathcal{Z}$-nonsingular. Then we construct a matrix R'_{I_0} of $\mathcal{A}^{1\times 2}$ such that (5.10) is $\mathcal{Z}$-nonsingular. To do so, we follow temporarily the proof of the Lemma 4.3.

Consider the first matrix of (5.8) as the matrix of (4.2), that is,

$$\begin{aligned}
A &= \sum_{I\in\mathcal{I}^\sharp} a_I \lambda_I^\omega D_I \widetilde{X}_I = [\,0\,]\,,\\
B &= -a_{I_0} \lambda_{I_0}^\omega \det(\lambda_{I_0}^\omega D_{I_0}) \widetilde{N}_{I_0}\\
&= -\alpha_{I_1} \lambda_{I_1} \langle(1+z)(1+2z)(1-3z)\rangle \begin{bmatrix} \lambda_{I_1} \\ \alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle \end{bmatrix}.
\end{aligned}$$

Then we choose a full-size a minor of $[\,A^t \quad B^t\,]^t$ having as few rows from B as possible. In this example, we can choose both entries in B. Here we choose the $(1,1)$-entry of B, so that

$$a = -\alpha_{I_1} \lambda_{I_1}^2 \langle(1+z)(1+2z)(1-3z)\rangle. \tag{5.11}$$

Thus we have $k = 1$, $i_1 = 1$, and $j_1 = 1$, where the notations $k, i_1, \ldots, i_k$, and $j_1, \ldots, j_k$ are as in the proof of Lemma 4.3. Hence R in the proof is given as $R = [\,1 \quad 0\,]$. We can confirm that $A + RB = [\,a\,]$ which is $\mathcal{Z}$-nonsingular by observing that every factor of the right-hand side of (5.11) has a nonzero constant term.

From here on we proceed with following again the proof of Theorem 3.3. The notation R used above corresponds to the notation R'_{I_0} in the proof of Theorem 3.3 (that is, $R'_{I_0} = [\,1 \quad 0\,]$). The matrix R_{I_1} is given as follows:

$$R_{I_1} = (R_{I_0} =) \lambda_{I_1}^\omega \operatorname{adj}(\lambda_{I_1}^\omega D_{I_1}) R'_{I_1} = \lambda_{I_1} [\,1 \quad 0\,]\,.$$

Then new $\widetilde{X}_{I_1}$ and $\widetilde{Y}_{I_1}$ are given as follows:

$$\begin{aligned}\widetilde{X}_{I_1} &:= \widetilde{X}_{I_1} - R_{I_1}\widetilde{N}_{I_1} = [\,-\lambda_{I_1}^2\,],\\ \widetilde{Y}_{I_1} &:= \widetilde{Y}_{I_1} + R_{I_1}\widetilde{D}_{I_1} = [\,1+\alpha_{I_1}\lambda_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle \quad 0\,].\end{aligned}$$

Therefore, a stabilizing controller C of the form (5.6) is obtained as

$$\begin{aligned}C &= (\lambda_{I_1}D_{I_1}\widetilde{X}_{I_1} + \lambda_{I_2}D_{I_2}\widetilde{X}_{I_2})^{-1}(\lambda_{I_1}D_{I_1}\widetilde{Y}_{I_1} + \lambda_{I_2}D_{I_2}\widetilde{Y}_{I_2})\\ &= \frac{-1}{\alpha_{I_1}\lambda_{I_1}^2\langle(1+z)(1+2z)(1-3z)\rangle}\\ &\qquad \begin{bmatrix}\alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle(1+\alpha_{I_1}\lambda_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle)\\ \alpha_{I_2}\langle(1+z)(1+2z)(1-3z+z^2)\rangle\end{bmatrix}^t.\end{aligned}$$

