SIAM J. CONTROL AND OPTIMIZATION
Vol. 22, No. 3, May 1984

DUAL APPROXIMATIONS IN OPTIMAL CONTROL*

WILLIAM W. HAGER† AND GEORGE D. IANCULESCU‡

Abstract. We analyze a dual approximation for the solution to an optimal control problem. The differential equation is handled with a Lagrange multiplier while other constraints are treated explicitly. An algorithm for solving the dual problem is presented.

Key words. duality, approximation, finite elements, sensitivity, optimal control

TABLE OF CONTENTS

1. Introduction. In computing the solution to an optimal control problem, most of the difficulty centers around the differential equation. In this paper, we consider a dual approach where the differential equation is handled with a Lagrange multiplier, while other constraints are treated explicitly. This scheme was first studied by Rockafellar [47], who establishes existence results and optimality conditions for primal and dual solutions. We now analyze the following numerical aspects of the dual procedure:

(1) Existence of finite dimensional approximations.
(2) Regularity of dual solutions.
(3) Relations between dual multipliers and primal solutions.
(4) Error estimates for piecewise polynomial approximation.
(5) Techniques for solving the dual problem.

The first error estimate for a dual approximation to a control problem is given by Bosarge and Johnson [5], who study unconstrained problems with quadratic cost and linear system dynamics. For piecewise polynomials of degree k, they show that the $\mathscr{L}^2$ error in the approximating control and state is order k. In [14] we introduce linear inequality state and control constraints, and analyze a *full dual* scheme where multipliers are attached to each constraint. The dual optimization is related to an

* Received by the editors November 29, 1979, and in revised form February 3, 1983. This research was supported partly by the National Science Foundation under grants MCS 8101892 and MCS 7825526, by the Office of Naval Research under grant N00014-76-C-0369, by the Hertz Foundation, and by the Ford Foundation. The analysis of the dual scheme and the algorithm for solving the dual problem are contributed by the first author. The four numerical examples are contributed by the second author.

† Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16802.
‡ Jet Propulsion Laboratory, Pasadena, California 91109.

energy projection, and the error is at best order 1.5 due to discontinuities in the control's derivative. Later [16] our estimates are extended to general convex problems, and examples are analyzed in [20]. Mathis and Reddien [32] notice that Bosarge and Johnson's estimate for the control error is not optimal, and sharpen this bound using a duality argument [37]. Some dual approximations to systems described by partial differential equations are developed by Mossino [35], [36] and Bosarge, Johnson and Smith [6].

Other techniques for constrained control problems are contained in [54], [25], [3], [10], [24], [56] and [39]. Penalty methods for variational problems were introduced by Courant [9], and applied to control problems by Russell [54], Lasdon, Warren and Rice [25], and others [3], [10]. These methods have wide applicability, although the penalized problem is ill-conditioned as the penalty grows—see Luenberger [29]. Jacobson and Lele [24] note that some state constraints can be removed by Valentine's device [57]. Thompson and Volz [56] show that control problems with linear dynamics, quadratic cost and a single linear inequality state constraint can be solved using a nonsymmetric Riccati equation. Pironneau and Polak [39] present a dual method of centers for problems with inequality endpoint constraints and affine inequality control constraints.

Advantages of the dual scheme are its speed and generality; problems with endpoint, control and state constraints can be handled. Although our convergence theory assumes that the system dynamics is linear, the scheme applies to nonlinear systems. Unfortunately, there are cases [31], [46] where the dual does not solve the primal. A cure for "duality gaps" is the multiplier methods [4], [21], [41], [48], [49] which combine penalty and duality techniques.

2. The method. A control problem is the constrained minimization of a functional $C(x, u)$ over a collection of controls

$$u: \mathcal{T} \to R^m$$

and a collection of states

$$x: \mathcal{T} \to R^n$$

where R denotes the real numbers and R^n is the n-fold Cartesian product $R \times R \times \cdots \times R$. For convenience, let us assume that $\mathcal{T}$ is the interval $[0, 1] \subset R$. Throughout this paper, Lebesgue measure is used for $\mathcal{T}$, and measurable functions are equal if they are equal almost everywhere. Let $\mathcal{Z}$ denote the set of pairs (x, u) where x is absolutely continuous and u is summable.

The admissible set for the control problem is described by two types of constraints. First there is the *system dynamics* $M(x, u) = 0$ where $M: \mathcal{Z} \to \mathcal{L}^1$ is a differential operator which we assume is linear:

$$M(x, u)(t) := x'(t) - A(t)x(t) - B(t)u(t);$$

here $\mathcal{L}^1$ is the space of summable functions $f: \mathcal{T} \to R^n$, $A(t)$ is an $n \times n$ matrix for each $t \in \mathcal{T}$ whose individual elements are summable, and $B(t)$ is an $n \times m$ matrix for every $t \in \mathcal{T}$ whose elements are essentially bounded. Second, there may be constraints such as

$$x(0) = a \qquad \text{(initial condition)},$$

$$x(1) = b \qquad \text{(target)},$$

$$|u(t)| \leqq 2 \qquad \text{(control constraint), or}$$

$$x(t) \geqq 0 \qquad \text{(state constraint)}.$$

We assume these conditions are embedded in the cost functional by setting $C(x, u) = \infty$ when the constraint is violated. This convention is discussed in Rockafellar's paper [52]. Hence $C: \mathscr{Z} \to \bar{R}$ where $\bar{R}$ is the extended reals $R \cup \{\infty\}$, and the control problem takes the form

$$\text{(P)} \qquad \begin{aligned} &\text{minimize} \quad C(z) \\ &\text{subject to} \quad M(z) = 0, \quad z \in \mathscr{Z}. \end{aligned}$$

Of course, z denotes the pair (x, u). Since the cost is minimized, we are only concerned with those z for which $C(z)$ is finite. The *effective domain* of C is given by

$$\operatorname{dom} C := \{z \in \mathscr{Z}: C(z) < \infty\}.$$

It is assumed that C *proper* and there exists a *feasible function* for (P); that is, the effective domain of C is nonempty and there exists $z \in \operatorname{dom} C$ such that $M(z) = 0$.

Now we formulate the dual of (P). Letting $\mathscr{L}^\infty$ be the space of essentially bounded functions $f: \mathscr{T} \to R^n$, the *dual functional* $L: \mathscr{L}^\infty \to R \cup \{-\infty\}$ is defined by

$$L(p) = \inf \{C(z) + \langle p, M(z)\rangle: z \in \mathscr{Z}\}$$

where $\langle \cdot, \cdot \rangle$ is the usual $\mathscr{L}^2$ *inner product*:

$$\langle f, g\rangle := \int_{\mathscr{T}} f(t) \cdot g(t)\, dt$$

for all measurable $f, g: \mathscr{T} \to R^n$. Here $\cdot$ is the *Euclidean dot product.* The dual problem becomes:

$$\text{(D)} \qquad \begin{aligned} &\text{maximize} \quad L(p) \\ &\text{subject to} \quad p \in \mathscr{L}^\infty. \end{aligned}$$

Since L is maximized, the effective domain of the dual functional is given by

$$\operatorname{dom} L = \{p \in \mathscr{L}^\infty: L(p) > -\infty\}.$$

Clearly, from the definition of L,

$$\sup \{L(p): p \in \mathscr{L}^\infty\} \leqq \inf \{C(z): z \in \mathscr{Z}, M(z) = 0\}. \tag{2.1}$$

This inequality is sometimes called weak duality [31]. The stronger statement, "There exists a solution to (D) and (2.1) is an equality," follows from the

BOUNDEDNESS ASSUMPTION. *There exists* $\rho > 0$ *such that*

$$\sup_{\substack{w \in \mathscr{L}^1 \\ \|w\|_{\mathscr{L}^1} \leqq \rho}} \inf_{\substack{z \in \mathscr{Z} \\ M(z) = w}} C(z) < \infty$$

where

$$\|w\|_{\mathscr{L}^1} = \int_{\mathscr{T}} |w(t)|\, dt$$

and $|\cdot|$ *is the Euclidean norm.*

THEOREM 2.1. *If C is convex and the boundedness hypothesis is satisfied, then there exists a solution p to the dual problem, and*

$$L(p) = \inf \{C(z): z \in \mathscr{Z}, M(z) = 0\}.$$

Notice that the dual problem has a solution even though the primal problem may have no solution. Theorem 2.1 is an elementary application of general duality principles (see Theorem A.3 in Appendix 1).

If p solves the dual problem, any solution to the primal problem minimizes

$$C(z)+\langle p, M(z)\rangle \tag{2.2}$$

over $z \in \mathscr{Z}$. Thus, to solve the primal problem, we can first solve the dual and then find those $z \in \mathscr{Z}$ which attain the minimum in (2.2). This procedure is modified for numerical computations. We replace the dual feasible set by a closed subset $\mathscr{S}_h$ of a finite dimensional space giving us the approximation:

$$(\mathrm{D}_h) \qquad \begin{aligned} &\text{maximize} \quad L(p) \\ &\text{subject to} \quad p \in \mathscr{S}_h. \end{aligned}$$

An important issue is whether there exists a solution to (D_h). By Theorem 2.2 below, the boundedness hypothesis assures existence. If p_h solves (D_h), we take as an approximation to a primal solution any $z_h \in \mathscr{Z}$ for which

$$L(p_h) = C(z_h)+\langle p_h, M(z_h)\rangle.$$

The paper's main focus is the second issue: Is z_h "close" to a primal solution? Under a uniform convexity hypothesis, the answer is yes.

These convergence properties are related to the smoothness of dual solutions. If an optimal p lies just in $\mathscr{L}^\infty$, the dual problem may be hard since the approximation of essentially bounded functions using standard $\mathscr{S}_h$ is not easy. In the following sections, we observe that p has some smoothness. This section concludes with an existence theorem for (D_h). Appendix 1 proves a more general result.

THEOREM 2.2. *Suppose that $\mathscr{S}_h \subset \mathscr{L}^\infty$ where $\mathscr{S}_h$ is also a closed subset of a finite dimensional space. If the boundedness hypothesis holds, then (D_h) has a solution.*

3. Bounded variation. Under appropriate assumptions, the classical minimum principle [40], [26] for the control problem

$$\begin{aligned} &\text{minimize} \quad \int_{\mathscr{T}} f(x(t), u(t), t)\, dt \\ &\text{subject to} \quad M(x, u) = 0, \quad (x, u) \in \mathscr{Z}, \quad x(0) = a, \\ &\qquad\qquad u(t) \in U \subset R^m \quad \text{almost everywhere,} \end{aligned}$$

states that an optimal solution satisfies

$$h(u(t), t) = \min\{h(v, t): v \in U\} \quad \text{almost everywhere}$$

where

$$h(v, t) = f(x(t), v, t)+q(t)^T B(t) v$$

and

$$\begin{aligned} &q'(t) = -A(t)^T q(t) - \nabla_x f(x(t), u(t), t)^T \quad \text{almost everywhere,} \\ &q(1) = 0. \end{aligned}$$

Above, T denotes *transpose*, ∇_x is the *gradient* with respect to the state argument, and q is called the *costate*.

We expect that dual solutions are related to the costate, but observe that q is differentiable while dom $L \subset \mathscr{L}^\infty$. However, a fairly weak hypothesis guarantees some smoothness for elements in dom L. Let $\mathscr{C}^\infty \subset \mathscr{L}^\infty$ be the subspace of infinitely differentiable functions. We introduce the sets

$$\mathscr{D}^\gamma = \{y \in \mathscr{C}^\infty : \|y\|_{\mathscr{C}} \leqq \gamma\}, \quad \mathscr{E} = \{y \in \mathscr{C}^\infty : y(0) = y(1) = 0\}, \quad \mathscr{E}^\gamma = \mathscr{E} \cap \mathscr{D}^\gamma,$$

where

$$\|y\|_{\mathscr{C}} = \sup\{|y(t)| : t \in \mathscr{T}\},$$

and the

INTERIORITY ASSUMPTION. *There exist* $\bar{z} = (\bar{x}, \bar{u}) \in \mathscr{Z}$ *and* $\gamma > 0$ *such that*

$$\sup\{C(\bar{x} + \psi, \bar{u}) : \psi \in \mathscr{E}^\gamma\} < \infty.$$

Finally, recall that elements of $\mathscr{L}^\infty$ are equivalence classes of functions equal almost everywhere, and let $\mathscr{B} \subset \mathscr{L}^\infty$ denote the subspace of functions with bounded variation that are right continuous on $(0, 1)$.

THEOREM 3.1. *If* $p \in \text{dom } L$ *and the interiority assumption holds, then* $p \cap \mathscr{B}$ *is nonempty.*

Proof. Inserting $z = (\bar{x} - \psi, \bar{u})$ into the relation

$$C(z) + \langle p, M(z) \rangle \geqq L(p) \quad \forall z \in \mathscr{Z},$$

and utilizing the interiority hypothesis, we have

$$\sup\{\langle p, \psi' \rangle : \psi \in \mathscr{E}^\gamma\} < \infty. \tag{3.1}$$

Given $\mathscr{S} \subset \mathscr{C}^\infty$ and $f \in \mathscr{L}^1$, let us define

$$f^\dagger(\mathscr{S}) = \sup\{\langle f, \psi' \rangle : \psi \in \mathscr{S}\}.$$

Hence, $p^\dagger(\mathscr{E}^\gamma) < \infty$ by (3.1). Any $\phi \in \mathscr{D}^1$ can be written as

$$\phi = \psi + \delta$$

where δ is the linear function agreeing with $\phi(t)$ at $t = 0$ and 1, and

$$\psi = (\phi - \delta) \in \mathscr{E}^2.$$

Since $\|\delta'\|_{\mathscr{C}} \leqq 2$, it follows that

$$\langle p, \phi' \rangle \leqq \langle p, \psi' \rangle + 2\|p\|_{\mathscr{L}^1},$$

and taking the supremum over $\phi \in \mathscr{D}^1$ yields

$$p^\dagger(\mathscr{D}^1) \leqq p^\dagger(\mathscr{E}^2) + 2\|p\|_{\mathscr{L}^1} = \frac{2}{\gamma} p^\dagger(\mathscr{E}^\gamma) + 2\|p\|_{\mathscr{L}^1} < \infty.$$

The next lemma completes the proof. □

LEMMA 3.2. *If* $f \in \mathscr{L}^1$ *and* $f^\dagger(\mathscr{D}^1) < \infty$, *then* $f \cap \mathscr{B}$ *is nonempty.*

Proof. Let $\mathscr{C}$ be the space of continuous functions $y : \mathscr{T} \to R^n$ and define $\Lambda : \mathscr{C}^\infty \to R$ by $\Lambda(\phi) = \langle f, \phi' \rangle$. Since

$$\Lambda(\phi) \leqq f^\dagger(\mathscr{D}^1)\|\phi\|_{\mathscr{C}},$$

Λ can be extended to a continuous linear functional $\tilde{\Lambda} : \mathscr{C} \to R$. By the Riesz representation theorem, there exists $g \in \mathscr{B}$ such that $g(1) = 0$ and

$$\tilde{\Lambda}(\phi) = \int_0^1 \phi(t) \cdot dg(t) \quad \forall \phi \in \mathscr{C}.$$

In particular, if $\phi \in \mathscr{C}^\infty$ and $\phi(0) = 0$, integration by parts gives us

$$\tilde{\Lambda}(\phi) = -\langle g, \phi' \rangle = \Lambda(\phi) = \langle f, \phi' \rangle.$$

Therefore, $f(t) = -g(t)$ almost everywhere. □

4. Absolute continuity, I. Although there exist dual solutions with bounded variation, the following example, called the obstacle problem, shows that a continuous dual solution may not exist:

$$\text{minimize} \quad \int_{\mathscr{T}} u(t)^2 \, dt$$

subject to

$$x'(t) = u(t) \quad \text{almost everywhere,}$$
$$x(t) \geqq \alpha(t) \quad \text{for all } t \in \mathscr{T},$$
$$x(0) = x(1) = 0, \quad (x, u) \in \mathscr{Z},$$

where $\alpha \in \mathscr{A}$ is given data. It turns out that the optimal state is the profile of an elastic string lying in the (t, x) plane with ends fastened at $(0, 0)$ and $(1, 0)$ and stretched over the obstacle $\alpha(t)$; moreover, a solution to the dual problem is the *derivative* of the optimal state. Hence, a dual solution can be discontinuous when the obstacle has discontinuous derivatives.

For problems with "smooth" data, we have already established the existence of a Lipschitz continuous dual solution [15]. On the other hand, the next section refines our earlier work [19] and shows that combinations of multipliers are absolutely continuous even if the data is rough. In this section, the existence of absolutely continuous solutions is established for control constrained problems. We say that the sequence $\{\psi_k\} \subset \mathscr{L}^\infty$ *converges pointwise* to $\psi \in \mathscr{L}^\infty$ if

$$\lim_k \psi_k(t) = \psi(t) \quad \text{almost everywhere}$$

and the essential supremum of ψ_k over $\mathscr{T}$ is bounded independently of k. In particular, the sequence is called a 0-*sequence* if $\psi = 0$. A functional F defined on $\Psi \subset \mathscr{L}^\infty$ is 0-*stable* on Ψ if there exists $N < \infty$ such that

$$\overline{\lim_k} F(\psi_k) < N$$

for each 0-sequence $\{\psi_k\} \subset \Psi$. Here $\overline{\lim}_k$ is an abbreviation for $\limsup_{k \to \infty}$. Finally, let us introduce the

0-STABILITY ASSUMPTION. *For some* $\bar{z} = (\bar{x}, \bar{u}) \in \mathscr{Z}$, $C(\bar{x} + \cdot, \bar{u})$ *is* 0-*stable on* $\mathscr{E}$.

Letting $\mathscr{B}\mathscr{E} \subset \mathscr{B}$ be the subspace of functions that are continuous at $t = 0$ and 1, we have:

THEOREM 4.1. *If* $p \in \mathscr{B}\mathscr{E} \cap \operatorname{dom} L$ *and the* 0-*stability hypothesis is satisfied, then* p *is absolutely continuous.*

By Remark 1 in § 8, absolute continuity for a dual solution is also deduced from [47, Thm. 4] in some cases. In his proof of [47, Thm. 4], Rockafellar uses both an "attainability" and an "integrability" assumption. Attainability is related to, but weaker than, our boundedness condition, while integrability implies that the cost functional satisfies a growth condition, a requirement not present in our analysis.

