JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 28, No. 3, JULY 1979

A Combination of Penalty Function and Multiplier Methods for Solving Optimal Control Problems[1]

S. T. GLAD[2]

Communicated by M. R. Hestenes

Abstract. The properties of combined multiplier and penalty function methods are investigated using a second-order expansion and results known for the Riccati equation. It is shown that the lower bound of the values of the penalty constant necessary to obtain a minimum is given by a certain Riccati equation. The convergence rate of a common updating rule for the multipliers is shown to be linear.

Key Words. Optimal control, multiplier methods, penalty functions, Riccati equation, convergence rate.

1. Introduction

Recently, methods for solving optimal control problems that do not require the explicit solution of the differential equations have received some attention. Hestenes has suggested (Refs. 1 and 2) a method that is a combination of multiplier and penalty function methods. Rupp (Ref. 3) and Di Pillo *et al.* (Ref. 4) have used similar techniques. The combined penalty function and multiplier approach has also been used by Rupp (Ref. 5) for isoperimetric constraints and by Nahra (Ref. 6), Mårtensson (Ref. 7), and O'Doherty and Pierson (Ref. 8) for terminal constraints. Properties of these methods will be investigated here, using a second-order expansion and properties of the Riccati equation. The properties of the free endpoint problem are considered in Section 3, and those of the fixed endpoint problem are considered in Section 4. Finally, convergence properties of an iterative method for both free endpoint and fixed endpoint problems are investigated in Section 5.

[1] This work has been supported by the Swedish Institute of Applied Mathematics.

[2] Research Associate, Department of Automatic Control, Lund University, Lund, Sweden.

2. Problem Formulation

The optimal control problem to be studied here can be formulated as follows. Minimize the functional

$$I(x, u)=\int_0^T L(x(t), u(t), t)\, dt+F(x(T)), \tag{1}$$

subject to

$$\dot{x}(t)=f(x(t), u(t), t), \qquad 0\leq t\leq T,$$

$$x(0)=a, \qquad \psi(x(T))=0.$$

Here, x and u are functions belonging to $C_1^n[0, T]$ and $C_0^m[0, T]$, respectively. $C_0^k[0, T]$ denotes the class of continuous k-vector-valued functions on $[0, T]$, while $C_1^k[0, T]$ denotes the class of continuously differentiable functions. The functions f and ψ are vector-valued with n and r components, respectively.

The following assumptions are made.

(i) L and f are three times continuously differentiable with respect to x and u.

(ii) L and f, together with their first and second derivatives with respect to x and u, are continuous with respect to t.

(iii) F and ψ are three times continuously differentiable.

(iv) The minimization problem has a solution, denoted by $(\bar{x}(t), \bar{u}(t))$.

Define the Hamiltonian

$$H(x(t), u(t), p(t), t)=L(x(t), u(t), t)+p^T(t)f(x(t), u(t), t), \tag{2}$$

where p is a continuous function of time. The standard necessary conditions for the optimal control problem are given by the following theorem.

Theorem 2.1. Let $\bar{x}$, $\bar{u}$ be the solution to the problem defined by (1), and assume that the following regularity conditions are satisfied.

(i) The matrix $\psi_x(\bar{x}(T))$ has rank r.

(ii) Given any vector z, it is possible to find a continuous function v such that

$$\dot{h}(t)=f_x(\bar{x}(t), \bar{u}(t), t)h(t)+f_u(\bar{x}(t), \bar{u}(t), t)v(t), \qquad h(0)=0,$$

has a solution satisfying

$$h(T)=z,$$

i.e., the linearized system is controllable. Then, there is an n-dimensional vector-valued function $\bar{p}(t)$ and an r-dimensional vector $\bar{b}$ such that, for all

$t \in [0, T]$,

$$\begin{aligned}-\dot{\bar{p}}(t) &= H_x(\bar{x}(t), \bar{u}(t), \bar{p}(t), t),\\ \bar{p}(T) &= F_x^T(\bar{x}(T)) + \psi_x^T(\bar{x}(T))\bar{b},\\ H_u(\bar{x}(t), \bar{u}(t), \bar{p}(t), t) &= 0.\end{aligned} \tag{3}$$

Proof. See Luenberger (Ref. 9).

3. Optimal Control Problem with Only Differential Equation Constraints

In this case, the problem can be written as follows. Minimize the functional

$$I(x, u) = \int_0^T L(x(t), u(t), t)\, dt + F(x(T)),$$

subject to

$$\begin{aligned}\dot{x}(t) &= f(x(t), u(t), t), \qquad 0 \leq t \leq T,\\ x(0) &= a.\end{aligned} \tag{4}$$

The idea in Hestenes (Ref. 1 and Ref. 2) is to form the augmented function

$$\begin{aligned}J(x, u, p, c) = \int_0^T &\{L(x(t), u(t), t) + p^T(t)[f(x(t), u(t), t) - \dot{x}(t)]\\ &+ (c/2)[f(x(t), u(t), t) - \dot{x}(t)]^T\\ &\cdot [f(x(t), u(t), t) - \dot{x}(t)]\}\, dt + F(x(T)).\end{aligned} \tag{5}$$

Here, c is a positive real number and p is a continuous function. The functions x and u are now allowed to take arbitrary values, not necessarily satisfying the differential equation $\dot{x} = f(x, u, t)$. The condition $x(0) = a$, however, is still applied.

It has been proved by Hestenes (Ref. 1) that $J(x, u, \bar{p}, c)$ has a local minimum at $(\bar{x}, \bar{u})$ if c is large enough. Here, a different proof, based on the Riccati equation, will be given. It has the advantage of showing what the lower bound of c is. The connection with the sufficiency conditions in Bryson and Ho (Ref. 10) is also given.

To show that $J(x, u, \bar{p}, c)$ has a local minimum at $(\bar{x}, \bar{u})$, an expansion is used. Let

$$x(t) = \bar{x}(t) + h(t), \qquad u(t) = \bar{u}(t) + k(t).$$

Since the function x belongs to $C_1^n[0, T]$, with $x(0) = a$, and u belongs to $C_0^m[0, T]$, only variations h and k satisfying $h(0) = 0$, with h belonging to $C_1^n[0, T]$ and k belonging to $C_0^m[0, T]$, are of interest. The functions h and k satisfying these conditions will be called admissible. The following norms are used:

$$\|h\|_1 = \sup_{0 \leq t \leq T} \|h(t)\| + \sup_{0 \leq t \leq T} \|\dot{h}(t)\|,$$

$$\|k\|_0 = \sup_{0 \leq t \leq T} \|k(t)\|,$$

where $\|\cdot\|$ denotes an arbitrary vector norm. In what follows, x and u will often be written in place of $x(t)$ and $u(t)$, to simplify the notation.