The matrix $H(P,C)$ over $\mathcal{A}$ with the stabilizing controller C above is expressed as follows:

$$H(P,C) = \begin{bmatrix}h_{11} & h_{12} & h_{13}\\ h_{21} & h_{22} & h_{23}\\ h_{31} & h_{32} & h_{33}\end{bmatrix},$$

where

$$\begin{aligned}h_{11} &= -\alpha_{I_1}\lambda_{I_1}^2\langle(1+z)(1+2z)(1-3z)\rangle\\ &\qquad\qquad +\alpha_{I_2}\langle(1+z)(1-3z+z^2)(1+2z+4z^2)\rangle,\\ h_{12} &= -\alpha_{I_2}\langle(1+2z)(1+z+z^2)(1-3z+z^2)\rangle,\\ h_{13} &= \lambda_{I_1}^3,\\ h_{21} &= -\alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle\\ &\qquad\qquad (1+\lambda_{I_1}\alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle),\\ h_{22} &= \alpha_{I_1}(\langle(1+2z)(1-3z)(1+z+z^2)\rangle(1+\alpha_{I_1}\lambda_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle)\\ &\qquad\qquad -\lambda_{I_1}^2\langle(1+z)(1+2z)(1-3z)\rangle),\\ h_{23} &= \alpha_{I_1}\lambda_{I_1}^2\langle(1+z)(1-3z)(1+2z+4z^2)\rangle,\\ h_{31} &= \alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle(1+\alpha_{I_1}\lambda_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle),\\ h_{32} &= \alpha_{I_2}\langle(1+z)(1+2z)(1-3z+z^2)\rangle,\\ h_{33} &= -\alpha_{I_1}\lambda_{I_1}^2\langle(1+z)(1+2z)(1-3z)\rangle.\end{aligned}$$

Before finishing this example, let us show that the $\mathcal{A}$-module $\mathcal{T}_P$ is *not* free. We show it by contradiction. Suppose that $\mathcal{T}_P$ is free. Then the $\mathcal{A}$-module $M_r(T)$ is also free. Since the matrix D, a part of T, is nonsingular, the rank of $M_r(T)$ is m. Let V be a matrix in $(\mathcal{A})_m$ whose rows are m distinct generators of $M_r(T)$. As in (5.5), we have matrices $\widetilde{Y}$, $\widetilde{X}$, N', D' over $\mathcal{A}$ such that

$$[\,N^t \quad D^t\,]^t = [\,N'^t \quad D'^t\,]^t V,\quad \begin{bmatrix}\widetilde{Y} & \widetilde{X}\end{bmatrix}[\,N^t \quad D^t\,]^t = V,\quad \widetilde{Y}'N' + \widetilde{X}'D' = E_1.$$

However, the last matrix equation is inconsistent with the fact that the plant P does not have coprime factorization. Therefore, $\mathcal{T}_P$ is not free. Nevertheless we note that $\mathcal{T}_P$ is projective by Theorem 3.3. □

Example 5.3. Let us follow part (b) in the proof of Theorem 3.3. Suppose that (i) of Theorem 3.3 holds, that is, the modules $\mathcal{T}_P$ and $\mathcal{W}_P$ are projective.

Consider again the plant P of (3.2). Let $F_1 = \{\lambda_{I_1}, \lambda_{I_2}\}$, where λ_{I_1} and λ_{I_2} are given as in (3.8). Then we have known that $\Sigma_{f\in F_1} f = 1$ and that there exists a right-coprime factorization of the plant over $\mathcal{A}_f$ for every $f \in F_1$. By Lemma 4.2, the transposed plant P^t is stabilizable. We can construct its stabilizing controller by analogy to Example 5.2. Further we see that for both λ_{I_1} and λ_{I_2}, the transposed plant P^t has right-coprime factorizations over $\mathcal{A}_{\lambda_{I_1}}$ and $\mathcal{A}_{\lambda_{I_2}}$; that is, P has left-coprime factorizations over $\mathcal{A}_{\lambda_{I_1}}$ and $\mathcal{A}_{\lambda_{I_2}}$. Thus let $F_2 = \{\lambda_{I_1}, \lambda_{I_2}\}$. For $\lambda_{I_1} \in F_2$, we have the matrices $\widetilde{N}_{I_1}$ $\widetilde{D}_{I_1}$, Y_{I_1}, X_{I_1} over $\mathcal{A}_{\lambda_{I_1}}$ such that $\widetilde{N}_{I_1} Y_{I_1} + \widetilde{N}_{I_1} X_{I_1} = E_2$ and

$$\begin{aligned}
\widetilde{N}_{I_1} &= \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right], \quad Y_{I_1} = [\,1 \quad 0\,], \quad X_{I_1} = \left[\begin{smallmatrix}0&0\\0&1\end{smallmatrix}\right],\\
\widetilde{D}_{I_1} &= \begin{bmatrix} \lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1+2z)(1-3z)\rangle & 0 \\ \lambda_{I_1}^{-1}\alpha_{I_1}\langle(1+z)(1-3z)(1+2z+4z^2)\rangle & 1 \end{bmatrix}.
\end{aligned}$$