Now let us prove the theorem. Inserting $z=(\bar{x}-\psi_k, \bar{u})$ into the relation

$$C(z)+\langle p, M(z)\rangle \geqq L(p) \quad \forall z \in \mathscr{Z},$$

and invoking the 0-stability hypothesis,

$$\overline{\lim_k} \langle p, \psi_k' \rangle \leqq N - L(p) - \langle p, M(\bar{z})\rangle \tag{4.1}$$

for all 0-sequences $\{\psi_k\} \subset \mathscr{E}$. Moreover, if c is a scalar, $\{c\psi_k\}$ is a 0-sequence satisfying (4.1). Since c is arbitrary,

$$\lim_k \langle p, \psi_k' \rangle = 0. \tag{4.2}$$

For convenience, let us assume that $n=1$ and let λ be the regular Borel measure corresponding to p [53, Thm. 8.14]. We show that λ is absolutely continuous with respect to Lebesgue measure μ; that is, $\lambda(E)=0$ for every Lebesgue measurable set E such that $\mu(E)=0$. Given a closed set $E \subset (0,1)$, it is well known [22, p. 4] that there exists a sequence $\{\psi_k\} \subset \mathscr{E}$ such that $0 \leqq \psi_k \leqq 1$ and

$$\lim_k \psi_k(x) = K_E(x)$$

for every $x \in \mathscr{T}$ where K_E is the characteristic function of E. Of course, $\{\psi_k\}$ is a 0-sequence if $\mu(E)=0$. By the dominated convergence theorem,

$$\lim_k \int_{\mathscr{T}} \psi_k(t)\, d\lambda(t) = \lambda(E).$$

On the other hand, integrating (4.2) by parts,

$$\lim_k \int_{\mathscr{T}} \psi_k(t)\, d\lambda(t) = 0 \tag{4.3}$$

for all 0-sequences $\{\psi_k\} \subset \mathscr{E}$. Since λ is a regular Borel measure, we conclude that λ is absolutely continuous with respect to μ and hence p is absolutely continuous [53, Thm. 8.16]. □

Defining the Stieltjes integral

$$\langle p, f\rangle_{\mathscr{C}} = \int_{\mathscr{T}} f(t) \cdot dp(t)$$

for $f \in \mathscr{C}$ and $p \in \mathscr{B}$, observe that (4.3) is equivalent to the statement, "$\langle p, \cdot\rangle_{\mathscr{C}}$ is 0-stable on $\mathscr{E}$," so we have:

COROLLARY 4.2. *If $p \in \mathscr{B}\mathscr{E}$ and $\langle p, \cdot\rangle_{\mathscr{C}}$ is 0-stable on $\mathscr{E}$, then p is absolutely continuous.*

5. Absolute continuity, II. Now let us characterize the feasible dual functions for state constrained problems. If $D: \mathscr{Z} \to \bar{R}$ and $\mathscr{S}$ is a set of states, we say that $(\mathscr{S}, D)$ is an *extension* of C if $C=D$ on dom C and

$$\text{dom } C = \{(x,u) \in \text{dom } D: x \in \mathscr{S}\}.$$

For example, in the obstacle problem,

$$\mathscr{S} = \{x \in \mathscr{A}: x(t) \geqq \alpha(t)\ \forall t \in \mathscr{T}\}$$

where $\mathscr{A} \subset \mathscr{B}$ is the subspace of absolutely continuous functions and

$$D(x, u) = \begin{cases} \langle u, u \rangle & \text{if } x(0) = x(1) = 0, \\ \infty & \text{otherwise} \end{cases}$$

is an extension of C.

The following theorem introduces a multiplier for $\mathscr{S}$. Defining the norm

$$\|x\|_{\mathscr{A}} = |x(0)| + \int_{\mathscr{T}} |x'(t)|\, dt \quad \text{for } x \in \mathscr{A},$$

observe that $\mathscr{A}$ and $R^n \times \mathscr{L}^1$ are isomorphic with elements in the respective spaces related by the rule

$$x \to (x(0), x').$$

Hence, any $f \in \mathscr{A}^*$, the space of bounded linear functionals on $\mathscr{A}$, can be expressed in the form

$$f(x) = c \cdot x(0) + \langle \omega, x' \rangle := \langle (c, \omega), x \rangle_{\mathscr{A}}$$

where $c \in R^n$ and $\omega \in \mathscr{L}^\infty$.

THEOREM 5.1. *Suppose that $p \in \operatorname{dom} L$, and $(\mathscr{S}, D)$ is an extension of C where $\mathscr{S}$ and D are convex, and $\mathscr{S} \subset \mathscr{A}$ has nonempty interior. Then there exists $\gamma = (c, \omega) \in R^n \times \mathscr{L}^\infty$ such that*

$$D(z) + \langle p, M(z) \rangle + \langle \gamma, x - y \rangle_{\mathscr{A}} \geqq L(p) \tag{5.1}$$

for all $z = (x, u) \in \mathscr{Z}$ and $y \in \mathscr{S}$. Furthermore, if D satisfies the 0-stability hypothesis and $(p + \omega) \in \mathscr{B}\mathscr{C}$, then $p + \omega$ is absolutely continuous.

Since $D(z) \leqq C(z)$ for all $z \in \mathscr{Z}$, (5.1) implies that

$$C(z) + \langle p, M(z) \rangle + \langle \gamma, x - y \rangle_{\mathscr{A}} \geqq L(p) \tag{5.2}$$

for all $z = (x, u) \in \mathscr{Z}$ and $y \in \mathscr{S}$. The existence of γ satisfying (5.1) follows directly from Fenchel's duality theorem [46], [28] and the fact that

$$L(p) = \inf \{D(z) + \langle p, M(z) \rangle : z = (x, u) \in \mathscr{Z}, x \in \mathscr{S}\}.$$

If D satisfies the 0-stability hypothesis and $(p + \omega) \in \mathscr{B}\mathscr{C}$, it is easy to deduce from (5.1) that $\langle p + \omega, \cdot \rangle_{\mathscr{C}}$ is 0-stable on $\mathscr{C}$. By Corollary 4.2, $p + \omega$ is absolutely continuous.

If $x: \mathscr{T} \to R^n$, we write $x \leqq 0$ if $x_i(t) \leqq 0$ for every t and i. Similarly, x is nondecreasing if $x(t) - x(s) \leqq 0$ for all $t \leqq s$. Recall that spaces like $\mathscr{C}$ and $\mathscr{A}$ consist of functions $f: \mathscr{T} \to R^n$. To denote the corresponding space of functions $f: \mathscr{T} \to R^s$, we attach the subscript s to the space.

LEMMA 5.2. *Suppose that $K: \mathscr{A} \to \mathscr{C}_s$ is convex and define the set*

$$\mathscr{S} = \{x \in \mathscr{A} : K(x) \leqq 0\}.$$

If there exists $\bar{x} \in \mathscr{A}$ such that $K(\bar{x})_i(t) < 0$ for every t and i, then for each $\gamma \in \mathscr{A}^*$, there is a nondecreasing $\nu \in \mathscr{B}_s$ such that

$$\langle \nu, K(x) \rangle_{\mathscr{C}} \geqq \inf \{\langle \gamma, x - y \rangle_{\mathscr{A}} : y \in \mathscr{S}\} \tag{5.3}$$

for all $x \in \mathscr{A}$.

Combining Theorem 5.1 and Lemma 5.2,

$$C(z) + \langle p, M(z) \rangle + \langle \nu, K(x) \rangle_{\mathscr{C}} \geqq L(p) \tag{5.4}$$

for all $z=(x, u)\in\mathscr{Z}$. To prove the lemma, we apply Theorem A.3 from Appendix 1 to the problem

$$\begin{aligned} &\text{maximize} && \langle\gamma, y\rangle_{\mathscr{A}} \\ &\text{subject to} && K(y)\leqq 0, \quad y\in\mathscr{A}. \end{aligned}$$

Hence, there exists $\nu\in\mathscr{B}_s$ satisfying (5.3) and

$$\langle\nu, f\rangle_{\mathscr{C}}\geqq 0 \tag{5.5}$$

for all nonnegative $f\in\mathscr{C}_s$. If λ is the regular Borel measure corresponding to ν, (5.5) implies that λ is positive or ν is nondecreasing. □

In the last lemma, we generated ν for given γ. Now, let us produce γ for given ν.

PROPOSITION 5.3. *Assume that $K:\mathscr{A}\to\mathscr{C}_s$ is convex and differentiable, C is convex, $\nu\in\mathscr{B}_s$ is nondecreasing, and* (5.4) *holds for some $p\in\operatorname{dom} L$. If $\hat{z}=(\hat{x}, \hat{u})\in\mathscr{Z}$ has the property that $C(\hat{z})+\langle p, M(\hat{z})\rangle=L(p)$ and*

$$\hat{x}\in\mathscr{S}:=\{x\in\mathscr{A}: K(x)\leqq 0\},$$

then the $\gamma\in\mathscr{A}^*$ defined by

$$\langle\gamma, y\rangle_{\mathscr{A}}=\langle\nu, K'[\hat{x}]y\rangle_{\mathscr{C}} \quad \forall y\in\mathscr{A}$$

satisfies (5.2).

Proof. Under our hypotheses, the functional $f(\cdot):=\langle\nu, K(\cdot)\rangle_{\mathscr{C}}$ is differentiable on $\mathscr{A}$ and

$$f'[x](y)=\langle\nu, K'[x]y\rangle_{\mathscr{C}}.$$

Since $\hat{x}\in\mathscr{S}$ and ν is nondecreasing, it also follows that $f(\hat{x})\leqq 0$. The inequality (5.4) and the relations $C(\hat{z})+\langle p, M(\hat{z})\rangle=L(p)$ and $f(\hat{x})\leqq 0$ imply that $f(\hat{x})=0$ and $\hat{z}$ minimizes $C(z)+\langle p, M(z)\rangle+f(x)$ over $z=(x, u)\in\mathscr{Z}$. Applying Lions' characterization [27, p. 12] for the minimizer of the sum of convex and differentiable functions gives us:

$$C(z)+\langle p, M(z)\rangle+f'[\hat{x}](x-\hat{x})\geqq L(p) \tag{5.6}$$

for all $z=(x, u)\in\mathscr{Z}$. Since ν is nondecreasing, f is convex and we have the standard inequality [31, p. 84]:

$$f(y)\geqq f(x)+f'[x](y-x) \tag{5.7}$$

for all $x, y\in\mathscr{A}$. Inserting $x=\hat{x}$ and recalling that $f(\hat{x})=0$, (5.7) yields

$$f'[\hat{x}](y-\hat{x})\leqq 0 \quad \forall y\in\mathscr{S}.$$

Relation (5.6) completes the proof. □

In some cases, the γ produced by Proposition 5.3 can be described more precisely. Suppose that $G(t)$ is an $s\times n$ matrix for each $t\in\mathscr{T}$, and define $Gx:\mathscr{T}\to R^s$ by

$$(Gx)(t)=G(t)x(t)$$

where $x:\mathscr{T}\to R^n$.

LEMMA 5.4. *If the elements of G are absolutely continuous and $\nu\in\mathscr{B}_s$, then*

$$\langle\nu, Gx\rangle_{\mathscr{C}}=\langle G^T\nu, x\rangle_{\mathscr{C}}-\langle G'x, \nu\rangle$$

for every $x\in\mathscr{A}$. Hence, for suitable b and $c\in R^n$, we have

$$\langle\nu, Gx\rangle_{\mathscr{C}}=c\cdot x(0)+\langle\omega, x'\rangle \quad \forall x\in\mathscr{A}$$

where

$$\omega(t) = b - G(t)^T \nu(t) + \int_0^t G'(\sigma)^T \nu(\sigma)\, d\sigma.$$

These identities are left for exercises. Given more information about the constraints, feasible dual functions can be described more precisely. For example, if the states are unconstrained at $t = 1$, then $p(1) + \omega(1) = 0$ under the hypotheses of Theorem 5.1. Of course, a dual solution may be smoother than a typical feasible element. In an earlier paper [15] we show that when the cost is strictly convex in the control and constraints are smooth enough, there exist optimal Lipschitz continuous functions p, ω, ν, x and u. Moreover, x and $p+\omega$ have Lipschitz continuous derivatives. (To be more precise, ν is only Lipschitz continuous on the open interval $(0, 1)$.)

The analysis of primal and dual solutions is different from the arguments in §§ 3–5. In [15] we start with the control minimum principal and adjoint equation, and use the implicit function theorem to estimate $|\nu(t_1) - \nu(t_2)|$ and $|u(t_1) - u(t_2)|$ in terms of smoother variables, x and $p+\omega$. Greater smoothness for ν and u implies better regularity for x and $p+\omega$. Malanowski [30] extends these results to problems with nonlinear system dynamics. Returning to the question concerning the relation between p and the costate, we show in [19] that $p+\omega$ corresponds to the usual costate.

6. Interiority. Suppose that

$$\mathcal{S} = \{x \in \mathcal{A}: x(t) \in X(t)\ \forall t \in \mathcal{T}\}$$

where $X(t) \subset R^n$ for each $t \in \mathcal{T}$. A map such as X from $\mathcal{T}$ to subsets of another space is called a *multifunction* [51]. If $X(t)$ is convex for every $t \in \mathcal{T}$, we say that X is *convex-valued.* In this section, properties of $\mathcal{S}$ are studied under the

POINTWISE INTERIORITY ASSUMPTION. *The interior of $X(t)$ is nonempty for every $t \in \mathcal{T}$ and the set*

$$\dot{X} := \{(t, x): t \in \mathcal{T}, x \in \operatorname{int} X(t)\} \subset \mathcal{T} \times R^n$$

is open.

Above, "int" denotes interior.

Rockafellar [45, Lemma 2] shows that X is lower semicontinuous when X is convex-valued and pointwise interiority holds, and in proving [45, Thm. 5], it is seen that $\mathcal{S}$ has nonempty interior. Rockafellar's development utilizes a continuous selection theorem of Michael [34, Thm. 3.2]. The fact that $\mathcal{S}$ has nonempty interior is also deduced from an appropriate partition of unity, as we now demonstrate.

The *support* of a function $f: \mathcal{T} \to R^n$ is defined by

$$\operatorname{supp} f = \text{closure}\,\{t \in \mathcal{T}: f(t) \neq 0\}.$$

Given a collection $\mathcal{O}$ of open sets whose union is $\mathcal{T}$, there exists [1, p. 51] a finite set $\Psi \subset \mathcal{C}_1^\infty$ of nonnegative functions such that

$$\sum_{\psi \in \Psi} \psi(t) = 1 \quad \forall t \in \mathcal{T},$$

and for every $\psi \in \Psi$,

$$\operatorname{supp} \psi \subset U$$

for some $U \in \mathcal{O}$. The set Ψ is called an *infinitely differentiable partition of unity* subordinate to $\mathcal{O}$. Defining the set

$$\dot{\mathcal{S}} = \{x \in \mathcal{C}^\infty : (t, x(t)) \in \dot{X} \ \forall t \in \mathcal{T}\},$$

we have:

LEMMA 6.1. *If X is convex-valued and pointwise interiority holds, then $\dot{\mathcal{S}}$ is nonempty.*

Proof. Given $f: \mathcal{T} \to R^n$, define

$$\mathcal{O}_f = \{t \in \mathcal{T} : f(t) \in \operatorname{int} X(t)\}.$$

By pointwise interiority, $\mathcal{O}_f$ is open when f is continuous. If $\mathcal{F} \subset \mathcal{C}^\infty$ is the collection of constant functions, then

$$\mathcal{T} = \bigcup_{f \in \mathcal{F}} \mathcal{O}_f.$$

Let Ψ be a partition of unity subordinate to $\{\mathcal{O}_f : f \in \mathcal{F}\}$. For each $\psi \in \Psi$, there exists $f(\psi) \in \mathcal{F}$ such that

$$\operatorname{supp} \psi \subset \mathcal{O}_{f(\psi)}.$$

Observe that $x \in \mathcal{C}^\infty$ given by

$$x(t) = \sum_{\psi \in \Psi} f(\psi)\psi(t)$$

is a convex combination of points in the interior of $X(t)$ for every $t \in \mathcal{T}$. □

LEMMA 6.2. *If pointwise interiority holds, $x \in \mathcal{C}$, and $x(t) \in \operatorname{int} X(t)$ for each $t \in \mathcal{T}$, then there exists $\rho > 0$ such that*

$$\{y \in R^n : |y - x(t)| \leqq \rho\} \subset X(t)$$

for every $t \in \mathcal{T}$.

Proof. Since $\dot{X}^c$, the complement of $\dot{X}$, and $\{(t, x(t)) : t \in \mathcal{T}\}$ are disjoint closed sets, the distance between them is positive. □

Lemmas 6.1 and 6.2 and the inequality

$$\|x\|_{\mathcal{C}} \leqq \|x\|_{\mathcal{A}} \quad \forall x \in \mathcal{A}$$

imply that $\mathcal{S}$ has nonempty interior when X is convex-valued and pointwise interiority holds. Defining the set

$$\mathcal{S}^\infty = \{x \in \mathcal{L}^\infty : x(t) \in X(t) \text{ almost everywhere}\},$$

we have:

THEOREM 6.3. *If pointwise interiority holds and X is convex-valued, then for each $x \in \mathcal{S}^\infty$, there exists a sequence $\{x_k\} \subset \dot{\mathcal{S}}$ converging pointwise to x. Moreover, for any finite set $\{(t_j, a_j)\} \subset \dot{X}$ where the t_j are distinct, it can be arranged so that $x_k(t_j) = a_j$ for every j and k.* (Pointwise convergence is defined in § 4.)