The expansion can now be written as

$$\begin{aligned} J(\bar{x}+h, \bar{u}+k, \bar{p}, c) = \int_0^T & \{H(\bar{x}+h, \bar{u}+k, \bar{p}, t) - p^T(\dot{x}+\dot{h}) \\ & + (c/2)[f(\bar{x}+h, \bar{u}+k, t) - \dot{x} - \dot{h}]^T[f(\bar{x}+h, \bar{u}+k, t) \\ & - \dot{x} - \dot{h}]\}\, dt + F(\bar{x}(T) + h(T)), \end{aligned}$$

where the Hamiltonian

$$H = L + p^T f$$

is used. Expanding H, f, F in a Taylor series gives

$$\begin{aligned} J(\bar{x}+h, \bar{u}+k, \bar{p}, c) = J(\bar{x}, \bar{u}, \bar{p}, c) + \int_0^T & (H_x h + H_u k - \bar{p}^T \dot{h})\, dt + F_x h(T) \\ & + \frac{1}{2} \int_0^T (h^T H_{xx} h \\ & + 2h^T H_{xu} k + k^T H_{uu} k)\, dt \\ & + \frac{1}{2} \int_0^T c(h^T f_x^T f_x h + k^T f_u^T f_u k + \dot{h}^T \dot{h} + 2h^T f_x^T f_u k \\ & \qquad - 2h^T f_x^T \dot{h} - 2k^T f_u^T \dot{h})\, dt \\ & + \tfrac{1}{2} h^T(T) F_{xx} h(T) + R(h, k), \end{aligned}$$

where H_x, H_u, H_{xx}, etc., are evaluated along $(\bar{x}, \bar{u})$,

$$|R(h, k)| \leq \varepsilon(h, k) \int_0^T (h^T h + \dot{h}^T \dot{h} + k^T k)\, dt,$$

and $\varepsilon(h, k) \to 0$ as $(h, k) \to 0$ in the norms given above.

Since $\bar{p}$ satisfies the conditions of Theorem 2.1, it follows from an integration by parts that the linear terms disappear. Then,

$$\begin{aligned} J(\bar{x}+h, \bar{u}+k, \bar{p}, c) - J(\bar{x}, \bar{u}, \bar{p}, c) = {} & \frac{1}{2}\int_0^T \{h^T(H_{xx} + cf_x^Tf_x)h \\ & + 2h^T(H_{xu} + cf_x^Tf_u)k \\ & + k^T(H_{uu} + cf_u^Tf_u)k + c\dot{h}^T\dot{h} \\ & - 2ch^Tf_x^T\dot{h} - 2ck^Tf_u^T\dot{h}\}\, dt \\ & + h^T(T)F_{xx}h(T) + R(h, k) \\ = {} & \delta^2 J(h, k) + R(h, k). \end{aligned}$$

To prove that $\delta^2 J$ is positive, we transform it into a perfect square. First, observe that, for any continuously differentiable matrix function S, it is true that

$$\int_0^T (h^T\dot{S}h + 2h^TS\dot{h})\, dt - h^T(T)S(T)h(T) = 0$$

for all continuously differentiable h satisfying

$$h(0) = 0.$$

The addition of a term of this form to get a perfect square is used in the calculus of variations; see Gelfand and Fomin (Ref. 11). Adding this quantity to $\delta^2 J$ gives

$$\begin{aligned} \delta^2 J = {} & \frac{1}{2}\int_0^T \{h^T(H_{xx} + cf_x^Tf_x + \dot{S})h + 2h^T(H_{xu} + cf_x^Tf_u)k \\ & + k^T(H_{uu} + cf_u^Tf_u)k + c\dot{h}^T\dot{h} + 2h^T(S - cf_x^T)\dot{h} - 2ck^Tf_u^T\dot{h}\}\, dt \\ & + \tfrac{1}{2}h^T(T)[F_{xx} - S(T)]h(T) \\ = {} & \frac{1}{2}\int_0^T \begin{bmatrix} k + H_{uu}^{-1}(H_{ux} + f_u^TS)h \\ \dot{h} + [f_uH_{uu}^{-1}(H_{ux} + f_u^TS) + (1/c)S - f_x]h \end{bmatrix}^T \\ & \cdot \begin{bmatrix} H_{uu} + cf_u^Tf_u & -cf_u^T \\ -cf_u & cI \end{bmatrix} \begin{bmatrix} k + H_{uu}^{-1}(H_{ux} + f_u^TS)h \\ \dot{h} + [f_uH_{uu}^{-1}(H_{ux} + f_u^TS) + (1/c)S - f_x]h \end{bmatrix} dt \\ & + \frac{1}{2}\int_0^T h^T[\dot{S} + H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_uH_{uu}^{-1}H_{ux})^TS \\ & + S(f_x - f_uH_{uu}^{-1}H_{ux}) - Sf_uH_{uu}^{-1}f_u^TS - (1/c)S^2]\, dt \\ & + \tfrac{1}{2}h^T(T)[F_{xx} - S(T)]h(T). \end{aligned} \tag{6}$$

Here, it is assumed that H_{uu} is nonsingular. Now, the following theorem is an immediate consequence.

Theorem 3.1. Let $(\bar{x}, \bar{u})$ be a solution to (1), and let $\bar{p}$ satisfy Eq. (3) of Theorem 2.1. Also, assume that $c > 0$, and

$$H_{uu}(\bar{x}(t), \bar{u}(t), \bar{p}(t), t) > 0, \qquad t \in [0, T],$$

and that the Riccati equation

$$\begin{aligned} -\dot{S} &= H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_u H_{uu}^{-1} H_{ux})^T S \\ &\quad + S(f_x - f_u H_{uu}^{-1} H_{ux}) - S[f_u H_{uu}^{-1} f_u^T + (1/c)I]S, \qquad S(T) = F_{xx}, \end{aligned} \tag{7}$$

where H_{xx}, etc., are evaluated along $\bar{x}$, $\bar{u}$, has a solution over the whole interval $[0, T]$. Then, $\delta^2 J(h, k) > 0$ for all admissible h and k that are not both identically zero.

Proof. First, it is shown that the matrix

$$\begin{bmatrix} H_{uu} + cf_u^T f_u & -cf_u^T \\ -cf_u & cI \end{bmatrix}$$

is positive definite for all $c > 0$. Form the expression

$$[z^T, w^T] \begin{bmatrix} H_{uu} + cf_u^T f_u & -cf_u^T \\ -cf_u & cI \end{bmatrix} \begin{bmatrix} z \\ w \end{bmatrix} = z^T H_{uu} z + c(f_u z - w)^T (f_u z - w) \geq 0.$$

Equality is attained only for $z = 0$, $w = 0$.

Since the second and third terms of $\delta^2 J$ in (6) disappear, it follows that $\delta^2 J \geq 0$. If $\delta^2 J = 0$, then

$$k + H_{uu}^{-1}(H_{ux} + f_u^T S)h = 0, \qquad 0 \leq t \leq T,$$

$$\dot{h} + [f_u H_{uu}^{-1}(H_{ux} + f_u^T S) + (1/c)S - f_x]h = 0, \qquad 0 \leq t \leq T.$$

Since $h(0) = 0$, it follows from the uniqueness theorem for linear differential equations that $h(t)$ is identically zero. Then, k is also identically zero. □

To show that J has a minimum at $(\bar{x}, \bar{u})$, it is not enough to know that $\delta^2 J$ is positive. It must also dominate the higher-order terms. The result of Theorem 3.1 can, however, be strengthened.