On the other hand, for $\lambda_{I_2} \in F_2$, we have the matrices $\widetilde{N}_{I_2}$ $\widetilde{D}_{I_2}$, Y_{I_2}, X_{I_2} over $\mathcal{A}_{\lambda_{I_2}}$ such that $\widetilde{N}_{I_2} Y_{I_2} + \widetilde{N}_{I_2} X_{I_2} = E_2$ and

$$\begin{aligned}
\widetilde{N}_{I_2} &= \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right], \quad Y_{I_2} = [\,1 \quad 0\,], \quad X_{I_2} = \left[\begin{smallmatrix}0&1\\0&0\end{smallmatrix}\right],\\
\widetilde{D}_{I_2} &= \begin{bmatrix} 0 & \lambda_{I_2}^{-1}\alpha_{I_2}\langle(1+z)(1+2z)(1-3z+z^2)\rangle \\ 1 & -\lambda_{I_2}^{-1}\alpha_{I_2}\langle(1+2z)(1+z+z^2)(1-3z+z^2)\rangle \end{bmatrix}.
\end{aligned}$$

Now we let $F = \{\lambda_{I_1}^2, \lambda_{I_1}\lambda_{I_2}, \lambda_{I_2}^2\}$ $(= \{f_1 f_2 \,|\, f_1 \in F_1, f_2 \in F_2\})$. Then F still generates $\mathcal{A}$ since $\lambda_{I_1}^2 + 2\lambda_{I_1}\lambda_{I_2} + \lambda_{I_2}^2 = 1$.

In the following we consider the case $f = \lambda_{I_1}^2$. Then using the matrix K of (4.3), we have (5.1) with $\nu = 1$, $K_f = K$, and $V_f = \lambda_{I_1}\Delta_{I_1}T$.

Then the ideal $I_{m\mathcal{A}}(f^\nu K_f)$ is generated by

$$\lambda_{I_1}^3, \quad \alpha_{I_1}\lambda_{I_1}^2\langle(1+z)(1-3z)(1+2z+4z^2)\rangle, \quad \alpha_{I_1}\lambda_{I_1}^2\langle(1+z)(1+2z)(1-3z)\rangle.$$

Thus since each of them is in Λ_{PI_1}, (5.2) holds. Further we can observe that for any integer ξ greater than 1, (5.3) holds since $\lambda_{I_1}^3 \in I_{m\mathcal{A}}(f^\nu K_f)$.

For the other cases $f = \lambda_{I_1}\lambda_{I_2}$ and $f = \lambda_{I_2}^2$, we can follow the relations of (5.2) and (5.3) analogously. Details are left to interested readers.

Remark 5.4. Since Anantharam's example in [1] is artificial, we do not present here the construction of a stabilizing controller. However, we can construct it as part (c) in the proof of Theorem 3.3 (Since Anantharam in [1] did not consider the causality, we let $\mathcal{Z} = \{0\}$ so that $\mathcal{P} = \mathcal{F}$.)

6. Causality of stabilizing controllers.

In this section, we present two facts: (i) for a stabilizable causal plant, there exists at least one stabilizing causal controller and (ii) the stabilizing controller of the strictly causal plant is causal, which inherits Theorem 4.1 in section III of [14, p. 888] and Proposition 1 of [11].

PROPOSITION 6.1. *For every stabilizable causal plant, there exists at least one stabilizing causal controller of the plant.*

Proof. In the construction of the stabilizing controller in part (c) of the proof of Theorem 3.3, the denominator matrix of (5.6) is $\mathcal{Z}$-nonsingular. Suppose that the obtained stabilizing controller is expressed as $\widetilde{B}^{-1}\widetilde{A}$ with the matrices $\widetilde{A}$ and $\widetilde{B}$ over $\mathcal{A}$ such that $\widetilde{B}$ is $\mathcal{Z}$-nonsingular. Then since the relation $\widetilde{B}^{-1}\widetilde{A} = (\det(\widetilde{B})E_m)^{-1}(\mathrm{adj}(\widetilde{B})\widetilde{A})$ holds, every entry of $\widetilde{B}^{-1}\widetilde{A}$ is causal. $\square$