Proof. Given $x \in \mathcal{L}^\infty$ and $\varepsilon > 0$, we exhibit $w \in \dot{\mathcal{S}}$ such that

$$\mu\{t \in \mathcal{T} : |w(t) - x(t)| > \varepsilon\} \leqq \varepsilon \tag{6.1}$$

where μ is Lebesgue measure and $\|w\|_{\mathcal{C}}$ is bounded independently of ε. To simplify notation, let x also denote a particular element in its equivalence class for which

$$x(t) \in X(t) \quad \forall t \in \mathcal{T}$$

and $\|x\|_{\mathscr{C}}$ is finite. By Lusin's theorem and regularity properties of Borel measure [53, Thms. 2.23 and 2.17], there exist $y \in \mathscr{C}$ and a closed set $K \subset \mathscr{T}$ such that $\mu(K^c) \leqq \varepsilon$, $\|y\|_{\mathscr{C}} \leqq \|x\|_{\mathscr{C}}$, and

$$x(t) = y(t) \quad \forall t \in K.$$

Recalling Lemmas 6.1 and 6.2 and the fact that $\mathscr{C}^\infty$ is a dense subset of $\mathscr{C}$, there is $z \in \mathscr{C}^\infty$ such that

$$\|y - z\|_{\mathscr{C}} \leqq \varepsilon,$$

and $z(t) \in \operatorname{int} X(t)$ for every $t \in K$. By pointwise interiority, the set

$$\mathscr{O} = \{t \in \mathscr{T}: z(t) \in \operatorname{int} X(t)\}$$

is open. Let $\{\psi_1, \psi_2\}$ be a partition of unity subordinate to $\{\mathscr{O}, K^c\}$ and define

$$w(t) = \psi_1(t) z(t) + \psi_2(t) \dot{x}(t)$$

where $\dot{x} \in \dot{\mathscr{S}}$. Since $z(t) \in \operatorname{int} X(t)$ on $\operatorname{supp} \psi_1$, and $\psi_1 + \psi_2$ is identically 1, it follows that $w \in \dot{\mathscr{S}}$. Since $\psi_1 = 1$ on K, (6.1) is established.

Next, given $(\sigma, a) \in \dot{X}$, let $\mathscr{O} \subset \mathscr{T}$ be an open interval containing σ such that $\mu(\mathscr{O}) \leqq \varepsilon$ and

$$\mathscr{O} \times \{a\} \subset \dot{X}.$$

Letting $\{\phi_1, \phi_2\}$ be a partition of unity subordinate to $\{\mathscr{O}, \{\sigma\}^c\}$, define

$$v(t) = a\phi_1(t) + w(t)\phi_2(t).$$

Observe that $v \in \dot{\mathscr{S}}$, $v(\sigma) = a$, and $v(t) = w(t)$ except on a set of measure $\leqq \varepsilon$. The second part of the theorem follows almost immediately. □

7. Pointwise minimization. The next section provides a convenient representation for the dual functional when the cost and the constraints assume a special form. Here we review some theorems on measurability, drawing on Rockafellar's work [51], and develop preliminary results. Let us consider the following problem:

$$\inf \{I(x): x \in \mathscr{L}^\infty\}$$

where $I: \mathscr{L}^\infty \to \bar{R}$ is defined by

$$I(x) = \int_{\mathscr{T}} f(x(t), t)\, dt$$

for some $f: R^n \times \mathscr{T} \to \bar{R}$. We assume that I is proper, and the integrand is measurable and majorizes a summable function whenever $x \in \mathscr{L}^\infty$.

Classically, $f(x(\cdot), \cdot)$ is measurable when $x(\cdot)$ is measurable if the *Carathéodory conditions* hold; that is, $f(\cdot, t)$ is continuous for each fixed $t \in \mathscr{T}$ and $f(x, \cdot)$ is measurable for each fixed $x \in R^n$. On the other hand, we may wish to embed constraints in the cost functional. For example, the constraint

$$x(t) \in X(t)$$

almost everywhere can be incorporated in the cost through the definition

$$f(x, t) = \infty \quad \text{if } x \notin X(t).$$

The *normal integrand*, introduced by Rockafellar [51], is a natural one-sided extension of the Carathéodory integrand. The integrand $f: R^n \times \mathscr{T} \to \bar{R}$ is normal if $f(x, t)$ is

lower semicontinuous in x for each fixed $t \in \mathcal{T}$, and f is measurable on $R^n \times \mathcal{T}$ with respect to the σ-algebra generated by products of Borel sets in R^n and Lebesgue sets in $\mathcal{T}$. Therefore, it follows that $f(x(\cdot), \cdot)$ is measurable whenever $x(\cdot)$ is measurable. An important property of normal integrands is contained in the following lemma, an immediate consequence of [51, Thm. 2K]:

LEMMA 7.1. *If f is a normal integrand on $R^n \times \mathcal{T}$, then*

$$\inf\{I(x): x \in \mathcal{L}^\infty\} = \int_{\mathcal{T}} \inf\{f(x, t): x \in R^n\}\, dt$$

and the integrand above is measurable. Furthermore, there exists a measurable function $x: \mathcal{T} \to R^n$ such that

$$x(t) \in \arg\min\{f(x, t): x \in R^n\}$$

wherever the minimum is attained.

As noted earlier, constraints can be embedded in the cost. Given $X: \mathcal{T} \to 2^{R^n}$, let us define

$$\bar{f}(x, t) = \begin{cases} f(x, t) & \text{if } x \in X(t), \\ \infty & \text{otherwise.} \end{cases}$$

By [51, Props., 2H and 2L], $\bar{f}$ is normal provided f is normal and X is *closed-valued* and *measurable*; that is, $X(t)$ is closed for each $t \in \mathcal{T}$ and for all closed sets $K \subset R^n$,

$$\{t \in \mathcal{T}: X(t) \cap K \text{ is nonempty}\}$$

is measurable. If X is closed-valued and convex-valued, then X is measurable under the pointwise interiority hypothesis. This follows from Theorem 6.3 and Castaing's characterization of a closed-valued measurable multifunction in terms of the closure of a countable collection of measurable functions [7], [51].

Now consider the problem

$$\hat{C} = \inf\{E(x) + I(x): x \in \mathcal{C}^\infty\}$$

where $E: \mathcal{C}^\infty \to \bar{R}$ and for some finite set $\Omega \subset \mathcal{T}$,

$$E(x) = E(y)$$

whenever $x, y \in \mathcal{C}^\infty$ and $x(t) = y(t)$ for each $t \in \Omega$. For example, $E(x)$ might be expressed in terms of $x(1)$. Let us define

$$X(t) = \{x \in R^n. f(x, t) < \infty\},$$

and let $\dot{\mathcal{S}}$ and $\mathcal{S}^\infty$ be the sets defined in § 6.

LEMMA 7.2. *Suppose that X is convex-valued, pointwise interiority holds,*

$$\inf\{E(x): x \in \dot{\mathcal{S}}\} = \inf\{E(x): x \in \mathcal{S}^\infty \cap \mathcal{C}^\infty\}, \tag{7.1}$$

and

$$I(x) = \lim_k I(x^k) \tag{7.2}$$

for each sequence $\{x^k\} \subset \mathcal{S}^\infty$ converging pointwise to some $x \in \operatorname{dom} I$. Then

$$\hat{C} = \inf\{E(x): x \in \dot{\mathcal{S}}\} + \inf\{I(x): x \in \mathcal{L}^\infty\}.$$

Proof. Since $I(x) = \infty$ if $x \notin \mathcal{S}^\infty$,

$$\hat{C} = \inf\{E(x) + I(x): x \in \mathcal{S}^\infty \cap \mathcal{C}^\infty\}.$$

Given $x \in \mathscr{S}^\infty$ and $y \in \dot{\mathscr{S}}$, Theorem 6.3 provides a sequence $\{x^k\} \subset \dot{\mathscr{S}}$ converging pointwise to x and

$$x^k(t) = y(t) \quad \forall t \in \Omega.$$

If $x \in \operatorname{dom} I$,

$$E(y) + I(x) = \lim_k \{E(x^k) + I(x^k)\} \geqq \inf\{E(x) + I(x): x \in \dot{\mathscr{S}}\}. \tag{7.3}$$

Combining (7.1) and (7.3),

$$\begin{aligned} \inf\{E(y): y \in \mathscr{S}^\infty \cap \mathscr{C}^\infty\} + \inf\{I(x): x \in \mathscr{S}^\infty\} & \\ = \inf\{E(y): y \in \dot{\mathscr{S}}\} + \inf\{I(x): x \in \mathscr{S}^\infty\} & \\ \geqq \inf\{E(x) + I(x): x \in \dot{\mathscr{S}}\} & \\ \geqq \inf\{E(x) + I(x): x \in \mathscr{S}^\infty \cap \mathscr{C}^\infty\} = \hat{C}. & \end{aligned}$$

Since the reverse inequalities are trivial, the proof is complete. □

If $E(y) = e(y(1))$ where $e: R^n \to \bar{R}$, then (7.1) is satisfied if $\operatorname{dom} e \subset \operatorname{int} X(1)$. Moreover, under these hypotheses,

$$\inf\{E(x): x \in \dot{\mathscr{S}}\} = \inf\{e(a): a \in R^n\}.$$

Relation (7.2) holds if $f(\cdot, t)$ is continuous on $X(t)$ and if for each $\rho > 0$ there is a summable function $g: \mathscr{T} \to R$ such that

$$g(t) \geqq |f(x, t)|$$

whenever $x \in X(t)$ and $|x| \leqq \rho$. If f is a normal integrand on $R^n \times \mathscr{T}$, Lemma 7.1 gives us

$$\inf\{I(x): x \in \mathscr{L}^\infty\} = \int_{\mathscr{T}} \inf\{f(x, t): x \in R^n\}\, dt.$$

8. Dual formulations. Let us evaluate the dual functional when the primal cost has the form

$$C(x, u) = e(x(0), x(1)) + \int_{\mathscr{T}} f(x(t), u(t), t)\, dt$$

where $e: R^{2n} \to \bar{R}$ and $f: R^{n+m} \times \mathscr{T} \to \bar{R}$ is a normal integrand which majorizes a summable function whenever x is essentially bounded and u is summable, and the integral is finite for some $(x, u) \in \mathscr{L}^\infty_n \times \mathscr{L}^\infty_m$. We define

$$H(a, q, z) = e(x(0), x(1)) - a_0 \cdot x(0) - a_1 \cdot x(1) + \int_{\mathscr{T}} [f(z(t), t) - q(t) \cdot z(t)]\, dt$$

where $z = (x, u) \in \mathscr{L}^\infty_n \times \mathscr{L}^1_m$, $q \in \mathscr{L}^1_n \times \mathscr{L}^\infty_m$, and $a = (a_0, a_1) \in R^n \times R^n$. Corresponding to e and f, we have the conjugate functions $e^*(a) = \inf\{e(b) - a \cdot b: b \in R^{n+m}\}$ and

$$f^*(q) = \int_{\mathscr{T}} \inf\{f(z, t) - q(t) \cdot z: z \in R^{n+m}\}\, dt.$$

The integrand of f^* is measurable by Lemma 7.1 and the fact that the sum of normal and Carathéodory integrands is normal [51, Prop. 2M]. By these definitions, the following inequalities are clearly satisfied:

$$e^*(a) + f^*(q) \leqq \inf\{H(a, q, z): z \in \mathscr{Z}\} \leqq \inf\{H(a, q, x, u): x \in \mathscr{C}^\infty, u \in \mathscr{L}^\infty_m\}.$$

Now set $E(x) = e(x(0), x(1)) - a_0 \cdot x(0) - a_1 \cdot x(1)$ and for fixed $u \in \mathscr{L}_m^\infty$ define

$$I(x) = \int_{\mathscr{T}} f(x(t), u(t), t)\, dt.$$

We assume that for each fixed $u \in \mathscr{L}_m^\infty$ where the domain of I in $\mathscr{L}^\infty$ is nonempty, there exists an element of u's equivalence class such that the hypotheses of Lemma 7.2 are satisfied. Referring to the discussion after Lemma 7.2, it is also assumed that for this element of u's equivalence class, we have the identity $\inf\{E(x)\colon x \in \hat{\mathscr{S}}\} = e^*(a)$. Then Lemmas 7.1 and 7.2 give us:

$$\begin{aligned}\inf\{H(a, q, x, u)\colon x \in \mathscr{C}^\infty, u \in \mathscr{L}_m^\infty\} &= e^*(a) + \inf\{H(a, q, x, u)\colon x \in \mathscr{L}_n^\infty, u \in \mathscr{L}_m^\infty\} \\ &= e^*(a) + f^*(q).\end{aligned}$$

Combining these relations, it follows that

$$e^*(a) + f^*(q) = \inf\{H(a, q, z)\colon z \in \mathscr{Z}\}.$$

We say that (P) has a *pointwise representation* if this equality holds for every $a \in R^{2n}$ and $q \in \mathscr{L}_n^1 \times \mathscr{L}_m^\infty$.

LEMMA 8.1. *If* (P) *has a pointwise representation, then for all* $p \in \mathscr{A}$,

$$L(p) = e^*(a) + f^*(q) \tag{8.1}$$

where

$$q(t) = \begin{pmatrix} p'(t) + A(t)^T p(t) \\ B(t)^T p(t) \end{pmatrix} \tag{8.2}$$

and

$$a = \begin{pmatrix} p(0) \\ -p(1) \end{pmatrix}.$$

Proof. Starting with the definition of L and integrating by parts,

$$L(p) = \inf\{H(a, q, z)\colon z \in \mathscr{Z}\}$$

where a and q are given above. The conclusion follows immediately. □

Remark 1. Rockafellar [47] uses (8.1) to define $L(p)$ when p is absolutely continuous. Since the dual solution may be discontinuous, he shows that the dual function can be extended to the space of functions with bounded variation.

We now examine four problems which will be solved in § 11.

Problem I.

$$\text{minimize} \quad \frac{1}{2}\int_0^1 [x(t)^2 + u(t)^2]\, dt$$

subject to

$$x'(t) = u(t), \quad u(t) \leqq a \quad \text{almost everywhere},$$
$$x(0) = c, \quad (x, u) \in \mathscr{Z}.$$

Here a and c are given scalars. Defining the functions $f\colon R^2 \times \mathscr{T} \to \bar{R}$ and $e\colon R^2 \to \bar{R}$ by

$$f(x, u, t) = \begin{cases} \frac{1}{2}(x^2 + u^2) & \text{if } u \leqq a, \\ \infty & \text{if } u > a \end{cases}$$

and

$$e(x, y) = \begin{cases} 0 & \text{if } x = c, \\ \infty & \text{if } x \neq c, \end{cases}$$

we can write Problem I as

$$\text{minimize} \quad e(x(0), x(1)) + \int_{\mathscr{T}} f(x(t), u(t), t)\, dt$$

$$\text{subject to} \quad x'(t) = u(t) \quad \text{almost everywhere.}$$

If a_0 and a_1 are given scalars and u is a fixed element of $\mathscr{L}^\infty$, let us define

$$E(x) = e(x(0), x(1)) - a_0 \cdot x(0) - a_1 \cdot x(1),$$

and

$$I(x) = \int_{\mathscr{T}} f(x(t), u(t), t)\, dt.$$

If the domain of I is nonempty, then there exists an element of u's equivalence class such that $X(t) = R$ for every $t \in \mathscr{T}$ where $X(t)$ is introduced in § 7. Hence (7.1) and the pointwise interiority assumption are satisfied trivially. Likewise, (7.2) holds since $f(\cdot, \cdot, t)$ is continuous on its effective domain. Finally, it is easy to check that $\inf\{E(x): x \in \mathscr{S}\} = e^*(a)$. Therefore, by the discussion at the start of this section, Problem I has a pointwise representation, and by Lemma 8.1, the dual function is

$$L(p) = e^*(p(0), -p(1)) + f^*(p', p)$$

for every $p \in \mathscr{A}$. The conjugate functions e^* and f^* are easily evaluated:

$$e^*(x, y) = \begin{cases} -cx & \text{if } y = 0, \\ -\infty & \text{if } y \neq 0, \end{cases}$$

$$f^*(p', p) = -\int_{\mathscr{T}} l(p'(t), p(t), t)\, dt,$$

$$l(x, y, t) = \begin{cases} \frac{1}{2}[x^2 + y^2] & \text{if } y \leqq a, \\ \frac{1}{2}[x^2 + a(2y - a)] & \text{if } y > a. \end{cases}$$

Although the dual problem is to maximize $L(p)$ over $p \in \mathscr{L}^\infty$, Theorems 3.1 and 4.1 tell us that we only need consider $p \in \mathscr{A}$. Since $e^*(x, y) = -\infty$ when $y \neq 0$, we can also impose the explicit dual constraint $p(1) = 0$. In summary, the dual of Problem I can be written

$$\text{maximize} \quad -\left\{ cp(0) + \int_{\mathscr{T}} l(p'(t), p(t), t)\, dt \right\}$$

$$\text{subject to} \quad p(1) = 0, \quad p \in \mathscr{A}.$$

Next let us consider

Problem II.

$$\text{minimize} \quad \frac{1}{2} \int_0^1 [x(t)^2 + u(t)^2]\, dt$$

subject to

$$x'(t) = u(t), \; u(t) \leqq a \quad \text{almost everywhere,}$$

$$x(t) \leqq b \quad \text{for all } t \in \mathscr{T}, \; x(0) = c, \; (x, u) \in \mathscr{Z}$$

where $c < b$. Again, defining the functions $f: R^2 \times \mathcal{T} \to \bar{R}$ and $e: R^2 \to \bar{R}$ by

$$f(x, u, t) = \begin{cases} \frac{1}{2}(x^2 + u^2) & \text{if } u \leqq a \text{ and } x \leqq b, \\ \infty & \text{if } u > a \text{ or } x > b \end{cases}$$

and

$$e(x, y) = \begin{cases} 0 & \text{if } x = c \text{ and } y \leqq b, \\ \infty & \text{if } x \neq c \text{ or } y > b, \end{cases}$$

we can cast Problem II in the form

$$\text{minimize} \quad e(x(0), \dot{x}(1)) + \int_{\mathcal{T}} f(x(t), u(t), t)\, dt$$

$$\text{subject to} \quad x'(t) = u(t) \quad \text{almost everywhere.}$$

Let us define E and I as we did for Problem I. If the domain of I is nonempty, then there exists an element of u's equivalence class such that $X(t) = \{x \in R: x \leqq b\}$ for every $t \in \mathcal{T}$. Since $\dot{X} = \{(t, x) \in \mathcal{T} \times R: x < b\}$ is an open subset of $\mathcal{T} \times R$, the pointwise interiority assumption holds. To verify (7.1), suppose that $x \in \mathcal{C}^\infty$, $x(0) = c$, and $x(t) \leqq b$ for each $t \in \mathcal{T}$. Then the sequence $\{x_k\}$ defined by

$$x_k(t) = x(t) - \frac{t}{k}$$

lies in $\dot{\mathcal{S}}$ and $\lim_k E(x_k) = E(x)$. Hence (7.1) holds. Since $f(\cdot, \cdot, t)$ is continuous on its effective domain, (7.2) is satisfied. Again, it is easy to see that $\inf\{E(x): x \in \dot{\mathcal{S}}\} = e^*(a)$. By the discussion at the start of this section, Problem II has a pointwise representation. Applying Lemma 8.1, the dual function can be expressed

$$L(p) = e^*(p(0), -p(1)) + f^*(p', p)$$

for each $p \in \mathcal{A}$ where

$$e^*(x, y) = \begin{cases} -(cx + by) & \text{if } y \geqq 0, \\ -\infty & \text{if } y < 0, \end{cases}$$

$$f^*(p', p) = -\int_{\mathcal{T}} [l_x(p'(t), p(t), t) + l_u(p'(t), p(t), t)]\, dt,$$

$$l_x(x, y, t) = \begin{cases} \frac{1}{2}x^2 & \text{if } x \leqq b, \\ \frac{1}{2}b(2x - b) & \text{if } x > b, \end{cases}$$

$$l_u(x, y, t) = \begin{cases} \frac{1}{2}y^2 & \text{if } y \leqq a, \\ \frac{1}{2}a(2y - a) & \text{if } y > a. \end{cases}$$

Although the dual maximization is over $p \in \mathcal{L}^\infty$, it follows from our regularity analysis [15] that there exists a Lipschitz continuous dual solution to Problem II. Consequently, the dual problem reduces to

$$\text{maximize} \quad -\{cp(0) - bp(1) + \int_{\mathcal{T}} [l_x(p'(t), p(t), t) + l_u(p'(t), p(t), t)]\, dt\}$$

$$\text{subject to} \quad p(1) \leqq 0 \quad p \in \mathcal{A}.$$

The derivation of the dual for the final two examples is similar to Problems I and II so we just summarize the conclusions. The primal version of the next problem is found in [24] and [33].