Theorem 3.2. With the same assumptions as in Theorem 3.1, there exists a constant $\eta > 0$ such that

$$\delta^2 J(h, k) \geq \eta \int_0^T (h^T h + \dot{h}^T \dot{h} + k^T k)\, dt.$$

Proof. Consider the expression

$$A(h, k, \eta) = \delta^2 J(h, k) - \eta \int_0^T (h^T h + \dot{h}^T \dot{h} + k^T k)\, dt,$$

where $\eta > 0$. The value of $A(h, k, \eta)$ is the same as the value of $\delta^2 J(h, k)$ with

$$h^T(H_{xx} + cf_x^T f_x)h \text{ replaced by } h^T(H_{xx} + cf_x^T f_x - \eta I)h,$$
$$c\dot{h}^T\dot{h} \text{ replaced by } (c-\eta)\dot{h}^T\dot{h}$$
$$k^T(H_{uu} + cf_u^T f_u)k \text{ replaced by } k^T(H_{uu} + cf_u^T f_u - \eta I)k.$$

It then follows from Lemma 7.5 in Appendix A that, if η is chosen sufficiently small, then the Riccati equation corresponding to $A(h, k, \eta)$ exists over the interval $[0, T]$. Since the matrix

$$\begin{bmatrix} H_{uu} + cf_u^T f_u - \eta I & -cf_u^T \\ -cf_u & (c-\eta)I \end{bmatrix}$$

is still positive definite for sufficiently small η, it follows that $A(h, k, \eta) \geq 0$, and the theorem is proved. □

This leads directly to the following result, showing that J has a local minimum at $(\bar{x}, \bar{u})$.

Theorem 3.3. If the assumptions of Theorem 3.1 are satisfied, then

$$J(\bar{x} + h, \bar{u} + k, \bar{p}, c) > J(\bar{x}, \bar{u}, \bar{p}, c)$$

for all admissible h and k, not both identically zero, and with $\|h\|_1$ and $\|k\|_0$ sufficiently small.

Proof. From Theorem 3.2, one has

$$\begin{aligned} &J(\bar{x} + h, \bar{u} + k, \bar{p}, c) - J(\bar{x}, \bar{u}, \bar{p}, c) \\ &\quad = \delta^2 J(h, k) + R(h, k) \\ &\quad \geq [\eta - |\varepsilon(h, k)|] \int_0^T (h^T h + \dot{h}^T \dot{h} + k^T k)\, dt > 0, \end{aligned}$$

if $\|h\|_1$ and $\|k\|_0$ are sufficiently small and h and k are not both identically zero. □

It is interesting to note that the magnitude of c that is required depends only on the Riccati equation (7), provided $c > 0$. The interesting question is of course: is there any c for which (7) has a solution over $[0, T]$? First, note the following result.

Theorem 3.4. If (7) has a solution on $[0, T]$ for $c = c_1$, then it has a solution for any $c \geq c_1$.

Proof. Let $c_2 > c_1$, and define

$$P_1 = f_u H_{uu}^{-1} f_u^T + (1/c_1)I,$$
$$P_2 = f_u H_{uu}^{-1} f_u^T + (1/c_2)I.$$

Then,

$$P_1 - P_2 \geq 0.$$

It now follows from Lemma 7.2 in Appendix A that

$$S_2(t) \geq S_1(t),$$

where S_1 and S_2 are the solutions corresponding to c_1 and c_2, respectively. Since, from Lemma 7.4, the only way the solution S can fail to exist on an interval $[t_1, T]$ is by going off to minus infinity, it follows that S_2 exists on any interval where S_1 exists. □

Corollary 3.1. Either there are no values of c for which (7) has a solution on $[0, T]$, or else there is a number c_0 such that S exists on $[0, T]$ for $c > c_0$ and goes to minus infinity for some $t_1 \in [0, T]$ when $c < c_0$.

Proof. Take $c_0 = \inf\{\text{all } c > 0 \text{ such that } S \text{ exists on the whole interval}\}$. □

Theorem 3.5. Let the Riccati equation

$$\begin{aligned} -\dot{S} = H_{xx} - H_{xu} H_{uu}^{-1} H_{ux} + (f_x - f_u H_{uu}^{-1} H_{ux})^T S \\ + S(f_x - f_u H_{uu}^{-1} H_{ux}) - S f_u H_{uu}^{-1} f_u^T S, \qquad S(T) = F_{xx}, \end{aligned} \tag{8}$$

have a solution defined in the whole interval $[0, T]$. Then, there exists a $c_0 \geq 0$ such that (7) also has a solution over $[0, T]$ for all $c > c_0$.

Proof. Since the difference between the matrices $f_u H_{uu}^{-1} f_u^T$ and $[f_u H_{uu}^{-1} f_u^T + (1/c)I]$ can be made arbitrarily small by choosing c large, the result follows from Lemma 7.5 in Appendix A. □

An immediate consequence is the following theorem.

Theorem 3.6. Let $(\bar{x}; \bar{u})$ be the solution to (1) and let $\bar{p}$ satisfy Eq. (3) of Theorem 1.1. Also, assume that

$$H_{uu}(\bar{x}(t), \bar{u}(t), \bar{p}(t), t) > 0, \qquad t \in [0, T],$$

and that the Riccati equation (8) has a solution over $[0, T]$. Then, there exists a constant $c_0 \geq 0$ such that $J(x, u, \bar{p}, c)$ has a local minimum at $(\bar{x}, \bar{u})$ for all $c > c_0$.

Proof. It follows directly from Theorems 3.5 and 3.3. □

The assumptions made in this theorem are the standard second-order sufficiency conditions of problem (1); see, e.g., Bryson and Ho (Ref. 10). If J has a minimum with respect to arbitrary (x, u), then it also has a minimum with respect to the special choice of (x, u) which satisfies the differential equation $\dot{x} = f(x, u, t)$. Since $J = I$ for these (x, u), Theorem 3.6 actually forms an alternative proof of the sufficiency conditions.

So far, it has been shown that, when $H_{uu} > 0$, the existence of a solution to (7) over $[0, T]$ is a sufficient condition for J to have a local minimum at $(\bar{x}, \bar{u})$. The condition is almost necessary in the sense explained in the following theorem.

Theorem 3.7. Let $\bar{p}$ satisfy (3), and assume that $J(x, u, \bar{p}, c)$ has a local minimum at $(\bar{x}, \bar{u})$ for some $c > 0$. Assume that

$$H_{uu}(\bar{x}(t), \bar{u}(t), \bar{p}(t), t) > 0, \qquad t \in [0, T].$$

Then, the Riccati equations (7) and (8) have a solution over $[\varepsilon, T]$ for all $\varepsilon > 0$.