Proposition 6.2. *For every stabilizable strictly causal plant, all stabilizing controllers of the plant must be causal.*

Proof. Suppose that the plant P is stabilizable and strictly causal. Suppose further that C is a stabilizing controller of P. We employ the notation from part (c) of the proof of Theorem 3.3. Thus $a_{I_0}, \lambda_{I_0} \in \mathcal{A}\backslash\mathcal{Z}$ and $\widetilde{Y}_{I_0}N_{I_0} + \widetilde{X}_{I_0}D_{I_0} = E_m$ with $P = N_{I_0}D_{I_0}^{-1} \in \mathcal{F}(\mathcal{A}_{\lambda_{I_0}})$ from (5.5). Let $\mathcal{Z}_{\lambda_{I_0}} = \{z/1 \cdot u \in \mathcal{A}_{\lambda_{I_0}} \mid z \in \mathcal{Z}, u \text{ is a unit of } \mathcal{A}_{\lambda_{I_0}}\}$. Then this $\mathcal{Z}_{\lambda_{I_0}}$ is again a principal ideal of $\mathcal{A}_{\lambda_{I_0}}$.

Observe here that Lemma 8.3.2 of [13] and its proof hold even over a general commutative ring. According to its proof, there exist matrices $\widetilde{A}$ and $\widetilde{B}$ over $\mathcal{A}_{\lambda_{I_0}}$ such that $C = \widetilde{B}^{-1}\widetilde{A}$ and $\widetilde{A}N_{I_0} + \widetilde{B}D_{I_0} = E_m$ ($\widetilde{A}$ and $\widetilde{B}$ correspond to T and S, respectively, in the proof of Lemma 8.3.2 of [13]). Observe also that every entry of N_{I_0} is in $\mathcal{Z}_{\lambda_{I_0}}$. Thus reviewing the proof of Lemma 3.5 of [14], in which the calligraphic H and K in [14] correspond to $\mathcal{A}_{\lambda_{I_0}}$ and $\mathcal{Z}_{\lambda_{I_0}}$, respectively, we have $\det \widetilde{B} \in \mathcal{A}_{\lambda_{I_0}} \backslash \mathcal{Z}_{\lambda_{I_0}}$. This implies that $\widetilde{B}^{-1}\widetilde{A} \in \mathcal{P}^{m\times n}$ by noting that $\lambda_{I_0} \in \mathcal{A}\backslash\mathcal{Z}$. Thus C is causal. □

7. Further work. In this paper we have presented criteria for feedback stabilizability. We have also presented a construction of a stabilizing controller to which Sule's method cannot be applied. Recently the first author [8] has developed a parameterization of stabilizing controllers, which is based on the results of this paper and which does not require coprime factorizability. This can be applied to models to which Youla parameterization cannot be applied.

REFERENCES

[1] V. Anantharam, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1030–1031.

[2] M. F. Atiyah and I. G. MacDonald, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, MA, 1969.

[3] C. A. Desoer, R. W. Liu, J. Murray, and R. Saeks, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 399–412.

[4] K. O. Geddes, S. R. Czapor, and G. Labahn, *Algorithms for Computer Algebra*, Kluwer Academic Publishers, Norwell, MA, 1992.

[5] J. D. Greenfield, ed., *Microprocessor Handbook*, John Wiley and Sons, New York, 1985.

[6] T. Lam, *Serre's Conjecture*, Lecture Notes in Math. 635, Springer-Verlag, New York, 1978.

[7] B. R. McDonald, *Linear Algebra over Commutative Rings*, Monogr. Textbooks Pure Appl. Math. 87, Marcel Dekker, New York, 1984.

[8] K. Mori, *Parameterization of stabilizing controllers over commutative rings*, SIAM J. Control Optim., submitted.

[9] K. Mori and K. Abe, *Improvement of generalized elementary factors and new criteria of the feedback stabilizability over integral domain*, in Proceedings of the European Control Conference TH-M D3 (533), 1997.

[10] S. Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.

[11] V. R. Sule, *Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control Optim., 32 (1994), pp. 1675–1695.

[12] V. R. Sule, *Corrigendum: Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control Optim., 36 (1998), pp. 2194–2195.

[13] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

[14] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880–894.