Problem III.

$$\text{minimize} \quad \int_0^1 [x_1(t)^2 + x_2(t)^2 + .005u(t)^2]\, dt$$

$$\text{subject to} \quad x_1'(t) = x_2(t), \quad x_2'(t) = -x_2(t) + u(t) \quad \text{almost everywhere},$$
$$x_1(0) = 0, \quad x_2(0) = -1, \quad (x_1, x_2, u) \in \mathscr{Z}.$$

In addition, two different state constraints are considered:

Case A. $x_2(t) \leqq \alpha(t)$ for all $t \in \mathscr{T}$,

Case B. $x_1(t) \leqq \alpha(t)$ for all $t \in \mathscr{T}$

where $\alpha(t) = 2(1-2t)^2 - \frac{1}{2}$ (see Figs. 1 and 2). In Case A the dual is

$$\text{maximize} \quad \left\{ p_2(0) + \tfrac{3}{2}p_2(1) - \int_{\mathscr{T}} l(p'(t), p(t), t)\, dt \right\}$$

$$\text{subject to} \quad p_1(1) = 0, \quad p_2(1) \leqq 0, \quad p = (p_1, p_2) \in \mathscr{A}$$

where

$$l(w, x, y, z, t) = \begin{cases} 50z^2 + \frac{1}{4}(w^2 + \beta^2) & \text{if } \beta \leqq 2\alpha(t), \\ 50z^2 + \frac{1}{4}w^2 + \alpha(t)(\beta - \alpha(t)) & \text{if } \beta > 2\alpha(t) \end{cases}$$

and $\beta = x + y - z$. In Case B the dual is

$$\text{maximize} \quad \left\{ p_2(0) + \tfrac{3}{2}p_1(1) - \int_{\mathscr{T}} l(p'(t), p(t), t)\, dt \right\}$$

$$\text{subject to} \quad p_1(1) \leqq 0, \quad p_2(1) = 0, \quad p = (p_1, p_2) \in \mathscr{A}$$

where

$$l(w, x, y, z, t) = \begin{cases} 50z^2 + \frac{1}{4}(w^2 + \beta^2) & \text{if } w \leqq 2\alpha(t), \\ 50z^2 + \frac{1}{4}\beta^2 + \alpha(t)(w - \alpha(t)) & \text{if } w > 2\alpha(t) \end{cases}$$

and $\beta = x + y - z$.

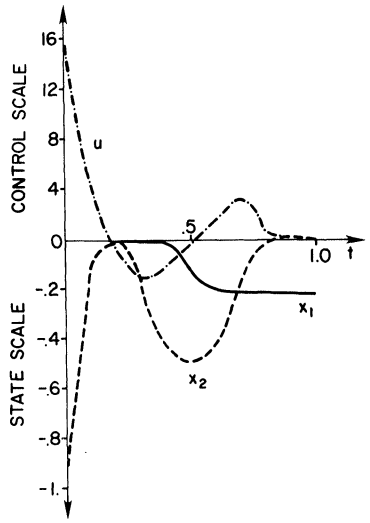

FIG. 1. *Solution to Problem* IIIA.

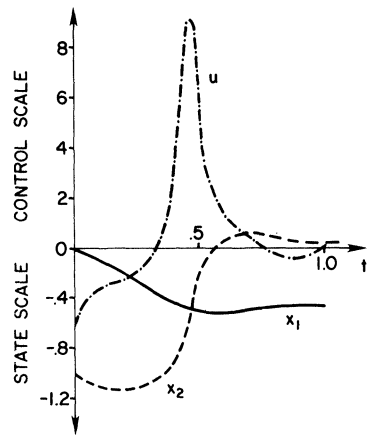

FIG. 2. *Solution to Problem* IIIB.

The primal version of the following problem is found in [56] (see Fig. 3).

Problem IV.

$$\text{minimize} \quad \frac{1}{2}\left\{x_2(1)^2+\int_0^1 [x_1(t)^2+u(t)^2]\,dt\right\}$$

$$\text{subject to} \quad x_1'(t)=x_2(t), \quad x_2'(t)=u(t) \quad \text{almost everywhere},$$

$$x_1(0)=-1, \quad x_2(0)=0,$$

$$x_2(t)\leqq \tfrac{1}{20} \quad \text{for all } t\in\mathcal{T}, \quad (x_1, x_2, u)\in\mathcal{X}.$$

The dual is

$$\text{maximize} \quad -\left\{\phi(p_2(1))-p_1(0)+\int_{\mathcal{T}} l(p'(t), p(t), t)\,dt\right\}$$

$$\text{subject to} \quad p_1(1)=0, \quad p_1(t)+p_2'(t)\geqq 0 \quad \text{almost everywhere},$$

$$p=(p_1, p_2)\in\mathcal{A}$$

where

$$l(w, x, y, z, t)=\tfrac{1}{2}[w^2+z^2+.05(x+y)],$$

$$\phi(x)=\begin{cases}\frac{1}{2}x^2 & \text{if } x\geqq -\frac{1}{20},\\ -\frac{1}{20}(x+\frac{1}{40}) & \text{if } x<-\frac{1}{20}.\end{cases}$$

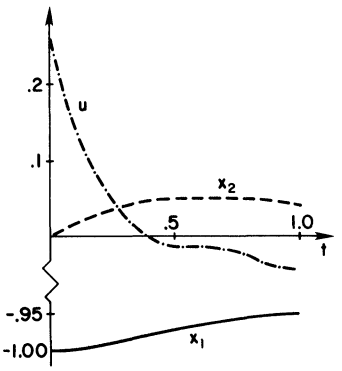

FIG. 3. *Solution to Problem* IV.

To conclude this section, let us examine the relations between solutions to the primal and the dual problems. If q is related to $p\in\mathcal{A}$ through (8.2), and $z\colon\mathcal{T}\to R^{n+m}$ is a measurable function such that

$$f(z(t), t)-q(t)\cdot z(t)=\min\{f(z, t)-q(t)\cdot z\colon z\in R^{n+m}\}$$

almost everywhere, we say that (p, z) is a *min-pair.*

THEOREM 8.2. *If $p\in\mathcal{L}^\infty$, $z=(x, u)$ is feasible in* (P), *and $L(p)=C(z)$, then p is optimal in the dual problem and z is optimal in the primal problem. Moreover, if $p\in\mathcal{A}$ and* (P) *has a pointwise representation, then (p, z) is a min-pair and*

$$e^*(p(0), -p(1))=e(x(0), x(1))-\begin{pmatrix}p(0)\\ -p(1)\end{pmatrix}\cdot\begin{pmatrix}x(0)\\ x(1)\end{pmatrix}. \tag{8.3}$$

Proof. The first part of the theorem, the optimality of p and z, is a standard property of the dual functional. Let us consider the second half. Since $M(z)=0$ and (P) has a pointwise representation, Lemma 8.1 gives us:

$$e^*(a)+f^*(q)=L(p)=C(z)=C(z)+\langle p, M(z)\rangle=H(a, q, z)$$

where the last equality comes from integrating by parts. Examining the definition of H, (p, z) in a min-pair and (8.3) is satisfied. □

If z is optimal in the primal problem, p is optimal in the dual problem, and $L(p)=C(z)$, then we say that (p, z) is an *optimal pair.* Hence the preceding theorem states that an optimal pair (p, z) with $p\in\mathscr{A}$ is a min-pair when (P) has a pointwise representation. For p and $\omega\in\mathscr{A}$, let us define

$$q^{\omega}(t)=q(t)+\begin{pmatrix}\omega'(t)\\ 0\end{pmatrix}$$

where q is given by (8.2).

THEOREM 8.3. *If* (P) *has a pointwise representation, p and* $\omega\in\mathscr{A}$, *and*

$$z=\arg\min\{C(\xi)+\langle p, M(\xi)\rangle+\langle(c, \omega), x\rangle_{\mathscr{A}}: \xi=(x, u)\in\mathscr{Z}\} \tag{8.4}$$

for some $c\in R^n$, *then*

$$f(z(t), t)-q^{\omega}(t)\cdot z(t)=\inf\{f(\xi, t)-q^{\omega}(t)\cdot\xi: \xi\in R^{n+m}\} \tag{8.5}$$

almost everywhere.

Proof. Since (P) has a pointwise representation, we integrate by parts to get

$$\min\{C(\xi)+\langle p, M(\xi)\rangle+\langle(c, \omega), x\rangle_{\mathscr{A}}: \xi=(x, u)\in\mathscr{Z}\}=e^*(a)+f^*(q^{\omega})$$

for some $a\in R^{2n}$. Since z attains the minimum, (8.5) holds. □

9. Fundamental inequalities. Observe that estimating the error in z_h, the approximation to a primal solution introduced in § 2, is essentially a parametric programming problem in the parameter $p\in\mathscr{L}^{\infty}$. Defining

$$\Omega(p)=\{z\in\mathscr{Z}: L(p)=C(z)+\langle p, M(z)\rangle\},$$

we hope that $\Omega(p)$ approaches a primal solution as p approaches a dual solution. Fiacco and Hutzler [11] and Guddat [13] give good surveys of recent work on parametric programs. Exploiting the structure of the dual functional, we now obtain an estimate for the $\mathscr{L}^2$ error in z_h when the cost is strictly convex.

First, let us consider parametric programs in finite dimensions. We say that a functional $h: R^n\to\bar{R}$ is *uniformly convex* if h is convex and there is an $\alpha>0$ with the following property: For each $x\in\operatorname{dom} h$, there exists $z\in R^n$ such that

$$h(w+x)\geqq h(x)+z\cdot w+\alpha|w|^2\quad\forall w\in R^n \tag{9.1}$$

and

$$\lim_{s\downarrow 0}\frac{h(x+sw)-h(x)}{s}=z\cdot w \tag{9.2}$$

whenever $(x+w)\in\operatorname{dom} h$. The scalar α is called the *modulus of convexity*, and we let $h'(x)$ denote any z satisfying the relations above. Suppose that $\{g_{\lambda}: \lambda\in\Lambda\}$ is a collection of lower semicontinuous proper functions where $g_{\lambda}: R^n\to\bar{R}$ is uniformly convex with

modulus of convexity α independent of $\lambda \in \Lambda$. Under these hypotheses, there is a unique $\xi(\lambda) \in R^n$ for which

$$g_\lambda(\xi(\lambda)) = \inf\{g_\lambda(\xi): \xi \in R^n\}$$

whenever $\lambda \in \Lambda$.

LEMMA 9.1. *Assume that λ and $\mu \in \Lambda$. If $x := \xi(\lambda)$, then*

$$\alpha|x-y|^2 \leqq g_\lambda(y) - g_\lambda(x) \tag{9.3}$$

for every $y \in R^n$. Conversely, if $y := \xi(\mu) \in \operatorname{dom} g_\lambda$, then

$$g_\lambda(y) - g_\lambda(x) \leqq \frac{1}{4\alpha}|g'_\mu(y) - g'_\lambda(y)|^2 \tag{9.4}$$

for each $x \in \operatorname{dom} g_\mu$.

Proof. Taking $w = y - x$, (9.1) implies that

$$g_\lambda(y) - g_\lambda(x) \geqq g'_\lambda(x) \cdot w + \alpha|w|^2. \tag{9.5}$$

Let us assume that $y \in \operatorname{dom} g_\lambda$ since (9.3) is trivial otherwise. Recalling that x minimizes $g_\lambda(\cdot)$, we have the standard inequality [28, p. 178]:

$$g'_\lambda(x) \cdot (y - x) \geqq 0.$$

Hence (9.3) follows from (9.5).

Now consider (9.4). Since y minimizes $g_\mu(\cdot)$ and $x \in \operatorname{dom} g_\mu$, we also have the relation

$$g'_\mu(y)(x - y) \geqq 0,$$

or equivalently,

$$g'_\lambda(y) \cdot w \leqq (g'_\lambda(y) - g'_\mu(y)) \cdot w \tag{9.6}$$

where $w = y - x$. Interchanging y and x in (9.5) and combining with (9.6) gives us

$$g_\lambda(y) - g_\lambda(x) \leqq g'_\lambda(y) \cdot w - \alpha|w|^2 \leqq (g'_\lambda(y) - g'_\mu(y)) \cdot w - \alpha|w|^2.$$

Finally, utilizing the inequality

$$a \cdot b \leqq \frac{1}{4\alpha}|a|^2 + \alpha|b|^2,$$

we get (9.4). □

Let us return to the cost functional defined at the start of § 8, and impose the following condition on the integrand:

UNIFORM CONVEXITY ASSUMPTION. *For each $t \in \mathcal{T}$, $f(\cdot, t)$ is uniformly convex with modulus of convexity α independent of t.*

We define a function $g: R^{2(n+m)} \times \mathcal{T} \to \bar{R}$ by the rule

$$g(\xi, \lambda, t) = f(\xi, t) - \lambda \cdot \xi. \tag{9.7}$$

Lemma 7.1 and the uniform convexity hypothesis imply that for each $q \in \mathcal{L}^1_{n+m}$, there is a measurable function $z: \mathcal{T} \to R^{n+m}$ such that

$$g(z(t), q(t), t) = \min\{g(z, q(t), t): z \in R^{n+m}\}$$

almost everywhere. If $\|\cdot\|$ denotes the $\mathscr{L}^2$ norm defined by

$$\|z\| = \langle z, z\rangle^{1/2},$$

we have:

THEOREM 9.2. *Suppose that* (P) *has a pointwise representation, the uniform convexity hypothesis is satisfied, and* (p, z) *is an optimal pair where* $p \in \mathscr{A}$. *Then*

$$\alpha\|z - z_h\|^2 \leqq L(p) - L(p_h)$$

for all min-pairs (p_h, z_h).

Proof. Since (P) has a pointwise representation, Lemma 8.1 yields

$$L(p) = e^*(a) + f^*(q) \quad \text{and} \quad L(p_h) = e^*(a_h) + f^*(q_h).$$

Holding t fixed, we apply Lemma 9.1 to g from (9.7) taking $\lambda = q_h(t)$ and $\mu = q(t)$. Integrating (9.3) over $\mathscr{T}$ and utilizing Theorem 8.2,

$$\begin{aligned}\alpha\|z - z_h\|^2 &\leqq \int_{\mathscr{T}} [g(z(t), q_h(t), t) - g(z_h(t), q_h(t), t)]\, dt \\ &= \int_{\mathscr{T}} [g(z(t), q(t), t) - g(z_h(t), q_h(t), t)]\, dt + \int_{\mathscr{T}} z(t) \cdot (q(t) - q_h(t))\, dt \\ &= f^*(q) - f^*(q_h) + \langle z, q - q_h\rangle.\end{aligned}$$

Integrating the last term by parts,

$$\langle z, q - q_h\rangle = x(0) \cdot (p_h(0) - p(0)) - x(1) \cdot (p_h(1) - p(1))$$

since $M(z) = 0$. By Theorem 8.2,

$$\langle z, q - q_h\rangle \leqq e^*(a) - e^*(a_h).$$

Combining these relations, the proof is complete. □

THEOREM 9.3. *Under the hypotheses of Theorem* 9.2, *we have*:

$$L(p) - L(p_I) \leqq \frac{1}{4\alpha}\|q - q_I\|^2$$

for all $p_I \in \mathscr{A}$ *which agree with* $p(t)$ *at* $t = 0$ *and* 1 *where* q *is given by* (8.2) *and*

$$q_I(t) = \begin{pmatrix} p_I'(t) + A(t)^T p_I(t) \\ B(t)^T p_I(t) \end{pmatrix}. \tag{9.8}$$

Proof. As in the last theorem's proof, we hold t fixed and apply Lemma 9.1 to $g(\cdot, \cdot, t)$ taking $\lambda = q_I(t)$ and $\mu = q(t)$. Integrating (9.4) over $\mathscr{T}$ and utilizing Theorem 8.2,

$$\begin{aligned}\frac{1}{4\alpha}\|q - q_I\|^2 &\geqq \int_{\mathscr{T}} [g(z(t), q_I(t), t) - g(z_I(t), q_I(t), t)]\, dt \\ &= f^*(q) - f^*(q_I) + \langle z, q - q_I\rangle = f^*(q) - f^*(q_I) = L(p) - L(p_I).\end{aligned}$$

The last step comes from Lemma 8.1 and the fact that $p_I = p$ at the ends of $\mathscr{T}$. The preceding step utilizes the relation $\langle z, q - q_I\rangle = 0$, which is deduced from the identity $M(z) = 0$. □

Unfortunately, this upper bound from $L(p) - L(p_I)$ is too coarse for the error estimates in § 10. Since p' appears in the first component of q, and the derivative of

the optimal dual multiplier is often discontinuous for state constrained problems,

$$\|q-q_I\|_{\mathscr{C}}=O(1)$$

when p_I lies in typical piecewise polynomial spaces. Hence the upper bound is expressed in terms of the smoother variable q^ω introduced in § 8. Given $K: R^n\times\mathscr{T}\to R^s$ and $x:\mathscr{T}\to R^n$, let $K(x):\mathscr{T}\to R^s$ be defined by

$$K(x)(t)=K(x(t),t).$$

THEOREM 9.4. *Suppose that* (P) *has a pointwise representation, the uniform convexity hypothesis holds, K is twice continuously differentiable on $R^n\times\mathscr{T}$, and for each $t\in\mathscr{T}$, $K(\cdot,t)$ is convex and*

$$\operatorname{dom} f(\cdot,t)\subset\{y\in R^n: K(y,t)\leqq 0\}\times R^m.$$

If (p,z) is an optimal pair with $p\in\mathscr{A}$ and (8.4) *holds for some $\omega\in\mathscr{A}$, then*

$$L(p)-L(p_I)\leqq\frac{1}{4\alpha}\|q^\omega-q_I^\omega\|^2-\langle\nu_I,K(x)\rangle_{\mathscr{C}} \tag{9.9}$$

for all $p_I\in\mathscr{A}$ that agree with p at the ends of $\mathscr{T}$, and for all nondecreasing $\nu_I\in\mathscr{A}_s$ where

$$q_I^\omega(t)=q_I(t)-\begin{pmatrix}G(t)^T\nu_I'(t)\\0\end{pmatrix}, \tag{9.10}$$

$G(t)=\nabla_x K(x(t),t)$, *and q_I is defined in* (9.8).