Proof. For J to have a local minimum, it is necessary that $\delta^2 J(h, k) \geq 0$ for all admissible h and k. Since the solution of (7) exists on $[t_1, T]$ for some $t_1 < T$ (local existence theorem for differential equations, see Ref. 12), it follows that

$$\begin{aligned}\delta^2 J(h, k) = {} & \frac{1}{2}\int_0^{t_1} \{h^T(H_{xx} + cf_x^T f_x)h + 2h^T(H_{xu} + cf_x^T f_u)k \\ & + k^T(H_{uu} + cf_u^T f_u)k + c\dot{h}^T\dot{h} - 2ch^T f_x^T \dot{h} - 2ck^T f_u^T \dot{h}\}\, dt \\ & + \frac{1}{2}\int_{t_1}^{T} \begin{bmatrix} k + H_{uu}^{-1}(h_{ux} + f_u^T S)h \\ \dot{h} + [f_u H_{uu}^{-1}(H_{ux} + f_u^T S) + (1/c)S - f_x]h \end{bmatrix} \\ & \cdot \begin{bmatrix} H_{uu} + cf_u^T f_u & -cf_u^T \\ -cf_u & cI \end{bmatrix} \begin{bmatrix} k + H_{uu}^{-1}(H_{ux} + f_u^T S)h \\ \dot{h} + [f_u H_{uu}^{-1}(H_{ux} + f_u^T S) + (1/c)S - f_x]h \end{bmatrix} dt \\ & + \tfrac{1}{2} h^T(t_1) S(t_1) h(t_1).\end{aligned}$$

Now, choose

$$k(t) = 0, \qquad t \in [0, t_1],$$

$$h(t) = (t/t_1)a, \qquad t \in [0, t_1],$$

where a is an arbitrary constant vector, and let h and k be the solutions of

$$k = -H_{uu}^{-1}(H_{ux} + f_u^T S)h,$$

$$\dot{h} = -[f_u H_{uu}^{-1}(H_{ux} + f_u^T S) + (1/c)S - f_x]h, \qquad h(t_1) = a,$$

in $[t_1, T]$. For this choice of h and k, we have

$$\delta^2 J(h, k) = \tfrac{1}{2}a^T \int_0^{t_1} \{t^2(H_{xx} + cf_x^T f_x) + cI - ct(f_x + f_x^T)\}\, dta/t_1^2 + \tfrac{1}{2}a^T S(t_1)a.$$

Since the h and k used here can be approximated arbitrarily well with continuous k and continuously differentiable h, it follows that

$$\delta^2 J(h, k) \geq 0$$

also for this choice of h and k. Then,

$$a^T S(t_1)a \geq -a^T \int_0^{t_1} [t^2(H_{xx} + cf_x^T f_x) + cI - ct(f_x + f_x^T)]\, dta/t_1^2 \tag{9}$$

for any vector a. Now, suppose that S goes to minus infinity for $t = t_2$, $0 < t_2 < T$. Then (9) must be violated for some $t_1 \in [t_2, T]$. Consequently, the solution to (7) exists on $[\varepsilon, T]$ for any ε. From Theorem 3.5, this is true also for the solution to (8). □

Corollary 3.2. If the solution to (7) goes to minus infinity for some t in $(0, T)$, then $J(x, u, \bar{p}, c)$ does not have a local minimum at $(\bar{x}, \bar{u})$.

Example 3.1. Find the shortest distance between a point and a great circle on a unit sphere.

Let the given point be at the origin 0 of a latitude–longitude coordinate system with latitude θ and longitude α, and let the great circle be the meridian $\alpha = \alpha_1$. Then,

$$ds^2 = (d\theta)^2 + (\cos\theta\, d\alpha)^2,$$

and the problem is to minimize

$$I = \int_0^{\alpha_1} \sqrt{(u^2 + \cos^2\theta)}\, d\alpha,$$

where

$$\dot{\theta} = u, \qquad \theta(0) = 0.$$

The Hamiltonian is given by

$$H = \sqrt{(u^2 + \cos^2\theta)} + pu.$$

The first-order necessary conditions are

$$u/\surd(u^2+\cos^2\theta)+p=0,$$
$$\dot{p}=\cos\theta\sin\theta/\surd(u^2+\cos^2\theta), \qquad p(T)=0.$$

They are satisfied by

$$\bar{u}=0, \qquad \bar{\theta}=0, \qquad \bar{p}=0.$$

The second derivatives of H evaluated along $\bar{u}$, $\bar{\theta}$, $\bar{p}$ are

$$H_{uu}=1, \qquad H_{u\theta}=0, \qquad H_{\theta\theta}=-1.$$

The Riccati equation (8) then becomes

$$-dS/d\alpha=-1-S^2, \qquad S(\alpha_1)=0,$$

with solution

$$S(\alpha)=-\tan(\alpha_1-\alpha).$$

The second-order sufficiency conditions are satisfied if

$$0<\alpha_1<\pi/2.$$

The Riccati equation (7) becomes

$$-dS/d\alpha=-1-(1+1/c)S^2, \qquad S(\alpha_1)=0,$$

with the solution

$$S=-\tan[(\alpha_1-\alpha)\surd(1+1/c)]/\surd(1+1/c).$$

The lower bound of c is then

$$c_0=\alpha_1^2/(\pi^2/4-\alpha_1^2) \qquad \text{for } 0<\alpha_1<\pi/2.$$

4. Extension to Terminal Constraints

The problem with terminal constraints can be written as follows. Minimize the functional

$$I(x,u)=\int_0^T L(x(t),u(t),t)\,dt+F(x(T)),$$

subject to

$$\dot{x}(t)=f(x(t),u(t),t),$$
$$x(0)=a, \qquad \psi(x(T))=0.$$

Terminal constraints have been treated by Nahra (Ref. 6), Mårtensson (Ref. 7), and O'Doherty and Pierson (Ref. 8). They replaced $F(x(T))$ by

$$F(x(T))+b^T\psi(x(T))+\tfrac{1}{2}c_2\psi^T(x(T))\psi(x(T))$$

and iterated on the multipliers b. The combination of this idea with the methods of the preceding section will now be studied.
Define

$$\begin{aligned}J(x, u, p, b, c_1, c_2) = \int_0^T &\{L(x, u, t)+p^T[f(x, u, t)-\dot{x}]\\ &+\tfrac{1}{2}c_1[f(x, u, t)-\dot{x}]^T[f(x, u, t)-\dot{x}]\}\, dt\\ &+F(x(T))+b^T\psi(x(T))+\tfrac{1}{2}c_2\psi^T(x(T))\psi(x(T)).\end{aligned}$$

The following theorems, analogous to the ones of Section 2, can be proved.

Theorem 4.1. Let $\bar{p}$ and $\bar{b}$ satisfy Eq. (3). Assume that

$$c_1>0, \qquad c_2\geq 0,$$

$$H_{uu}(\bar{x}(t), \bar{u}(t), \bar{p}(t), t)>0, \qquad t\in[0, T],$$

and that the Riccati equation

$$\begin{aligned}-\dot{S} &= H_{xx}-H_{xu}H_{uu}^{-1}H_{ux}+(f_x-f_uH_{uu}^{-1}H_{ux})^TS+S(f_x-f_uH_{uu}^{-1}H_{ux})\\ &\quad -S[f_uH_{uu}^{-1}f_u^T+(1/c_1)I]S, \qquad (10)\\ S(T) &= F_{xx}+c_2\psi_x^T\psi_x+\Sigma\bar{b}_i(\psi_i)_{xx},\end{aligned}$$

has a solution over $[0, T]$. Then, $J(x, u, \bar{p}, \bar{b}, c_1, c_2)$ has a local minimum at $(\bar{x}, \bar{u})$.

Proof. It follows from Theorems 3.1 to 3.3, with F replaced by

$$F+\bar{b}^T\psi+\tfrac{1}{2}c_2\psi^T\psi.$$

Theorem 4.2. Assume that $J(x, u, \bar{p}, \bar{b}, c_1, c_2)$ has a local minimum at $(\bar{x}, \bar{u})$ and that

$$H_{uu}(\bar{x}, \bar{u}, \bar{p}, t)>0, \qquad t\in[0, T].$$

Then, the Riccati equation (10) has a solution over $[\varepsilon, T]$ for arbitrary $\varepsilon>0$.

Proof. It is analogous to the proof of Theorem 3.7. □

It is interesting to study some special cases. First, let ψ determine $x(T)$ completely.

Theorem 4.3. Let $\psi(x(T))$ have dimension n. Assume that the regularity conditions of Theorem 2.1 hold and that $\bar{p}$ and $\bar{b}$ satisfy Eq. (3). Assume the following.

(i) $H_{uu}(\bar{x}(t), \bar{u}(t), \bar{p}(t), t) > 0, \qquad t \in [0, T]$.

(ii) There exists a symmetric matrix S_0 such that the Riccati equation

$$\begin{aligned} -\dot{S} &= H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_uH_{uu}^{-1}H_{ux})^T S \\ &\quad + S(f_x - f_uH_{uu}^{-1}H_{ux}) - Sf_uH_{uu}^{-1}f_u^T S, \qquad S(T) = S_0, \end{aligned} \tag{11}$$

has a solution in $[0, T]$.

Then, there exist constants $c_1 > 0$ and $c_2 \geq 0$ such that $J(x, u, \bar{p}, \bar{b}, c_1, c_2)$ has a local minimum at $(\bar{x}, \bar{u})$.

Proof. There exists a value of c_2 such that

$$F_{xx} + c_2\psi_x^T\psi_x + \Sigma b_i(\psi_i)_{xx} \geq S_0.$$

The difference between $[f_uH_{uu}^{-1}f_u^T + (1/c_1)I]$ and $f_uH_{uu}^{-1}f_u^T$ can be made arbitrarily small by choosing c_1 large enough. The result then follows from Lemmas 7.1 and 7.5 in Appendix A. □

The simplest type of terminal constraint is $x_i\ (T) = d_i$ for some indices i. For easier notation, assume that the variables are ordered such that

$$\begin{aligned} &x_i(T) = d_i, \qquad i = 1, \ldots, r, \\ &x_i(T) \text{ free}, \qquad i = r+1, \ldots, n. \end{aligned} \tag{12}$$

Theorem 4.4. Let the terminal constraint be given by (12), and assume that $\bar{b}$ and $\bar{p}$ are defined by (3). Assume the following.

(i) $H_{uu}(\bar{x}(t), \bar{u}(t), \bar{p}(t), t) > 0, \qquad t \in [0, T]$.

(ii) There exists an $r \times r$ matrix A such that the Riccati equation (11) with

$$S_0 = \left[\begin{array}{c|c} A & 0 \\ \hline 0 & F_{x_ix_j} \end{array}\right]$$

has a solution on $[0, T]$.

Then, there exist constants c_1 and c_2 such that $J(x, u, \bar{p}, \bar{b}, c_1, c_2)$ has a local minimum at $(\bar{x}, \bar{u})$.

Proof. It is analogous to that of Theorem 4.3. □

Example 4.1. *Shortest Distance Between Two Points on a Sphere.* The difference between this example and Example 3.1 lies in the boundary condition $\theta(\alpha_1)=0$. The Riccati equation (11) becomes

$$-dS/d\alpha=-1-S^2, \qquad S(\alpha_1) \text{ arbitrary},$$

which has the solution

$$S=-\tan(\alpha_0-\alpha),$$

where α_0 can be chosen arbitrarily. To prolong the existence of S as much as possible, α_0 should be taken close to $\alpha_1-\pi/2$, which corresponds to large values of $S(\alpha_1)$. The sufficiency conditions are then satisfied on the interval $0\leq\alpha\leq\pi-\varepsilon$ for any $\varepsilon>0$.

The Riccati equation (10) gives

$$-dS/d\alpha=-1-(1+1/c_1)S^2, \qquad S(\alpha_1)=c_2,$$

with the solution

$$S=-\tan\left[(\alpha_0-\alpha)\sqrt{(1+1/c_1)}\right]/\sqrt{(1+1/c_1)},$$

where

$$\alpha_0=\alpha_1-\arctan[c_2\sqrt{(1+1/c_1)}]/\sqrt{(1+1/c_1)}.$$

The values c_1 and c_2 for which S exists on $[0, \alpha_1]$ are given by

$$c_2+\tan[\pi/2-\alpha_1\sqrt{(1+1/c_1)}]/\sqrt{(1+1/c_1)}\geq 0.$$

Example 4.2. This example is given by Bryson and Ho (Ref. 10). Consider the motion of a rocket in a constant gravitational field. Let x_1 denote the altitude and x_2 the vertical component of the velocity. Assume that the thrust direction forms the angle β with the horizontal and that its magnitude is constant and equal to am, where m is the mass of the rocket. Let g denote the gravitational acceleration. The objective is to choose the control variable β to give the rocket horizontal flight at the altitude h at the time T and to mximize the horizontal velocity component.

The equations of motion are

$$\begin{gathered}\dot{x}_1=x_2,\\ \dot{x}_2=a\sin\beta-g,\\ x_1(0)=0, \qquad x_2(0)=0,\\ x_1(T)=h, \qquad x_2(T)=0,\end{gathered}$$

and the loss function is

$$J=-a\int_0^T\cos\beta\,dt.$$

The Hamiltonian is

$$H = -a \cos \beta + p_1 x_2 + p_2(a \sin \beta - g).$$

The first-order necessary conditions are

$$\dot{p}_1 = 0, \qquad \dot{p}_2 = -p_1,$$

$$\sin \beta + p_2 \cos \beta = 0.$$

This gives a control strategy of the form

$$\tan \beta = At + B,$$

where A and B are determined by the boundary conditions. Along the optimal trajectory, we have

$$H_{xx} = 0, \qquad H_{x\beta} = 0,$$

$$H_{\beta\beta} = a \cos \beta - ap_2 \sin \beta = a/\cos \beta > 0,$$

$$f_x = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad f_u = \begin{bmatrix} 0 \\ a \cos \beta \end{bmatrix}.$$

The Riccati equation is

$$-\dot{S} = f_x^T S + S f_x - S[f_\beta H_{\beta\beta}^{-I} f_\beta^T + (1/c_1)I]S, \qquad S(T) = c_2 I.$$

For $c_2 = 0$, the solution is $S(t) = 0$ all t. This means that any $c_1 > 0$ and $c_2 \geq 0$ will be sufficient for J to have a local minimum at the solution to the problem.