Proof. By Theorem 8.3,

$$f(z(t),t)-q^\omega(t)\cdot z(t)=\inf\{f(\xi,t)-q^\omega(t)\cdot\xi:\xi\in R^{n+m}\} \tag{9.11}$$

almost everywhere. If (p_I,z_I) is a min-pair, we have the trivial relation

$$f(z_I(t),t)-q_I^\omega(t)\cdot z_I(t)\geqq\inf\{f(\xi,t)-q_I^\omega(t)\cdot\xi:\xi\in R^{n+m}\} \tag{9.12}$$

almost everywhere. Lemma 9.1 with $\lambda=q_I^\omega(t)$ and $\mu=q^\omega(t)$ gives us

$$\begin{aligned}&\inf\{f(\xi,t)-q_I^\omega(t)\cdot\xi:\xi\in R^{n+m}\}-\inf\{f(\xi,t)-q^\omega(t)\cdot\xi:\xi\in R^{n+m}\}\\ &\geqq(q^\omega(t)-q_I^\omega(t))\cdot z(t)-\frac{1}{4\alpha}|q^\omega(t)-q_I^\omega(t)|^2\\ &=(q(t)-q_I(t))\cdot z(t)-\frac{1}{4\alpha}|q^\omega(t)-q_I^\omega(t)|^2+(\omega'(t)-\omega_I'(t))\cdot x(t)\end{aligned} \tag{9.13}$$

where $\omega_I'(t):=-G(t)^T\nu_I'(t)$. If $\varphi:R^n\to R$ is convex and differentiable and $\varphi(y)\leqq 0$ for some $y\in R^n$, the convexity inequality

$$\varphi(y)\geqq\varphi(x)+\varphi'[x](y-x)$$

implies that

$$\varphi(x)\leqq\varphi'[x](x-y). \tag{9.14}$$

Subtracting (9.11) from (9.12), utilizing (9.13) and (9.14) and integrating over $\mathscr{T}$, we get

$$f^*(q_I)-f^*(q)\geqq\langle q-q_I,z\rangle+\langle\nu_I,K(x)\rangle_{\mathscr{C}}-\frac{1}{4\alpha}\|q^\omega-q_I^\omega\|^2.$$

Integrating by parts, $\langle q-q_I, z\rangle = 0$ since $M(z)=0$ and $p_I = p$ at the ends of $\mathcal{T}$. Finally, by Lemma 8.1,

$$L(p)-L(p_I)=f^*(q)-f^*(q_I).$$

Collecting results, the proof is complete. □

If $p_h \in \mathcal{S}_h \subset \mathcal{A}$ and

$$L(p_h)=\text{maximum}\,\{L(p): p\in\mathcal{S}_h\},$$

we have the trivial relation

$$L(p)-L(p_h)\leqq L(p)-L(p_I)$$

for all $p_I \in \mathcal{S}_h$ and $p \in \text{dom } L$. Therefore, if (p, z) is an optimal pair, and the hypotheses of Theorems 9.2 and 9.4 hold,

$$\alpha\|z-z_h\|^2 \leqq L(p)-L(p_h) \leqq \frac{1}{4\alpha}\|q^\omega - q_I^\omega\|^2 - \langle \nu_I, K(x)\rangle_{\mathscr{C}} \tag{9.15}$$

for all $p_I \in \mathcal{S}_h$ which agree with p at the ends of $\mathcal{T}$, and for all nondecreasing $\nu_I \in \mathcal{A}_s$. Moreover, if

$$q^\omega(t)=q(t)-\begin{pmatrix} G(t)^T\nu'(t) \\ 0 \end{pmatrix},$$

then

$$q^\omega(t)-q_I^\omega(t)=\begin{pmatrix} \delta q'(t)+G'(t)^T\,\delta\nu(t)+A(t)^T\,\delta p(t) \\ B(t)^T\,\delta p(t) \end{pmatrix} \tag{9.16}$$

where

$$\delta p(t)=p(t)-p_I(t),\quad \delta\nu(t)=\nu(t)-\nu_I(t),\quad \delta q(t)=\delta p(t)-G(t)^T\,\delta\nu(t).$$

10. Error estimates. We now estimate the error in piecewise polynomial approximation. Given an interval $J\subset R$, let $\mathcal{P}^k(J)$ be the space of polynomials defined on J with degree at most k. Associated with a collection of points from $\mathcal{T}$:

$$0=t_0<t_1<\cdots<t_N=1,$$

we have the spacing parameter

$$h=\text{maximum}\,\{t_j-t_{j-1}: j=1,2,\cdots,N\},$$

and we let $\mathcal{P}_h^k$ denote the n-fold Cartesian product of sets of functions $f:\mathcal{T}\to R$ whose restriction to each interval $J=(t_{j-1}, t_j)$ lies in $\mathcal{P}^k(J)$. The points $\{t_0, t_1, \cdots, t_N\}$ are called the *mesh.*

For any interval $J\subset R$, we let $W^{0,\infty}(J)$ denote the set of essentially bounded functions $f: J\to R^n$, and for $k\geqq 1$, $W^{k,\infty}(J)\subset W^{0,\infty}(J)$ is the subspace of functions with $k-1$ Lipschitz continuous derivatives. The space $W^{s,\infty}(\mathcal{T})$ is abbreviated $W^{s,\infty}$. The main results in this section are stated below:

THEOREM 10.1. *Suppose that* (p, z) *is an optimal pair,* ν *satisfies* (5.4), *and* (9.15) *holds. If* $\mathscr{C}\cap\mathcal{P}_h^1\subset\mathcal{S}_h$, *we have*

$$\alpha\|z-z_h\|^2\leqq L(p)-L(p_h)=O(h^2)$$

provided the following conditions hold:

(i) x *and* $(p-G^T\nu)\in W^{2,\infty}$, $G\in W^{2,\infty}$, $A\in\mathscr{L}^2$;

(ii) $p\in W^{1,\infty}$, $\nu\in W^{1,\infty}$ *is nondecreasing*;

(iii) $K(x)\in W^{2,\infty}$ *and* $K(x)\leqq 0$.

Earlier [14] we observe that an optimal pair is often quite smooth except at points where constraints change between binding and nonbinding. Let $\tilde{W}^{k,\infty}$ denote the collection of functions $f\in W^{k-1,\infty}$ for which there is $M>0$ and scalars $0=s_0<s_1<\cdots<s_M=1$ such that the restriction of f to each interval (s_{j-1}, s_j) has $k-1$ Lipschitz continuous derivatives.

THEOREM 10.2. *Suppose that* (p, z) *is an optimal pair,* ν *satisfies* (5.4) *and* (9.15) *holds. If* $\mathscr{C}\cap\mathscr{P}_h^2\subset\mathscr{S}_h$, *we have*:

$$\alpha\|z-z_h\|^2\leqq L(p)-L(p_h)=O(h^3)$$

provided the following conditions hold:

(i) x *and* $(p-G^T\nu)\in\tilde{W}^{3,\infty}$, $G\in\tilde{W}^{3,\infty}$, $A\in\mathscr{L}^\infty$;

(ii) $p\in\tilde{W}^{2,\infty}$, $\nu\in\tilde{W}^{2,\infty}$ *is nondecreasing*;

(iii) $K(x)\in W^{2,\infty}$ *and* $K(x)\leqq 0$;

(iv) *the sets*

$$T_j=\{t\in\mathscr{T}\colon K_j(x(t),t)<0\},\qquad j=1,2,\cdots,s,$$

are each composed of a finite number of intervals, and there exists $\beta>0$ *such that*

$$\nu_j'(t)\geqq\beta\quad\forall t\in T_j^c,\quad j=1,2,\cdots,s.$$

These theorems are based on Lemmas 10.3, 10.4 and 10.5 appearing below. First, let us recall a result concerning polynomial interpolation. For any interval $J\subset R$ and any $f\in W^{0,\infty}(J)$, let $|f|_J$ denote the essential supremum of $|f(t)|$ over $t\in J$. Then [8] and [55] exhibit various linear maps $I\colon W_1^{s,\infty}(J)\to\mathscr{P}^k(J)$ for which

$$\left|\frac{d^m}{dt^m}(f-f_I)\right|_J\leqq c\mu(J)^{s-m}|f^{(s)}|_J \tag{10.1}$$

whenever $m\leqq s\leqq k+1$ and $f\in W_1^{s,\infty}(J)$ where $\mu(J)$ is the measure of J and c is a constant independent of f and J. (Remember that the subscript 1 on the space $W_1^{s,\infty}(J)$ means that the elements of the space map J to R^1.) Throughout this section, J is an interval and c denotes a generic constant. The operator I is usually called the *interpolation operator*, and we write f_I rather than $I(f)$; the function f_I is called the *interpolant* of f. For illustration, the following operator satisfies (10.1) when $s\geqq 1$: Let f_I be the unique polynomial of degree at most k that agrees with f at $k+1$ evenly spaced points on J.

Suppose that $0=t_0<t_1<\cdots<t_N=1$ is a mesh on $\mathscr{T}$ and $f\colon\mathscr{T}\to R$, and the restriction of f to each interval $J=[t_{j-1}, t_j]$ lies in $W_1^{s,\infty}(J)$. If I is an interpolation operator satisfying (10.1), we let $f_I\colon\mathscr{T}\to R$ be the function composed of interpolants of f over each interval $J=[t_{j-1}, t_j]$. If $f_I\in W_1^{m,\infty}$, (10.1) implies that

$$\left|\frac{d^m}{dt^m}(f-f_I)\right|_{\mathscr{T}}\leqq ch^{s-m}|f^{(s)}|_{\mathscr{T}} \tag{10.2}$$

whenever $m\leqq s\leqq k+1$ and $f\in W_1^{s,\infty}$. Finally, defining

$$\langle f, g\rangle_{\mathscr{C}(J)}=\int_J g(t)\cdot df(t)$$

for $g\in\mathscr{C}$ and $f\in\mathscr{B}$, we have:

LEMMA 10.3. *Suppose that $J \subset R$ is an interval, $f \in W_1^{1,\infty}(J)$, $g \in W_1^{2,\infty}(J)$, $g \leqq 0$, and*

$$f'(t)g(t) = 0 \quad \text{almost everywhere.} \tag{10.3}$$

If the interpolation operator I satisfies (10.1), *then*

$$\langle f_I, g\rangle_{\mathscr{C}(J)} \leqq c\mu(J)^3 |f^{(1)}|_J |g^{(2)}|_J.$$

Moreover, if $f \in W_1^{2,\infty}(J)$ and $k \geqq 1$,

$$\langle f_I, g\rangle_{\mathscr{C}(J)} \leqq c\mu(J)^4 |f^{(2)}|_J |g^{(2)}|_J.$$

Proof. If $g(t) > 0$ almost everywhere, (10.3) implies that f is constant. Thus, $f = f_I$ by (10.1) and

$$\langle f_I, g\rangle_{\mathscr{C}(J)} = 0.$$

Now, let us suppose that g vanishes at σ in the interior of J. The relation $g \leqq 0$ implies that $g'(\sigma) = 0$. Expanding in a Taylor series about σ yields:

$$|g|_J \leqq \tfrac{1}{2}\mu(J)^2 |g^{(2)}|_J.$$

Utilizing (10.3), we get:

$$\langle f_I, g\rangle_{\mathscr{C}(J)} = \langle f_I - f, g\rangle_{\mathscr{C}(J)} \leqq \mu(J)|f_I' - f'|_J |g|_J \leqq \tfrac{1}{2}\mu(J)^3 |g^{(2)}|_J |f_I' - f'|_J.$$

Relation (10.1) completes the proof. □

LEMMA 10.4. *Suppose that $J \subset R$ is an interval, and the interpolation operator I acts on $f: J \to R$ to produce the polynomial of degree at most k that agrees with f at $k+1$ distinct points on J. Then we have:*

$$\left| \frac{d^m}{dt^m}[(fg)_I - fg_I] \right|_J \leqq c\mu(J)^{k+1-m} \sum_{i=0}^{k} |f^{(k+1-i)}|_J |g^{(i)}|_J$$

whenever $0 \leqq m \leqq k+1$ and $f', g \in W_1^{k,\infty}(J)$.

Proof. Since the interpolant is expressed in terms of function values,

$$(fg)_I = (fg_I)_I.$$

Hence (10.1) gives us:

$$\left| \frac{d^m}{dt^m}[(fg_I)_I - fg_I] \right|_J \leqq c\mu(J)^{k+1-m} |(fg_I)^{(k+1)}|_J.$$

By Leibniz's formula

$$(fg)^{(m)} = \sum_{i=0}^{m} \frac{m!}{i!(m-i)!} f^{(i)} g^{(m-i)},$$

we see that

$$|(fg_I)^{(k+1)}|_J \leqq c \sum_{i=0}^{k+1} |f^{(k+1-i)} g_I^{(i)}|_J.$$

Since $g \in W_1^{k,\infty}$, (10.1) implies that

$$|g_I^{(i)}|_J \leqq c|g^{(i)}|_J$$

for all $0 \leqq i \leqq k$. Furthermore, $g_I^{(k+1)}$ is identically zero since g_I is a polynomial of

degree at most k. These relations and the inequality

$$|fg|_J \leqq |f|_J |g|_J$$

complete the proof. □

LEMMA 10.5. *If $J \subset R$ is a closed interval and $f \in W_1^{2,\infty}(J)$, then the quadratic agreeing with f at the two ends and the midpoint of J is nondecreasing if*

$$\mu(J)|f''|_J \leqq 2 \text{ minimum } \{f'(t): t \in J\}.$$

Proof. Since a nontrivial interval can be mapped by an affine transformation onto $\mathcal{T}$, there is no loss of generality in assuming that $J = \mathcal{T}$. Let I be the interpolation operator described by the lemma. Since I is linear and $g_I = g$ if g is constant, we can also assume that $f(0) = 0$. In this case, observe that

$$f_I(t) = 4f(\tfrac{1}{2})t(1-t) + f(1)t(2t-1).$$

The derivative of this quadratic is linear and nonnegative on $\mathcal{T}$ if and only if it is nonnegative at $t = 0$ and 1. Omitting the arithmetic, f_I is nondecreasing if and only if

$$\tfrac{3}{4}f(1) \geqq f(\tfrac{1}{2}) \geqq \tfrac{1}{4}f(1). \tag{10.4}$$

Since f is continuously differentiable, there exists $\sigma \in \mathcal{T}$ such that

$$f'(\sigma) = f(1).$$

Suppose that $\sigma \leqq \frac{1}{2}$ (the case $\sigma > \frac{1}{2}$ is treated in a similar manner). The identity

$$f(\tfrac{1}{2}) = \tfrac{1}{2}f(1) + \int_0^{1/2} \int_\sigma^\tau f''(t)\, dt\, d\tau,$$

and the bound

$$\int_0^{1/2} \int_0^\tau |f''(t)|\, dt\, d\tau \leqq \tfrac{1}{8}|f''|_{\mathcal{T}}$$

imply that

$$|f(\tfrac{1}{2}) - \tfrac{1}{2}f(1)| \leqq \tfrac{1}{8}|f''|_{\mathcal{T}}. \tag{10.5}$$

Combining (10.4) and (10.5), f_I is nondecreasing if

$$|f''|_{\mathcal{T}} \leqq 2f(1) = 2f'(\sigma),$$

a condition clearly satisfied under the lemma's hypothesis. □

Now, let us prove Theorem 10.1. Let (p_I, ν_I) be the continuous piecewise linear function which agrees with (p, ν) at each mesh point (except that $\nu_I(0) = \nu(0^+)$ and $\nu_I(1) = \nu(1^-)$—see the remarks at the end of § 5). Relation (5.4) and the identity

$$L(p) = C(z) + \langle p, M(z) \rangle$$

imply that $\langle \nu, K(x) \rangle_{\mathcal{C}} \geqq 0$. Since ν is nondecreasing and $K(x) \leqq 0$, we conclude that

$$\nu'(t) \cdot K(x(t), t) = 0 \quad \text{almost everywhere.}$$

Therefore, by Lemma 10.3 and the assumed smoothness properties,

$$\langle \nu_I, K(x) \rangle_{\mathcal{C}} \leqq O(h^2).$$

Furthermore, by (10.1),

$$|\delta p|_{\mathcal{T}} = O(h) = |\delta \nu|_{\mathcal{T}}.$$

Finally, observe that δq can be expressed as follows:

$$\delta q = (p - G^T\nu) - (p - G^T\nu)_I + G^T\nu_I - (G^T\nu)_I. \tag{10.6}$$

Since the operator I is linear, Lemma 10.4 and the assumed regularity give us

$$|\delta q'|_{\mathscr{T}} = O(h).$$

Relations (9.15) and (9.16) complete the proof. □

The proof of Theorem 10.2 is similar. Recall that the sets T_j defined earlier are each composed of a finite number of intervals. Let $\{\sigma_1, \sigma_2, \cdots, \sigma_k\}$ denote the union over j of boundary points in T_j, and let $\{\sigma_{k+1}, \cdots, \sigma_l\}$ be the points separating intervals where $x^{(3)}$, $(p - G\nu)^{(3)}$, $G^{(3)}$, $p^{(2)}$ and $\nu^{(2)}$ are essentially bounded. We form an interpolant (p_I, ν_I) by pasting together local interpolants of (p, ν) over each grid interval J where the local interpolants are defined as follows:

(1) If $J \cap \{\sigma_j\}$ is nonempty, interpolate linearly between function values at the ends of J.