5. Iterative Algorithm

The results of the previous two sections are only useful if p and b have the correct values $\bar{p}$ and $\bar{b}$. Therefore, iterative methods of updating p and b in such a way that they converge to $\bar{p}$ and $\bar{b}$ must be studied. A natural way of updating p was suggested by Hestenes (Ref. 2) and used by di Pillo *et al.* (Ref. 4). The updating rule is

$$p(t)^{(i+1)} = p(t)^{(i)} + c_1[f(x(t)^{(i)}, u(t)^{(i)}, t) - \dot{x}(t)^{(i)}],$$

where $x^{(i)}$ and $u^{(i)}$ are the values that minimize the functional $J(x, u, p^{(i)}, b^{(i)}, c_1, c_2)$. The multiplier b is updated using a similar rule by Nahra (Ref. 6) and O'Doherty and Pierson (Ref. 8)

$$b^{(i+1)} = b^{(i)} + c_2 \psi(x^{(i)}(T)).$$

The convergence properties of these updating methods will now be investigated. First, consider the minimization of J for fixed p and b, where

$$\begin{aligned} J(x, u, p, b, c_1, c_2) = \int_0^T & \{L(x, u, t) + p^T[f(x, u, t) - \dot{x}] \\ & + (c_1/2)[f(x, u, t) - \dot{x}]^T[f(x, u, t) - \dot{x}]\}\, dt \\ & + F(x(T)) + b^T\psi(x(T)) + \tfrac{1}{2}c_2\psi(x(T))^T\psi(x(T)). \end{aligned} \tag{13}$$

This problem is of a standard form studied in the calculus of variations. Therefore, the minimum satisfies the Euler equations (see Gelfand and Fomin, Ref. 11)

$$\begin{aligned} -d(p + c_1(f - \dot{x}))/dt &= L_x^T + f_x^T p + c_1 f_x^T(f - \dot{x}), \\ [p + c_1(f - \dot{x})]_{t=T} &= F_x^T + \psi_x^T b + c_2\psi_x^T\psi, \\ L_u^T + f_u^T p + c_1 f_u^T(f - \dot{x}) &= 0. \end{aligned} \tag{14}$$

Introducing the definitions

$$\begin{gathered} p + c_1(f - \dot{x}) = \xi, \qquad b + c_2\psi = \zeta, \\ H(x, u, p, t) = L(x, u, t) + p^T f(x, u, t), \\ \varphi(x, b) = F(x) + b^T\psi(x), \end{gathered}$$

these equations can be written as

$$\begin{gathered} \dot{x} = f(x, u, t) + (1/c_1)(p - \xi), \\ -\dot{\xi} = H_x^T(x, u, \xi, t), \\ H_u(x, u, \xi, t) = 0, \\ x(0) = a, \qquad \psi(x(T)) = (1/c_2)(\zeta - b), \\ \xi(T) = \varphi_x^T(x(T), \zeta). \end{gathered}$$

Let h, k, η, θ, q, d denote the deviations from the optimum, i.e.,

$$\begin{gathered} h = x - \bar{x}, \qquad k = u - \bar{u}, \qquad \eta = \xi - \bar{p}, \\ \theta = \zeta - \bar{b}, \qquad q = p - \bar{p}, \qquad d = b - \bar{b}. \end{gathered}$$

Then, the equations are

$$\begin{gathered} \dot{h} = f(\bar{x} + h, \bar{u} + k, t) - f(\bar{x}, \bar{u}, t) + (1/c_1)(q - \eta), \\ -\dot{\eta} = H_x^T(\bar{x} + h, \bar{u} + k, \bar{p} + \eta, t) - H_x^T(\bar{x}, \bar{u}, \bar{p}, t), \\ H_u(\bar{x} + h, \bar{u} + k, \bar{p} + \eta, t) = 0, \\ h(0) = 0, \qquad \psi(\bar{x}(T) + h(T)) = (1/c_2)(\theta - d), \\ \eta(T) = \varphi_x^T(\bar{x}(T) + h(T), \bar{b} + \theta) - \varphi_x^T(\bar{x}(T), \bar{b}). \end{gathered} \tag{15}$$

The linearized version of these equations is

$$\begin{gathered}\dot{h} = f_x h + f_u k + (1/c_1)(q - \eta),\\ -\dot{\eta} = H_{xx} h + H_{xu} k + f_x^T \eta,\\ H_{uu} k + H_{ux} h + f_u^T \eta = 0,\\ h(0) = 0, \qquad \psi_x h(T) = (1/c_2)(\theta - d),\\ \eta(T) = \varphi_{xx} h(T) + \psi_x^T \theta,\end{gathered} \tag{16}$$

where H_{xx}, H_{xu}, etc., are evaluated along $(\bar{x}, \bar{u}, \bar{p})$.

If $H_{uu} > 0$, k can be expressed as

$$k = -H_{uu}^{-1} H_{ux} h - H_{uu}^{-1} f_u^T \eta.$$

This gives the following two-point boundary-value problem:

$$\begin{gathered}\dot{h} = (f_x - f_u H_{uu}^{-1} H_{ux}) h - [f_u H_{uu}^{-1} f_u^T + (1/c_1) I]\eta + (1/c_1) q,\\ -\dot{\eta} = (H_{xx} - H_{xu} H_{uu}^{-1} H_{ux}) h - (H_{xu} H_{uu}^{-1} f_u^T - f_x^T)\eta,\\ h(0) = 0, \qquad \psi_x h(T) = (1/c_2)(\theta - d),\\ \eta(T) = \varphi_{xx} h(T) + \psi_x^T \theta.\end{gathered} \tag{17}$$

Let

$$\Phi(t, s) = \begin{bmatrix} \Phi_{11}(t, s) & \Phi_{12}(t, s) \\ \Phi_{21}(t, s) & \Phi_{22}(t, s) \end{bmatrix}$$

be the fundamental matric of this system of linear differential equations, and let S be the solution of the associated Riccati equation

$$\begin{aligned}-\dot{S} = {} & H_{xx} - H_{xu} H_{uu}^{-1} H_{ux} + (f_x - f_u H_{uu}^{-1} H_{ux})^T S\\ & + S(f_x - f_u H_{uu}^{-1} H_{ux}) - S[f_u H_{uu}^{-1} f_u^T + (1/c_1) I] S,\\ S(T) = {} & F_{xx} + \Sigma \bar{b}_i (\psi_i)_{xx} + c_2 \psi_x^T \psi_x.\end{aligned} \tag{18}$$

Note that this Riccati equation is identical to (10). Assume that there exist c_1^0 and c_2^0 such that (18) has a solution on $[0, T]$ for $c_1 \geq c_1^0$, $c_2 \geq c_2^0$. In what follows, only values of c_1 and c_2 satisfying $c_1 \geq c_1^0$, $c_2 \geq c_2^0$ will be studied.