(2) If $J \cap \{\sigma_j\}$ is empty, use quadratic interpolation based on function values at the ends and middle of J.

By assumption,

$$\nu_j'(t) \geqq \beta > 0 \quad \forall t \in T_j^c,$$

$j = 1, 2, \cdots, s$. Hence, when h is small enough, Lemma 10.5 asserts that $(\nu_I)_j$ is nondecreasing on all mesh intervals which intersect the complement of T_j. On the other hand, we observed in the proof of Theorem 10.1 that $\langle \nu, K(x)\rangle_{\mathscr{C}} = 0$. Since ν is nondecreasing and $K(x) \leqq 0$, it follows that ν_j is constant on intervals contained in T_j. Therefore, $(\nu_I)_j = \nu_j$ on all mesh intervals contained in T_j, and ν_I is nondecreasing if h is small enough.

By (10.1) and the assumed smoothness properties,

$$|\delta p|_J = O(h^2) = |\delta \nu|_J$$

for all mesh intervals J such that $J \cap \{\sigma_j\}$ is empty, and

$$|\delta p|_J = O(h) = |\delta \nu|_J$$

otherwise. Similarly, the identity (10.6) and Lemma 10.4 give us

$$|\delta q'|_J = O(h^2)$$

if $J \cap \{\sigma_j\}$ is empty, and

$$|\delta q'|_J = O(h)$$

otherwise. And by Lemma 10.3,

$$\langle \nu_I, K(x)\rangle_{\mathscr{C}(J)} \leqq c\mu(J)^4$$

if $J \cap \{\sigma_j\}$ is empty, and

$$\langle \nu_I, K(x)\rangle_{\mathscr{C}(J)} \leqq c\mu(J)^3$$

otherwise. Since the measure of mesh intervals intersecting $\{\sigma_j\}$ is at most lh, relations (9.15) and (9.16) complete the proof. □

For problems without state constraints, the analysis is much easier. In [18] we give a simple treatment of quadratic cost problems with control constraints. Although smoothness considerations limit the $\mathscr{L}^2$ convergence rate to 1.5, higher rates are achieved when the grid points are free parameters in the optimization process—see [14].

11. Algorithms. Section 10 establishes the convergence of dual finite element approximations to constrained control problems. We now consider the practical side: How is the dual problem solved? When the dual optimization is unconstrained, steepest descent, conjugate gradient and quasi-Newton methods can be applied, but the cost functional is ill-conditioned, and computing time on an IBM 370 computer can be one hour for simple problems! Our main objective in this section is to present a new algorithm which *quickly* solves the dual problem. We also examine the tightness of the error estimates that were established in § 10.

To illustrate the conditioning problems that can arise when standard optimization techniques are applied to the dual problem, the following experiment is cited: Consider the approximation (D_h) to the dual of Problem II from § 8 where the approximating space $\mathcal{S}_h$ is a space of linear splines on a uniform mesh (see [8], [38], or [55] for a discussion of piecewise polynomial spaces). The time needed to solve this dual problem using: 1000 basis elements (which gives 5-place accuracy), an IBM 370 model 3033 computer, the IMSL conjugate gradient routine, the FORTRAN IV (H) optimizing compiler and the initial guess zero, is 1 hour. We now develop an algorithm which solves this dual problem in 1 second.

For the dual problems in § 8, observe that the dual integrand at time t is chosen from a finite set. For example, the dual integrand in Problem I is $l(x, y, t) = \frac{1}{2}[x^2 + y^2]$ if $y \leqq a$ and $l(x, y, t) = \frac{1}{2}[x^2 + a(2y - a)]$ if $y > a$. In general the dual integrand l is expressed in terms of a partition $\{\mathcal{R}_1, \cdots, \mathcal{R}_k\}$ of $R^{2n} \times \mathcal{T}$ and integrands $l_1, \cdots, l_k$ defined on $R^{2n} \times \mathcal{T}$. And the integrand $l(p'(t), p(t), t)$ of the dual functional satisfies

$$l(x, y, t) = l_i(x, y, t) \tag{11.1}$$

whenever $(x, y, t) \in \mathcal{R}_i$. For Problem I, we have:

$$l_1(x, y, t) = \tfrac{1}{2}[x^2 + y^2], \qquad l_2(x, y, t) = \tfrac{1}{2}[x^2 + a(2y - a)],$$

$$\mathcal{R}_1 = \{(x, y, t) \in R \times R \times \mathcal{T} \colon y \leqq a\}, \qquad \mathcal{R}_2 = \{(x, y, t) \in R \times R \times \mathcal{T} \colon y > a\}.$$

In formulating our algorithm for the dual problem, we assume that the dual function has the form

$$L(p) = \phi(p) + \int_{\mathcal{T}} l(p'(t), p(t), t)\, dt$$

where $\phi : \mathcal{A} \to R \cup \{-\infty\}$ and l satisfies (11.1) for some partition $\{\mathcal{R}_1, \cdots, \mathcal{R}_k\}$ of $R^{2n} \times \mathcal{T}$ and integrands $l_1, \cdots, l_k$ defined on $R^{2n} \times \mathcal{T}$. Now, given a partition $T = \{T_1, \cdots, T_k\}$ of $\mathcal{T}$ into measurable sets, let us define the functional

$$M(p, T) = \phi(p) + \sum_{i=1}^{k} \int_{T_i} l_i(p'(t), p(t), t)\, dt. \tag{11.2}$$

Any $p \in \mathcal{A}$ induces a partition $\{T_1, \cdots, T_k\}$ of $\mathcal{T}$ where $t \in T_i$ if and only if

$$(p'(t), p(t), t) \in \mathcal{R}_i.$$

Let S be the map that acts on p to produce the associated partition of $\mathcal{T}$. From these definitions, we see that

$$L(p) = M(p, S(p)).$$

For the examples in § 8, observe that the elements of $S(p)$ are measurable for each $p \in \mathcal{A}$. More generally, it can be shown that the elements of $S(p)$ are measurable if the multifunctions $\mathcal{R}_1, \cdots, \mathcal{R}_k$ are measurable—see [51]. Henceforth, we assume that the elements of $S(p)$ are measurable for every $p \in \mathcal{A}$.

Letting $K \subset \mathcal{A}$ denote a convex set of dual feasible functions which contains a solution to the dual problem, our algorithm for solving (D) is the following: Starting from some $p^0 \in K$, we generate a sequence $p^1, p^2, \cdots$ (we hope) converging to a dual solution where

$$p^{j+1} = \arg\max \{M(p, T^j): p \in K\}, \qquad T^j := S(p^j).$$

There is an analogous scheme for the dual approximation (D_h). If $\{\psi_1, \cdots, \psi_N\}$ is a basis for the finite element space $\mathcal{S}_h \subset \mathcal{A}$, we define

$$M^h(\alpha, T) = M\left(\sum_{i=1}^N \alpha_i \psi_i, T\right) \quad \text{and} \quad K^h = \left\{\alpha \in R^N: \sum_{i=1}^N \alpha_i \psi_i \in K\right\}.$$

Our scheme for solving (D_h) starts from some $\alpha^0 \in K^h$, and constructs iterations $\alpha^1, \alpha^2, \cdots$ by the rule

$$\alpha^{j+1} = \arg\max \{M^h(\alpha, T^j): \alpha \in K^h\}, \qquad T^j := S\left(\sum_{i=1}^N \alpha_i^j \psi_i\right). \tag{11.3}$$

We remark that if $M^h(\cdot, T^j)$ is Gateaux differentiable and α^{j+1} satisfies (11.3), then the following standard inequality holds [27, Thm. I.1.3]:

$$\frac{\partial M^h}{\partial \alpha}[\alpha^{j+1}, T^j](\alpha - \alpha^{j+1}) \leqq 0 \quad \forall \alpha \in K^h.$$

This algorithm has been tested on the problems presented in § 8. Experimentally, the convergence is fast; moreover, the iterations seem to converge from any starting point α^0. For example, starting from the initial guess $\alpha^0 = 0$ in Problem I and using the linear spline basis mentioned earlier, the relative change $|\alpha^{j+1} - \alpha^j|/|\alpha^j|$ is reduced to 10^{-10} after 5 iterations, independent of the number of basis elements; each iteration involves solving a symmetric, tridiagonal system and is easy to implement. The FORTAN code for Problem I has about 60 statements. Later we show under appropriate hypotheses that the scheme (11.3) is quadratically convergent near a solution to the dual problem (D_h).

First we observe that any fixed point for the iterations (11.3) solves the dual maximization problem (D_h). This result is based on the rule for differentiating under the integral sign. Below, $W^{1,\infty}$ denotes the space of Lipschitz continuous functions $p: \mathcal{T} \to R^n$ with the norm

$$\|p\|_{W^{1,\infty}} = \text{essential supremum } \{|p(t)| + |p'(t)|: t \in \mathcal{T}\}.$$

LEMMA 11.1. *Suppose that $T \subset \mathcal{T}$ is measurable and $g: R^{2n} \times \mathcal{T} \to R$ is continuously differentiable in its first $2n$ arguments on $R^{2n} \times \mathcal{T}$. If $G: W^{1,\infty} \to R$ is defined by*

$$G(p) = \int_T g(p'(t), p(t), t)\, dt,$$

then the Fréchet derivative of G is

$$\frac{\partial G}{\partial p}[p](q) = \int_T [\nabla_1 g(p'(t), p(t), t) q'(t) + \nabla_2 g(p'(t), p(t), t) q(t)]\, dt$$

where $\nabla_1 g$ and $\nabla_2 g$ denote g's gradient with respect to its first n and second n arguments respectively.

Note that every dual integrand presented in § 8 is continuously differentiable. In Appendix 2 we show that this continuity property holds for a broad class of problems.

To prove that any fixed point of the iterations (11.3) solves the dual maximization problem (D_h), let us assume that both the integrands l_i and the composite integrand l are continuously differentiable in their first $2n$ arguments on $R^{2n} \times \mathcal{T}$ and the function ϕ in (11.2) is differentiable. Defining the sets

$$\mathcal{R}_i(t) = \{r \in R^{2n} : (r, t) \in \mathcal{R}_i\},$$

we assume moreover that

$$\text{closure } \mathcal{R}_i(t) = \text{closure (interior } \mathcal{R}_i(t))$$

for each $t \in \mathcal{T}$. Hence, if $r \in \mathcal{R}_i(t)$, there exists a sequence $\{r_k\} \subset$ interior $\mathcal{R}_i(t)$ converging to r, and since $l_i(\cdot, t) = l(\cdot, t)$ near r_k, we have

$$\nabla l_i(r_k, t) = \nabla l(r_k, t).$$

Taking the limit as k goes to infinity, the continuous differentiability assumption implies that

$$\nabla l_i(r, t) = \nabla l(r, t) \tag{11.4}$$

for every $r \in \mathcal{R}_i(t)$. Now, let us define

$$L^h(\alpha) = L\left(\sum_{i=1}^{N} \alpha_i \psi_i\right).$$

If each ψ_i lies in $W^{1,\infty}$, then Lemma 11.1 and (11.4) yield

$$\frac{\partial L^h}{\partial \alpha}[\beta] = \frac{\partial M^h}{\partial \alpha}[\beta, T] \quad \text{where } T = S\left(\sum_{i=1}^{N} \beta_i \psi_i\right).$$

This observation, the concavity of the dual function and the following result combine to show that any fixed point of the iterations (11.3) solves (D_h).

THEOREM 11.2. *Suppose that $G: K \times K \to R$ where K is a convex subset of a vector space, $y \in K$, and $G(x, x)$ and $G(x, y)$ are Gateaux differentiable functions of x at $x = y$ which satisfy*

$$\left.\frac{\partial G(x, x)}{\partial x}\right|_{x=y} = \left.\frac{\partial G(x, y)}{\partial x}\right|_{x=y}.$$

If $G(x, x)$ is a convex function of $x \in K$ and y minimizes $G(x, y)$ over $x \in K$, then y minimizes $G(x, x)$ over $x \in K$. Conversely, if $G(x, y)$ is a convex function of $x \in K$ and y minimizes $G(x, x)$ over $x \in K$, then y minimizes $G(x, y)$ over $x \in K$.

Proof. First assume that $G(x, x)$ is a convex function of $x \in K$ and y minimizes $G(x, y)$ over $x \in K$. Since $G(\cdot, y)$ is Gateaux differentiable at y and K is convex, we have the standard variational inequality [27, Thm. I.1.3]:

$$\left.\frac{\partial G(x, y)}{\partial x}\right|_{x=y} (x - y) \geqq 0 \quad \forall x \in K. \tag{11.5}$$

The hypotheses for the gradient and (11.5) imply that

$$\left.\frac{\partial G(x, x)}{\partial x}\right|_{x=y} (x - y) \geqq 0 \quad \forall x \in K. \tag{11.6}$$

Since $G(x, x)$ is convex, it follows from (11.6) and [27, Thm. I.1.3] that y minimizes $G(x, x)$ over $x \in K$. Conversely, let us assume that y minimizes $G(x, x)$ over $x \in K$ and $G(x, y)$ is a convex function of $x \in K$. Again, by [27, Thm. I.1.3], the variational

inequality (11.6) holds, and by the hypotheses for the gradient, we conclude that (11.5) is satisfied. Since $G(\cdot, y)$ is convex on K, (11.5) implies that y minimizes $G(x, y)$ over $x \in K$. □

Before giving a convergence proof for the iterative scheme (11.3), let us examine Problems I and II to help motivate our theorem's hypotheses. Let $\mathcal{S}_h$ be the space of linear splines defined on a uniform mesh where $h = 1/N$ is the distance between grid points, and let $\{\psi_0, \cdots, \psi_N\}$ be the usual basis for $\mathcal{S}_h$ sketched in Fig. 4. Applying Lemma 11.1 to the dual functional for Problem I, we have

$$-\frac{\partial L^h(p)}{\partial \alpha_i} = \int_0^1 p'(t)\psi_i'(t)\,dt + \int_{T_1(\alpha)} p(t)\psi_i(t)\,dt + a\int_{T_2(\alpha)} \psi_i(t)\,dt \tag{11.7}$$

for $i = 1, \cdots, N$ where

$$p(t) = \sum_{i=0}^{N} \alpha_i \psi_i(t), \quad T_1(\alpha) = \{t \in \mathcal{T}: p(t) \leqq a\}, \quad T_2(\alpha) = \{t \in \mathcal{T}: p(t) > a\}.$$

The partial derivative of $-L^h$ with respect to α_0 is c plus the terms on the right side of (11.7) where c is the state's initial value in Problem I. If $p(t)$ equals a at just a finite set of $t \in (0, 1)$ and $0 = t_l(\alpha) < t_{l+1}(\alpha) < \cdots < t_r(\alpha) = 1$ denote these t where $p(t)$ is a union $\{0, 1\}$, then we can write

$$\int_{T_1(\alpha)} p(t)\psi_i(t)\,dt + a\int_{T_2(\alpha)} \psi_i(t)\,dt = \sum_{j\,\text{even}} \int_{t_j(\alpha)}^{t_{j+1}(\alpha)} p(t)\psi_i(t)\,dt + \sum_{j\,\text{odd}} a\int_{t_j(\alpha)}^{t_{j+1}(\alpha)} \psi_i(t)\,dt.$$

Thus for β in a neighborhood of the fixed coefficients $\{\alpha_0, \cdots, \alpha_N\}$, the gradient of L^h evaluated at β has the form $g(\beta, T(\beta))$ where $T(\beta)$ is a vector with components $t_l(\beta), \cdots, t_r(\beta)$. Our main observation is the following: Since the $\psi_i(t)$ are continuous functions of t and $p(t_j(\alpha)) = a$ for $j = l+1, \cdots, r-1$, we have:

$$\left.\frac{\partial g_i(\alpha, T(\beta))}{\partial \beta_k}\right|_{\beta=\alpha} = \sum_{j=l+1}^{r-1} a[\psi_i(t_j) - \psi_i(t_j)]\left.\frac{\partial t_j(\beta)}{\partial \beta_k}\right|_{\beta=\alpha} = 0.$$

More compactly, this result can be stated

$$\left.\frac{\partial g(\alpha, T(\beta))}{\partial \beta}\right|_{\beta=\alpha} = 0. \tag{11.8}$$

This identity also holds for state constrained problems, but the argument is a little different. For Problem II, the terms in the gradient of the dual function corresponding to the state constraint are

$$\sum_{j\,\text{even}} \int_{t_j(\alpha)}^{t_{j+1}(\alpha)} p'(t)\psi_i'(t)\,dt + \sum_{j\,\text{odd}} b\int_{t_j(\alpha)}^{t_{j+1}(\alpha)} \psi_i'(t)\,dt \tag{11.9}$$

where the two sums above correspond to intervals where $p'(t) \leqq b$ and $p'(t) > b$ respectively. Since p is a linear spline, p' is piecewise constant. Hence, if $p'(t) \neq b$ for

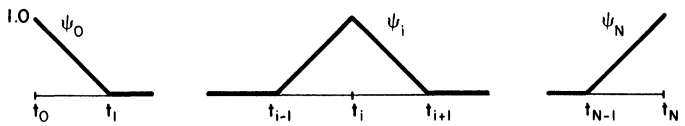

FIG. 4. *Linear spline basis.*

every $t \in \mathcal{T}$, then $t_j(\beta)$ is independent of β in a neighborhood of α. In summary, even though the integrands in (11.9) are discontinuous functions of t, the identity (11.8) still holds since $t_j(\beta)$ is independent of β in a neighborhood of α. With this motivation, we now present our local quadratic convergence result:

THEOREM 11.3. *Suppose that $g: R^n \times R^n \to R^n$ and K is a nonempty, closed convex subset of R^n, and consider the problem of finding $\alpha \in K$ such that*

$$g(\alpha, \alpha) \cdot (\beta - \alpha) \geqq 0 \tag{11.10}$$

for all $\beta \in K$. We assume that there exists a solution a^* to (11.10) *and that the following conditions are satisfied*:

(1) *$g(\alpha, \beta)$ is a continuous function of α and β near α^*, and $\partial g(\alpha, \beta)/\partial\alpha$ exists and is a continuous function of α and β near α^*.*

(2) *$g(\alpha^*, \cdot)$ is twice continuously differentiable near α^* and the first derivative vanishes at α^*.*

(3) *Either $K = R^n$ and $\partial g(\alpha, \alpha^*)/\partial\alpha|_{\alpha=\alpha^*}$ is nonsingular, or K is an arbitrary closed convex subset of R^n and $\partial g(\alpha, \alpha^*)/\partial\alpha|_{\alpha=\alpha^*}$ is positive definite.*

Then there exists a neighborhood $\mathcal{N}$ of α^* with the following properties: For each $\alpha^0 \in \mathcal{N}$, there is a unique sequence $\{\alpha^1, \alpha^2, \cdots\} \subset K \cap \mathcal{N}$ such that

$$g(\alpha^{j+1}, \alpha^j) \cdot (\beta - \alpha^{j+1}) \geqq 0 \tag{11.11}$$

for all $\beta \in K$ and $j = 0, 1, \cdots$, and for some constant c independent of j and $\alpha_0 \in \mathcal{N}$, we have:

$$|\alpha^{j+1} - \alpha^*| \leqq c|\alpha^j - \alpha^*|^2.$$

Proof. By Robinson [43, Thms. 2.1 and 3.1], there exist neighborhoods $\mathcal{N}_1$ and $\mathcal{N}_2$ of α^* such that the following problem has a unique solution $\alpha \in \mathcal{N}_1$ for each $\gamma \in \mathcal{N}_2$: find $\alpha \in K$ such that

$$g(\alpha, \gamma) \cdot (\beta - \alpha) \geqq 0 \tag{11.12}$$

for all $\beta \in K$. Shrink $\mathcal{N}_2$ so it is contained in a bounded region where $g(\alpha^*, \cdot)$ is twice continuously differentiable, and let $\Phi(\gamma) \in \mathcal{N}_1$ denote the solution of (11.12) corresponding to $\gamma \in \mathcal{N}_2$. By [43, Thm. 2.1] there also exists a constant μ such that

$$|\Phi(\gamma) - \Phi(\alpha^*)| \leqq \mu |g(\alpha^*, \gamma) - g(\alpha^*, \alpha^*)| \tag{11.13}$$

for all $\gamma \in \mathcal{N}_2$. Expanding $g(\alpha^*, \cdot)$ to first order about α^* and using the integral form for the remainder term, our second hypothesis implies that

$$|g(\alpha^*, \gamma) - g(\alpha^*, \alpha^*)| \leqq c|\gamma - \alpha^*|^2 \tag{11.14}$$

for some constant c independent of $\gamma \in \mathcal{N}_2$. Combining (11.13) and (11.14), we have for $\alpha^j \in \mathcal{N}_2$:

$$|\alpha^{j+1} - \alpha^*| = |\Phi(\alpha^j) - \Phi(\alpha^*)| \leqq c|\alpha^j - \alpha^*|^2. \tag{11.15}$$

Thus if α^0 is sufficiently close to α^*, the entire sequence $\{\alpha^j\}$ given by $\alpha^{j+1} = \Phi(\alpha^j)$ lies in $\mathcal{N}_1 \cap \mathcal{N}_2$ and (11.15) holds. □

Observe that the inequality (11.10) is essentially the relation (11.6) characterizing the solution to the dual approximation (D_h) while the iterations defined by (11.11) correspond to our scheme (11.3). The inequality (11.13) is a crucial step in our proof of Theorem 11.3. In Robinson's study [43] of the implicit function theorem for inequalities, he establishes this relation in a very general setting whenever the "strong regularity" assumption is satisfied. Moreover, in finite dimensions it follows from his

Theorem 3.1 that the strong regularity assumption holds under hypothesis 3 of our theorem. Hence a more general version of Theorem 11.3 can be established where R^n is replaced by a normed linear space and hypothesis 3 is replaced by the strong regularity assumption.

Now let us study the tightness of the error estimates established in § 10. Since the solutions to Problems I and II from § 8 can be determined analytically (see Appendix 3), the error in finite element approximations can be computed precisely. Taking $a = 1$ and $c = (1+3e)/2(1-e)$ in Problem I, the optimal control is 1 for $t \in [0, \frac{1}{2}]$. Thus the constraint $u(t) \leqq a$ is binding for the optimal control when $0 \leqq t \leqq \frac{1}{2}$. Taking $a = 1$, $b = 2\sqrt{e}/(1-e)$, and $c = (5e+3)/4(1-e)$ in Problem II, the optimal control is a for $t \in [0, \frac{1}{4}]$ and the optimal state is b for $t \in [\frac{3}{4}, 1]$. The solutions for these choices of parameters are shown in Figs. 5 and 6.

In § 10 we give the estimates

$$\|x - x^h\| + \|u - u^h\| = \begin{cases} O(h) & \text{for linear elements,} \\ O(h^{3/2}) & \text{for quadratic elements,} \end{cases}$$

where (x, u) solves the primal problem, (x^h, u^h) is the finite element approximation, and $\|\cdot\|$ is the $\mathscr{L}^2$ norm. Comparing the exact solution of Problems I and II to the finite element approximations, these estimates are tight. That is, there exists a constant $C > 0$ such that

$$\|x - x^h\| + \|u - u^h\| \geqq \begin{cases} Ch & \text{for linear elements,} \\ Ch^{3/2} & \text{for quadratic elements.} \end{cases}$$

Problem IV was solved using linear splines, and we obtained the solution reported in [56]; moreover, the error $\|x - x^h\| + \|u - u^h\|$ was proportional to h. Since Problem IV is not strictly convex, it appears that the convexity assumptions in § 9 can be mildly relaxed.

Although the estimate for $\|x - x^h\| + \|u - u^h\|$ is tight, we also observe in Figs. 7 and 8 that the control converges faster than the state. For linear elements,

$$\|u - u^h\| = O(h^{3/2}),$$

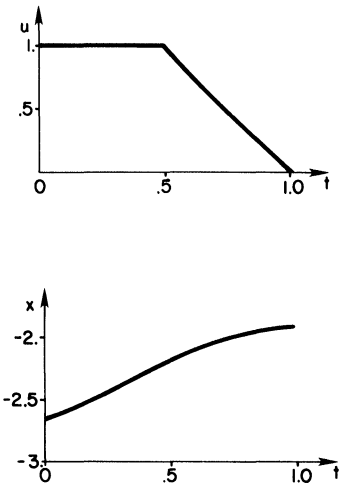

FIG. 5. *Solution to Problem* I.

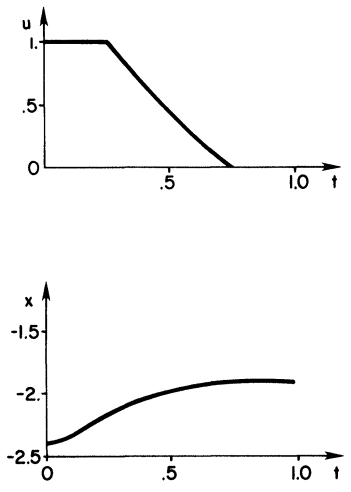

FIG. 6. *Solution to Problem* II.

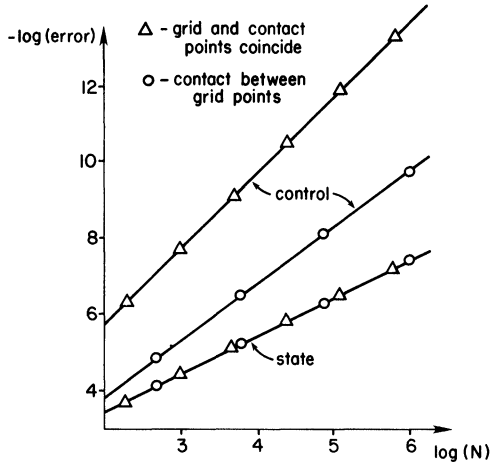

FIG. 7. $\mathscr{L}^2$ *error for Problem* I *and linear elements.*

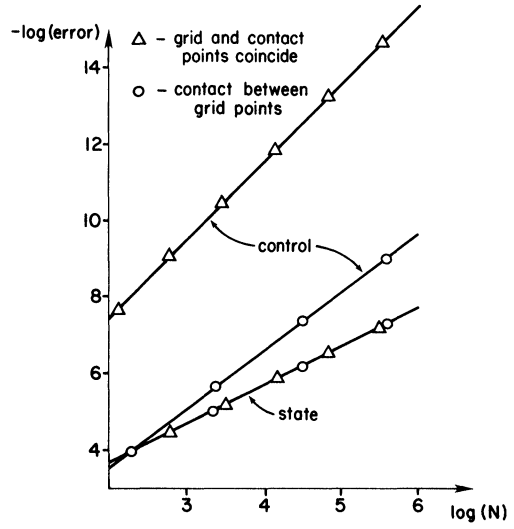

FIG. 8. $\mathscr{L}^2$ *error for Problem* II *and linear elements.*

and putting grid points at the contact points (where constraints in the primal problem change between binding and nonbinding) gives us the better result:

$$\|u-u^h\| = O(h^2).$$

For unconstrained problems, Mathis and Reddien [32] use a duality argument to show that the control convergence rate is h times the state rate. The extension of this result to constrained optimization is open. The convergence rate for quadratic elements also improves when the contacts are members of the grid:

$$\|x-x^h\|+\|u-u^h\| = O(h^2).$$

This property of the total error is established in [14] for a full dual scheme.

In summary better approximations to the primal solution are obtained as follows: Solve the dual problem on a fixed mesh and estimate the contacts; then insert grid points at the approximate contacts, and repeat the process. After the contacts converge, generate a better state by integrating forward the system dynamics with the approximate control as input.

Appendix 1. Existence. Suppose that $f: X \to [-\infty, +\infty)$ and $g: X \to Y$ where X is a set and Y is a normed vector space that is ordered by a convex cone $N \subset Y$ with vertex at the origin; that is, given a and $b \in Y$, we write $a \geqq b$ if $b - a \in N$. If Y^* denotes the space of bounded linear functionals on Y, N induces an ordering on Y^* relative to the convex cone

$$N^* = \{y^* \in Y^* : \langle y^*, y \rangle \geqq 0 \text{ for all } y \in N\}.$$

Above $\langle \cdot, \cdot \rangle$ denotes the usual pairing between Y^* and Y. Associated with the primal problem,

$$(P') \qquad \begin{aligned} &\text{maximize} \quad f(x) \\ &\text{subject to} \quad g(x) \geqq 0, \quad x \in X, \end{aligned}$$

is the dual problem,

$$(D') \qquad \begin{aligned} &\text{minimize} \quad L(\lambda) \\ &\text{subject to} \quad \lambda \geqq 0, \quad \lambda \in Y^*, \end{aligned}$$

where

$$L(\lambda) = \text{supremum}\,\{f(x) + \langle \lambda, g(x) \rangle : x \in X\}.$$

Although equality constraints are not explicitly stated in the primal problem, the inequality $g(x) \geqq 0$ becomes equality when $N = \{0\}$.

Under certain convexity hypotheses and constraint qualifications, a "typical duality theorem" asserts that there exists a solution λ to the dual problem and

$$L(\lambda) = \text{supremum}\,\{f(x) : g(x) \geqq 0, x \in X\}.$$

For example, see [28], [50], or Theorem A3 below. First, we observe that dual approximations exist without the convexity hypothesis. If S is a subset of $-N^*$, we consider the following approximation to the dual problem:

$$(D'_S) \qquad \text{minimize} \quad \{L(\lambda) : \lambda \in S\}.$$

Let us define the set

$$\Delta(y) = \{x \in X : g(x) \geqq y\},$$

and the ball

$$\mathcal{B}^\rho = \{y \in Y : \|y\| \leqq \rho\},$$

and let us introduce the following assumption for (P'):

BOUNDEDNESS ASSUMPTION. *There exists $\rho > 0$ such that $\Delta(y)$ is nonempty for all $y \in \mathcal{B}^\rho$ and*

$$M := \inf_{y \in \mathcal{B}^\rho} \sup_{x \in \Delta(y)} f(x) > -\infty.$$

THEOREM A.1. *If S is a closed subset of a finite dimensional space and the boundedness hypothesis is satisfied, there exists a solution to* (D'_S).

Proof. If $L(\lambda)$ is ∞ for all $\lambda \in S$, the theorem is trivial, so let us assume that $S \cap \operatorname{dom} L$ is nonempty. Beginning with the definition of the dual functional and utilizing the boundedness assumption,

$$L(\lambda) = \sup\{f(x) + \langle \lambda, g(x)\rangle : x \in X\} \geqq \sup\{f(x) + \langle \lambda, y\rangle : x \in \Delta(y)\} \geqq M + \langle \lambda, y\rangle$$

for each $y \in \mathscr{B}^\rho$. Maximizing over $y \in \mathscr{B}^\rho$, it follows that

$$\|\lambda\|_{Y^*} := \sup\{\langle \lambda, y\rangle : y \in \mathscr{B}^1\} \leqq (L(\lambda) - M)/\rho. \tag{A.1}$$

By the next lemma, L is lower semicontinuous. Since S is a closed subset of a finite dimensional space, (A.1) implies that the level sets

$$\{\lambda \in S : L(\lambda) \leqq \alpha\}$$

are compact. Hence, there exists a solution to (D'_S). □

LEMMA A.2. *L is lower semicontinuous with respect to both the norm topology of Y^* and the weak topology induced on Y^* by Y.*

Proof. This result is essentially contained in Rockafellar's work [50, Thm. 5] or [46, p. 104]. Consider the epigraph set

$$\operatorname{epi} L = \{(\alpha, \lambda) \in R \times Y^* : \alpha \geqq L(\lambda)\}.$$

Alternatively, we can view this set as the intersection of half spaces which are closed in the weak topology induced on Y^* by Y; in particular,

$$\operatorname{epi} L = \bigcap_{x \in X} \{(\alpha, \lambda) \in R \times Y^* : \alpha \geqq f(x) + \langle \lambda, g(x)\rangle\}.$$

Therefore, epi L is closed in both the norm and the weak topologies. Since lower semicontinuity of L is equivalent to the epigraph being closed, the proof is complete. □

Suppose that X is a convex subset of a vector space. We say that g is concave if

$$g(\alpha x_1 + (1-\alpha)x_2) \geqq \alpha g(x_1) + (1-\alpha) g(x_2)$$

for all $x_1, x_2 \in X$ and $0 \leqq \alpha \leqq 1$.

THEOREM A.3. *Suppose that X is a convex subset of a vector space and both f and g are concave. Under the boundedness hypothesis, there exists a solution λ to* (D′) *and*

$$L(\lambda) = v := \text{supremum}\{f(x) : g(x) \geqq 0, x \in X\}.$$

Moreover, for any solution x of (P′), *we have $\langle \lambda, g(x)\rangle = 0$.*

(This last result is called the complementary slackness condition.)

Proof. If $v = \infty$, the result is trivial. Since $v \geqq M > -\infty$ by the boundedness hypothesis, assume that v is finite. Let us define the convex sets:

$$E = \{(\alpha, y) \in R \times Y : \exists x \in \Delta(y) \text{ with } \alpha \leqq f(x)\}$$

and

$$D = \{(\alpha, y) \in R \times Y : \alpha = v, y \geqq 0\}.$$

By the boundedness assumption, $(M-1, 0) \in R \times Y$ is an interior point of E. Separating D from E with a hyperplane gives us $r \in R$ and $\mu \in Y^*$ such that

$$rv + \langle \mu, z\rangle \geqq r\alpha + \langle \mu, y\rangle \tag{A.2}$$

for all $(v, z) \in D$ and $(\alpha, y) \in E$. Clearly, $\mu \geqq 0$. Since $(M-1, 0)$ lies in the interior of E and $(\alpha, 0) \in E$ for all $\alpha < M$, we see that $r > 0$. Dividing by r, setting $\lambda = \mu/r$, and

inserting $(\alpha, y) = (f(x), g(x))$ and $z = 0$, (A.2) gives us

$$v \geqq f(x) + \langle \lambda, g(x) \rangle$$

for all $x \in X$, or equivalently, $L(\lambda) \leqq v$. Since $L(\lambda) \geqq v$ by weak duality, it follows that $L(\lambda) = v$. Finally, if (P) has a solution $x \in X$, then the inequality $v \geqq f(x) + \langle \lambda, g(x) \rangle = v + \langle \lambda, g(x) \rangle$ implies that $\langle \lambda, g(x) \rangle \leqq 0$. Since $\lambda \geqq 0$ and $g(x) \geqq 0$, we conclude that $\langle \lambda, g(x) \rangle = 0$. □

Appendix 2. Integrand regularity. In § 11, we note that the dual integrands $l(x, y, t)$ for Problems I–IV are continuously differentiable in x and y. In these examples, this result amounts to showing that the optimal cost of the quadratic program

$$\begin{aligned} &\text{minimize} \quad x^T Q x + q^T x \\ &\text{subject to} \quad Ax \leqq a, \quad x \in R^n \end{aligned}$$

depends smoothly on q. Here Q and A are matrices and q and a are vectors of the appropriate dimensions. Let us study the more general class of problems

$$\text{(A.3)} \qquad \text{minimize} \quad \{f(x, \xi) : g(x, \xi) \leqq 0, h(x, \xi) = 0, x \in R^n\}$$

where $f : R^n \times R^p \to R$, $g : R^n \times R^p \to R^m$, and $h : R^n \times R^p \to R^l$. Above, $\xi \in R^p$ is a fixed parameter, and the minimization is over $x \in R^n$. Our development is based on Lipschitz properties established earlier for the solution and the multiplier of (A.3). In [15] these properties are verified for quadratic programs, and in [15, Appendix], we indicate that these results extend to more general programs. This extension is now presented; as a corollary, we show that the optimal cost of (A.3) depends smoothly on ξ.