The two-point boundary-value problem (15) can be represented as an integral equation, using the technique of Falb and de Jong (Ref. 13). A short description is given in Appendix B. It is convenient to regard θ as a function on $[0, T]$ satisfying the differential equation $\dot{\theta} = 0$. The boundary conditions of the linearized problem (17) can then be written as

$$\begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} h(0) \\ \eta(0) \\ \theta(0) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ \phi_{xx} & -I & \psi_x^T \\ \psi_x & 0 & -(1/c_2) I \end{bmatrix} \begin{bmatrix} h(T) \\ \eta(T) \\ \theta(T) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -(1/c_2) d \end{bmatrix}.$$

The linearized problem is boundary compatible (see Definition 8.1 in Appendix B) if the following matrix is nonsingular

$$A=\begin{bmatrix} I & 0 & 0 \\ \varphi_{xx}\Phi_{11}-\Phi_{21} & \varphi_{xx}\Phi_{12}-\Phi_{22} & \psi_x^T \\ \psi_x\Phi_{11} & \psi_x\Phi_{12} & -(1/c_2)I \end{bmatrix}, \tag{19}$$

where

$$\Phi_{ij}=\Phi_{ij}(T,0).$$

A is nonsingular if

$$[\Phi_{22}(T,0)-(\varphi_{xx}+c_2\psi_x^T\psi_x)\Phi_{12}(T,0)]$$

is nonsingular. Since this matrix is related to the solution of the Riccati equation (18) by

$$\begin{aligned} S(t)=&[\Phi_{22}(T,t)-(\varphi_{xx}+c_2\psi_x^T\psi_x)\Phi_{12}(T,t)]^{-1}\\ &\cdot[(\varphi_{xx}+c_2\psi_x^T\psi_x)\Phi_{11}(T,t)-\Phi_{21}(T,t)], \end{aligned}$$

the nonsingularity follows from the assumption that $S(t)$ exists on $[0, T]$.

Note that the equation

$$H_u(\bar{x}+h,\bar{u}+k,\bar{p}+\eta,t)=0$$

defines k uniquely in terms of h and η if h and η are sufficiently small. This follows from the implicit function theorem (Ref. 9), since $H_{uu}(\bar{x},\bar{u},\bar{p},t)>0$. The solution of (15) can now be written as

$$\begin{aligned} \begin{bmatrix} h\\ \eta\\ \theta \end{bmatrix} =& K(t)\begin{bmatrix} 0\\ \varphi_x^T(\bar{x}(T)+h(T),\bar{b}+\theta)-\varphi_x^T(\bar{x}(T),\bar{b})-\varphi_{xx}h(T)-\psi_x^T\theta\\ \psi(\bar{x}(T)+h(T))-\psi_xh(T)-(1/c_2)d \end{bmatrix}\\ &+\int_0^T G(t,s)\begin{bmatrix} f(\bar{x}+h,\bar{u}+k,s)-f(\bar{x},\bar{u},s)-f_xh-f_uk+(1/c_1)q\\ H_x^T(\bar{x},\bar{u},\bar{p},s)-H_x^T(\bar{x}+h,\bar{u}+k,\bar{p}+\eta,s)+H_{xx}h+H_{xu}k+f_x^T\eta\\ 0 \end{bmatrix}ds, \end{aligned} \tag{20}$$

with k given by

$$H_u^T(\bar{x}+h,\bar{u}+k,\bar{p}+\eta,t)=0.$$

$K(t)$ and $G(t,s)$ are the Green's matrices associated with the linear two-point boundary-value problem (see Lemma 8.1 in Appendix B).

Theorem 5.1. There exist constants $\varepsilon>0$ and $\delta>0$ such that, for all continuous functions q and all d with $\|q\|_0\leq\delta$ and $\|d\|\leq\delta$, there exists a unique solution (h, η) to the integral equation (20) satisfying

$$\|h\|_0+\|\eta\|_0+\|\theta\|_0\leq \varepsilon.$$

Proof. The integral equation can be written as an operator equation

$$h = T_1(h, \eta, \theta)+A_1 \cdot q,$$
$$\eta = T_2(h, \eta, \theta),$$
$$\theta = T_3(h, \eta, \theta)+A_2 \cdot d,$$

where T_i are maps from $C_0^{3n}[0, T]$ to $C_0^n[0, T]$ and where A_1 and A_2 are linear maps. Let α be a real number, $0\leq\alpha<1$. Then, from Eq. (20), it follows that there exists an $\varepsilon>0$ and a $\delta>0$ such that

$$\begin{aligned}&\|T_i(h_1, \eta_1, \theta_1)-T_i(h_2, \eta_2, \theta_2)\|_0\\ &\quad\leq\alpha[\|h_1-h_2\|_0+\|\eta_1-\eta_2\|_0+\|\theta_1-\theta_2\|_0], \qquad i=1, 2, 3,\end{aligned}$$

for all $h_1, h_2, \eta_1, \eta_2, \theta_1, \theta_2$ satisfying

$$\|h_i\|_0+\|\eta_i\|_0+\|\theta_i\|_0\leq \varepsilon, \qquad i=1, 2.$$

It also follows that

$$\|T_1(0, 0, 0)+A_1 \cdot q\|\leq\|A_1\| \cdot \|q\|,$$
$$\|T_3(0, 0, 0)+A_2 \cdot d\|\leq\|A_2\| \cdot \|d\|.$$

Define

$$\eta=\max\,[\|A_1\| \cdot \|q\|, \|A_2\| \cdot \|d\|].$$

Choose δ such that

$$\eta/(1-\alpha)\leq\varepsilon$$

for q and d satisfying

$$\|q\|_0\leq\delta, \qquad \|d\|\leq\delta.$$

The conditions of the contraction mapping theorem (see, e.g., Ref. 13) are then satisfied, and the theorem is proved. □

To study the solution h, k, η for small values of q and b, it is desirable to have an approximate representation.

Theorem 5.2. Let (h, η) be the solution of the nonlinear problem (15). Then,

$$\begin{bmatrix}h\\ \eta\\ \theta\end{bmatrix}=K(t)\begin{bmatrix}0\\ 0\\ -(1/c_2)d\end{bmatrix}+\int_0^T (1/c_1)G(t, s)\begin{bmatrix}G\\ 0\\ 0\end{bmatrix}ds+r(q, d),$$

where

$$\|r(q, d)\|_0/(\|q\|_0+\|d\|)\to 0 \qquad \text{as } (q, d)\to 0.$$

Proof. From (20), it follows that

$$\|T_i(h, k, \eta)\|\le\tfrac{1}{2}K_1[\|h\|_0+\|\eta\|_0+\|\theta\|_0]^2;$$

consequently,

$$\|h\|_0+\|\eta\|_0+\|\theta\|_0\le K_1[\|h\|_0+\|\eta\|_0+\|\theta\|_0]^2+K_2(\|q\|_0+\|d\|),$$

for some constants K_1 and K_2. Let ε be the constant defined in Theorem 5.1, and let

$$\varepsilon_1=\min(1/2K_1, \varepsilon).$$

Then, for sufficiently small $\|q\|$ and $\|d\|$,

$$\eta/(1-\alpha)\le\varepsilon_1,$$

where α and η are defined as in Theorem 5.1. Consequently,

$$\|h\|_0+\|\eta\|_0+\|\theta\|_0\le 1/2K_1$$

for sufficiently small $\|q\|_0$ and $\|d\|$. This gives

$$\|h\|_0+\|\eta\|_0+\|\theta\|_0\le 2K_2[\|q\|_0+\|d\|].$$

Using this in the expression for $\|T_i(h, \eta, \theta)\|$ gives the desired bound on $r(q, d)$. □