Suppose that (A.3) has a unique solution $x(\xi)$ for ξ near 0. Under fairly weak assumptions, it has been shown [42] that the feasible set

$$\{x \in R^n : g(x, \xi) \leqq 0, h(x, \xi) = 0\}$$

is stable with respect to perturbations in ξ and hence by [44] $x(\xi)$ is a continuous function of ξ. Defining $z = x(0)$, we assume that $f(x, \xi)$, $g(x, \xi)$, and $h(x, \xi)$ are continuously Fréchet differentiable in x near $(x, \xi) = (z, 0)$. If $g_B(x, \xi)$ is the vector composed of g's components satisfying $g_i(x, \xi) = 0$, we also assume that the rows of

$$\begin{bmatrix} \nabla_1 g_B(x(\xi), \xi) \\ \nabla_1 h(x(\xi), \xi) \end{bmatrix}$$

are linearly independent for ξ near 0. Under these hypotheses, there exist unique multipliers $\lambda(\xi) \in R^m$ and $\mu(\xi) \in R^l$ satisfying the Kuhn–Tucker conditions [29, p. 233]:

$$\text{(A.4)} \qquad \nabla_1 \mathscr{L}(x(\xi), \xi) = 0, \quad \lambda(\xi) \geqq 0, \quad \lambda(\xi)^T g(x(\xi), \xi) = 0$$

where

$$\mathscr{L}(x, \xi) = f(x, \xi) + \lambda(\xi)^T g(x, \xi) + \mu(\xi)^T h(x, \xi).$$

LEMMA A.4. *If $x(\xi)$ is a continuous function of ξ near zero and the differentiability and independence assumptions stated above hold, then $\lambda(\xi)$ and $\mu(\xi)$ are continuous functions of ξ near zero.*

Proof. Let ξ be a fixed parameter near zero, let $I \subset \{1, \cdots, m\}$ be the set of indices i for which $g_i(x(\xi), \xi) = 0$, and let g_I be the vector with components g_i, $i \in I$. Consider the system

$$\text{(A.5)} \qquad \nabla_1 f(x, \eta) + \nabla_1 g_I(x, \eta)^T \lambda_I + \nabla_1 h(x, \eta)^T \mu = 0$$

in the unknowns x, λ_I, and μ. Since $x(\eta)$ depends continuously on η, $g_i(x(\eta), \eta) < 0$ if $i \notin I$ and η is near ξ. Assume that $|\eta - \xi|$ is so small that $g_i(x(\eta), \eta) < 0$ if $i \notin I$. By the Kuhn–Tucker conditions (A.4), $\lambda_i(\eta) = 0$ if $i \notin I$ and $(x(\eta), \lambda_1(\eta), \mu(\eta))$ satisfies (A.5). Since the rows of $\nabla_1 g_B(x(\xi), \xi)$ and $\nabla_1 h(x(\xi), \xi)$ are linearly independent and $x(\eta)$ is a continuous function of η near zero, it follows that the rows of $\nabla_1 g_I(x(\eta), \eta)$ and $\nabla_1 h(x(\eta), \eta)$ are uniformly independent of η near ξ. Hence (A.5) implies that $\lambda_I(\eta)$ and $\mu(\eta)$ are continuous functions of η near ξ. Since $\lambda_i(\eta) = 0$ if $i \notin I$, we conclude that

$$\lim_{\eta \to \xi} (x(\eta), \lambda(\eta), \mu(\eta)) = (x(\xi), \lambda(\xi), \mu(\xi)). \qquad \square$$

THEOREM A.5. *In addition to the hypotheses of Lemma* A.4, *we assume*:

(i) *f, g and h have partial derivatives $\partial^2/\partial x^2, \partial^2/\partial x\, \partial\xi$ and $\partial/\partial\xi$ which are continuous near $(z, 0)$, and*

(ii) *for each ξ near zero, we have*

$$y^T \nabla_{xx} \mathscr{L}(x(\xi), \xi) y > 0 \tag{A.6}$$

for every nonzero vector y such that

$$\nabla_1 g_B(x(\xi), \xi) y = 0 = \nabla_1 h(x(\xi), \xi) y.$$

Then $(x(\xi), \lambda(\xi), \mu(\xi))$ is a Lipschitz continuous function of ξ near zero.

Proof. Again, let ξ be a fixed parameter near zero, let $I \subset \{1, \cdots, m\}$ be the set of indices i for which $g_i(x(\xi), \xi) = 0$, and let g_I be the vector with components g_i, $i \in I$. Consider the system

$$\begin{aligned} &\nabla_1 f(x, \eta) + \nabla_1 g_I(x, \eta)^T \lambda_I + \nabla_1 h(x, \eta)^T \mu = 0, \\ &g_I(x, \eta) = 0, \\ &h(x, \eta) = 0, \end{aligned} \tag{A.7}$$

in the unknowns x, λ_I, and μ. Since (A.6) holds and the rows of $\nabla_1 g_B(x(\xi), \xi)$ and $\nabla_1 h(x(\xi), \xi)$ are linearly independent, it follows from [15, Lemma 3.2] that the Jacobian of the system (A.7) with respect to (x, λ_I, μ) is nonsingular at $\eta = \xi$ and $(x, \lambda_I, \mu) = (x(\xi), \lambda_I(\xi), \mu(\xi))$. By the implicit function theorem, (A.7) has a unique solution $(x, \lambda_I, \mu)(\eta)$ for η in a neighborhood of ξ which is a continuously differentiable function of η. Now, as ξ ranges over a neighborhood of zero, the solution $x(\xi)$ and the multipliers $\lambda(\xi)$ and $\mu(\xi)$ for the program (A.3) satisfy (A.7) for different choices of I. As in [15, § 3], it follows from [15, Thm. 2.3] that $(x(\xi), \lambda(\xi), \mu(\xi))$ is a Lipschitz continuous function of ξ near zero. $\square$

In a related paper [43], Robinson shows that the Kuhn–Tucker conditions (A.4) have a solution depending Lipschitz continuously on a parameter. His assumptions are similar to ours except that (A.6) is strengthened slightly while the assumption that (A.3) has a unique solution is dropped. Now consider the optimal cost $f(x(\xi), \xi)$. Since $x(\xi)$ depends Lipschitz continuously on ξ, we might expect that $f(x(\xi), \xi)$ depends just Lipschitz continuously on ξ. But the cost is smoother than expected:

COROLLARY A.6. *Under the hypotheses of Theorem* A.5, *$f(x(\xi), \xi)$ is a continuously differentiable function of ξ near 0. Moreover, if f, g and h have continuous second partial derivatives near $(z, 0)$, then the derivative of $f(x(\xi), \xi)$ is Lipschitz continuous.*

Proof. Since $x(\cdot)$ is differentiable almost everywhere, the chain rule gives us

$$\frac{\partial}{\partial\xi}f(x(\xi),\xi)=\frac{\partial}{\partial\xi}\mathscr{L}(x(\xi),\xi)$$
$$=\nabla_1\mathscr{L}(x(\xi),\xi)\frac{\partial x(\xi)}{\partial\xi}+g(x(\xi),\xi)^T\frac{\partial\lambda(\xi)}{\partial\xi}+h(x(\xi),\xi)^T\frac{\partial\mu(\xi)}{\partial\xi}$$
$$+\frac{\partial}{\partial\xi}\{f(x,\xi)+\lambda^Tg(x,\xi)+\mu^Th(x,\xi)\}\Big|_{\substack{x=x(\xi)\\ \lambda=\lambda(\xi)\\ \mu=\mu(\xi)}}.$$

By the Kuhn–Tucker conditions, $\nabla_1\mathscr{L}(x(\xi),\xi)=0$, and since $x(\xi)$ is feasible in (A.3), $h(x(\xi),\xi)=0$. Suppose that $g_i(x(\xi),\xi)<0$. Since $x(\eta)$ depends continuously on η, $g_i(x(\eta),\eta)<0$ for η near ξ. Hence the Kuhn–Tucker conditions also tell us that $\lambda_i(\eta)=0$ for η near ξ, and

$$\frac{\partial\lambda_i(\xi)}{\partial\xi}=0.$$

Combining these observations,

$$\frac{\partial f}{\partial\xi}(x(\xi),\xi)=\frac{\partial}{\partial\xi}\{f(x,\xi)+\lambda^Tg(x,\xi)+\mu^Th(x,\xi)\}\Big|_{\substack{x=x(\xi)\\ \lambda=\lambda(\xi)\\ \mu=\mu(\xi)}}. \tag{A.8}$$

Theorem A.5 completes the proof. □

Gauvin and Tolle [12] obtain (A.8) under weaker assumptions, although the feasible set is required to satisfy a uniform compactness condition. Also Armacost and Fiacco [2] give (A.8), but require the so-called strict complementary slackness condition which is not satisfied in our applications to the dual integrand.

Appendix 3. Exact solutions. Consider the problem

$$\text{minimize}\quad \frac{1}{2}\int_0^1[x(t)^2+u(t)^2]\,dt$$
$$\text{subject to}\quad x'(t)=u(t),\quad u(t)\leqq 1\quad\text{almost everywhere},$$
$$x(0)=\frac{1+3e}{2(1-e)},\quad (x,u)\in\mathscr{L}.$$

This problem's solution, computed in [23], is given below.

Region 1. $0\leqq t\leqq\frac{1}{2}$.

$$x(t)=t+\frac{1+3e}{2(1-e)},\quad u(t)=1,\quad p(t)=\frac{t^2}{2}+\frac{(1+3e)}{2(1-e)}t+\frac{13e-5}{8(e-1)}.$$

Region 2. $\frac{1}{2}\leqq t\leqq 1$.

$$x(t)=\frac{e^t+e^{2-t}}{\sqrt{e}(1-e)},\qquad u(t)=p(t)=\frac{e^t-e^{2-t}}{\sqrt{e}(1-e)}.$$

The optimal cost is

$$\frac{55e^2-2e-5}{48(e-1)^2}.$$

Next, consider the problem

$$\text{minimize} \quad \frac{1}{2}\int_0^1 [x(t)^2+u(t)^2]\,dt$$

$$\text{subject to} \quad x'(t)=u(t), \quad u(t)\leqq 1 \quad \text{almost everywhere},$$

$$x(t)\leqq \frac{2\sqrt{e}}{1-e} \quad \text{for all } t\in[0,1],$$

$$x(0)=\frac{5e+3}{4(1-e)}, \quad (x,u)\in\mathscr{L}.$$

This problem's solution, computed in [23], is given below.

Region 1. $0\leqq t\leqq \frac{1}{4}$.

$$x(t)=t-\frac{1}{4}+\frac{1+e}{1-e}, \quad u(t)=1, \quad p(t)=\frac{33}{32}+\frac{1+e}{1-e}\left(t-\frac{1}{4}\right)+\frac{1}{4}t(2t-1).$$

Region 2. $\frac{1}{4}\leqq t\leqq \frac{3}{4}$.

$$x(t)=\frac{e^{t-1/4}}{1-e}(1+e^{3/2-2t}), \qquad u(t)=p(t)=\frac{e^{t-1/4}}{1-e}(1-e^{3/2-2t}).$$

Region 3. $\frac{3}{4}\leqq t\leqq 1$.

$$x(t)=\frac{2\sqrt{e}}{1-e}, \qquad u(t)=p(t)=0.$$

The optimal cost is

$$\frac{49}{384}+\frac{e+1}{2(e-1)}+\frac{x(0)}{32}+\frac{x(0)^2+b^2}{8}$$

where $b=2\sqrt{e}/(1-e)$.

Acknowledgments. The authors are grateful for many helpful comments from the referee. In particular, a streamlined proof of Lemma 3.2 and Theorem 4.1 is mainly due to the referee. Experiments with the IMSL conjugate gradient routine reported at the start of § 11 were conducted by Y. Ting and E. Yeh, graduate students at the Pennsylvania State University.

REFERENCES

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

[2] R. L. ARMACOST AND A. V. FIACCO, *Sensitivity analysis for parametric nonlinear programming using penalty methods*, in Computers and Mathematical Programming, Special Publication 502, National Bureau of Standards, Washington, DC, pp. 261–269.

[3] A. V. BALAKRISHNAN, *On a new computing technique in optimal control*, this Journal, 6 (1968), pp. 149–173.

[4] D. P. BERTSEKAS, *Multiplier methods: A survey*, in Proc. IFAC 6th Triennial World Congress, Boston, 1975.

[5] W. E. BOSARGE, JR. AND O. G. JOHNSON, *Error bounds of high order accuracy for the state regulator problem via piecewise polynomial approximations*, this Journal, 9 (1971), pp. 15–28.

[6] W. E. BOSARGE, JR., O. G. JOHNSON AND C. L. SMITH, *A direct method approximation to the linear parabolic regulator problem over multivariate spline bases*, SIAM J. Numer. Anal., 10 (1973), pp. 35–49.

[7] C. CASTAING, *Sur les mult-applications mesurables*, Thèse, Univ. de Caën, France, 1967.
[8] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1980.
[9] R. COURANT, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.
[10] J. CULLUM, *Penalty functions and nonconvex continuous optimal control problems*, in Computing Methods in Optimization Problems, 2, L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 55–67.
[11] A. V. FIACCO and W. P. HUTZLER, *Basic results in the development of sensitivity and stability analysis in nonlinear programming*, Computers and Operations Research, to appear.
[12] J. GAUVIN AND J. W. TOLLE, *Differential stability in nonlinear programming*, this Journal, 15 (1977), pp. 294–311.
[13] J. GUDDAT, *On some actual questions in parametric optimization*, in Mathematical Methods in Operations Research, Bulgarian Academy of Sciences, Sofia, 1981, pp. 39–54.
[14] W. W. HAGER, *The Ritz–Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.
[15] ———, *Lipschitz continuity for constrained processes*, this Journal, 17 (1979), pp. 321–338.
[16] ———, *Convex control and dual approximations*, Control Cybernet., 8 (1979), *Part* I: pp. 5–22, *Part* II: pp. 73–86.
[17] ———, *Inequalities and approximation*, in Constructive Approaches to Mathematical Models, C. V. Coffman and G. J. Fix, eds., Academic Press, New York, 1979, pp. 189–202.
[18] W. W. HAGER AND G. D. IANCULESCU, *Semi-dual approximations in optimal control: Quadratic cost*, in Free Boundary Problems, Vol. II (Pavia, 1979), Ist. Naz. Alta Mat. Francesco Severi, Rome, 1980, pp. 321–332.
[19] W. W. HAGER AND S. K. MITTER, *Lagrange duality theory for convex control problems*, this Journal, 14 (1976), pp. 843–856.
[20] W. W. HAGER AND G. STRANG, *Free boundaries and finite elements in one dimension*, Math. Comp., 29 (1975), pp. 1020–1031.
[21] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
[22] L. HÖRMANDER, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1969.
[23] G. D. IANCULESCU, *Semi-dual approximations for convex optimal control problems*, Ph.D. dissertation, Carnegie–Mellon Univ., Pittsburgh, PA, 1979.
[24] D. H. JACOBSON AND M. M. LELE, *A transformation technique for optimal control problems with a state variable inequality constraint*, IEEE Trans. Automat. Control, 14 (1969), pp. 457–464.
[25] L. S. LASDON, A. D. WARREN AND R. K. RICE, *An interior penalty method for inequality constrained optimal control problems*, IEEE Trans. Automat. Control, 12 (1967), pp. 388–395.
[26] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control*, John Wiley, New York, 1967.
[27] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, S. K. Mitter, transl., Springer-Verlag, New York, 1971.
[28] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[29] ———, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.
[30] K. MALANOWSKI, *On the regularity of solutions to optimal control problems for systems with control appearing linearly*, Arch. Automat. Telemech., 23 (1978), pp. 227–242.
[31] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
[32] F. H. MATHIS AND G. W. REDDIEN, *Ritz–Trefftz approximations in optimal control*, this Journal, 17 (1979), pp. 307–310.
[33] R. K. MEHRA AND R. E. DAVIS, *A generalized gradient method for optimal control problems with inequality constraints and singular arcs*, IEEE Trans. Automat. Control, 17 (1972), pp. 69–79.
[34] E. MICHAEL, *Continuous selections*, I, Ann. of Math., 63 (1956), pp. 361–382.
[35] J. MOSSINO, *An application of duality to distributed optimal control problems with constraints on the state and the control*, J. Math. Anal. Appl., 50 (1975), pp. 223–242.
[36] ———, *Approximation numérique de problèmes de contrôle optimal avec contrainte sur le contrôle et sur l'état*, Calcolo, 13 (1976), pp. 21–62.
[37] J. A. NITSCHE, *Ein Kriterium für die Quasi-optimalitat des Ritzchen Verfahrens*, Numer. Math., 11 (1968), pp. 346–348.
[38] J. T. ODEN AND J. N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements*, Interscience, New York, 1976.
[39] O. PIRONNEAU AND E. POLAK, *A dual method for optimal control problems with initial and final boundary constraints*, this Journal, 11 (1973), pp. 534–549.
[40] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1965.

[41] M. J. D. POWELL, *A method of nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1972.
[42] S. M. ROBINSON, *Stability theory for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 12 (1976), pp. 497–513.
[43] ———, *Strongly regular generalized equations*, Mathematics Research Center Report 1877, Univ. of Wisconsin, Madison, WI, 1978.
[44] S. M. ROBINSON AND R. H. DAY, *A sufficient condition for continuity of optimal sets in mathematical programming*, J. Math. Anal. Appl., 45 (1974), pp. 506–511.
[45] R. T. ROCKAFELLAR, *Integrals which are convex functionals*, II, Pacific J. Math., 39 (1971), pp. 439–469.
[46] ———, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1972.
[47] ———, *State constraints in convex control problems of Bolza*, this Journal, 10 (1972), pp. 691–715.
[48] ———, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.
[49] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.
[50] ———, *Conjugate Duality and Optimization*, CBMS Regional Conference Series in Applied Mathematics 16, Society for Industrial and Applied Mathematics, Philadelphia, 1974.
[51] ———, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, Lucien Waelbroeak, ed., Springer-Verlag, New York, 1976, pp. 157–207.
[52] ———, *Duality in optimal control*, in Mathematical Control Theory, W. A. Coppel, ed., Springer-Verlag, New York, 1978, pp. 219–257.
[53] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[54] D. L. RUSSELL, *Penalty functions and bounded phase coordinate control*, this Journal, 2 (1965), pp. 409–422.
[55] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[56] D. D. THOMPSON AND R. A. VOLZ, *The linear quadratic cost problem with linear state constraints and the nonsymmetric Riccati equation*, this Journal, 13 (1975), pp. 110–145.
[57] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as added side conditions*, in Contributions to the Calculus of Variations, Univ. of Chicago Press, Chicago, 1937, pp. 407–448.