Corollary 5.1. Let $\tilde{h}$, $\tilde{\eta}$, $\tilde{\theta}$ denote the solution to the linearized boundary-value problem (17). Then, the solutions of the nonlinear problem and the linearized problem are related by

$$\begin{bmatrix} h \\ \eta \\ \theta \end{bmatrix}=\begin{bmatrix} \tilde{h} \\ \tilde{\eta} \\ \tilde{\theta} \end{bmatrix}+r(q, d),$$

where

$$\|r(q, d)\|_0/(\|q\|_0+\|d\|)\to 0 \qquad \text{as } (q, d)\to 0.$$

Proof. The solution to (17) is given by

$$\begin{bmatrix} \tilde{h} \\ \tilde{\eta} \\ \tilde{\theta} \end{bmatrix}=K(t)\begin{bmatrix} 0 \\ 0 \\ -(1/c_2)d \end{bmatrix}+\int_0^T G(t, s)\begin{bmatrix} (1/c_1)q \\ 0 \\ 0 \end{bmatrix}ds. \quad □ \qquad (21)$$

With this result, it is possible to investigate the convergence rate of the iterative method of updating the multipliers. As mentioned at the beginning of this section, the algorithm will be assumed to be the following.

Algorithm 5.1

(i) Choose starting values $p^{(0)}$, $b^{(0)}$; put $i=0$.

(ii) Minimize $J(x, u, p^{(i)}, b^{(i)}, c_1, c_2)$; let the result be $x^{(i)}$, $u^{(i)}$.

(iii) Update the multipliers

$$p^{(i+1)}(t)=p^{(i)}(t)+c_1[f(x^{(i)}, u^{(i)}, t)-\dot{x}^{(i)}(t)]$$
$$b^{(i+1)}=b^{(i)}+c_2\psi(x^{(i)}(t));$$

put $i=i+1$ and go to (ii).

It is assumed that c_1 and c_2 are held constant, and that

$$c_1\geq c_1^0, \qquad c_2\geq c_2^0.$$

Theorem 5.3. Let $p^{(i)}$ and $b^{(i)}$ be generated by Algorithm 5.1. Assume the following.

(i) $\bar{p}$ and $\bar{b}$ satisfy Eq. (3).

(ii) The linearized system

$$\dot{x}=f_x(\bar{x}, \bar{u}, t)h+f_u(\bar{x}, \bar{u}, t)k$$

is controllable.

(iii) $H_{uu}(\bar{x}, \bar{u}, \bar{p}, t)>0, \qquad t\in[0, T]$.

(iv) The Riccati equation (18) has a solution on $[0, T]$ for

$$c_1=c_1^0, \qquad c_2=c_2^0.$$

Then, there are constants

$$c_1\geq c_1^0, \qquad c_2\geq c_2^0$$

such that, if $p^{(0)}$ and $b^{(0)}$ are sufficiently close to $\bar{p}$ and $\bar{b}$, then

$$\|p^{(i+1)}-\bar{p}\|_0+\|b^{(i+1)}-\bar{b}\|\leq K[\|p^{(i)}-\bar{p}\|_0+\|b^{(i)}-\bar{b}\|],$$

where K is an arbitrary number in $(0, 1)$.

Proof. Using the notation

$$p^{(i)}-\bar{p}=q^{(i)}, \qquad x^{(i)}-\bar{x}=h^{(i)},$$
$$u^{(i)}-\bar{u}=k^{(i)}, \qquad b^{(i)}-\bar{b}=d^{(i)},$$

the updating formula can be written as

$$q^{(i+1)}=q^{(i)}+c_1[f_xh^{(i)}+f_uk^{(i)}-\dot{h}^{(i)}]+c_1R_1(h^{(i)}, k^{(i)}),$$
$$d^{(i+1)}=d^{(i)}+c_2\psi_xh^{(i)}(T)+c_2R_2(h^{(i)}(T)),$$

where

$$\|R_1(h, k)\|_0/(\|h\|_0+\|k\|_0)\to 0 \qquad \text{as } (h, k)\to 0,$$
$$\|R_2(z)\|/\|z\|\to 0 \qquad \text{as} \quad z\to 0.$$

If $\tilde{h}$, $\tilde{\eta}$, $\tilde{\theta}$, $\tilde{k}$ denote the solution to the linear, two-point boundary-value problem, then it follows from Theorem 5.2 that

$$\begin{aligned} q^{(i+1)} &= q^{(i)} + c_1(f_x\tilde{h}^{(i)} + f_u\tilde{k}^{(i)} - \dot{\tilde{h}}^{(i)}) + R_3(q^{(i)}, d^{(i)}), \\ d^{(i+1)} &= d^{(i)} + c_2\psi_x\tilde{h}^{(i)}(T) + R_4(q^{(i)}, d^{(i)}), \end{aligned} \tag{22}$$

where

$$\|R_i(q, d)\|_0/(\|q\|_0+\|d\|)\to 0 \qquad \text{as } (q, d)\to 0.$$

From (16), it follows that

$$\begin{aligned} f_x\tilde{h}^{(i)} + f_u\tilde{k}^{(i)} - \dot{\tilde{h}}^{(i)} &= (1/c_1)(\eta^{(i)} - q^{(i)}), \\ \psi_x\tilde{h}(T) &= (1/c_2)(\theta^{(i)} - d^{(i)}). \end{aligned}$$

Using these expressions in (22) results in

$$\begin{aligned} q^{(i+1)} &= \eta^{(i)} + R_3(q^{(i)}, d^{(i)}), \\ d^{(i+1)} &= \theta^{(i)} + R_4(q^{(i)}, d^{(i)}). \end{aligned}$$

From conditions (i)–(iv), it follows that the linear problem (17) has a solution for $c_1=\infty$, $c_2=\infty$. Then, $K(t)$ and $G(t, s)$ go to finite limits as $c_1\to\infty$, $c_2\to\infty$. From (21), it then follows that there are values $c_1\geq c_1^0$ and $c_2\geq c_2^0$ such that

$$\|\eta^{(i)}\|_0+\|\theta^{(i)}\|_0\leq (K/2)[\|q^{(i)}\|_0+\|d^{(i)}\|].$$

For these values of c_1 and c_2, choose δ such that

$$\|R_i(q^{(i)}, d^{(i)})\|_0\leq (K/4)[\|q^{(i)}\|_0+\|d^{(i)}\|]$$

for

$$\|q^{(i)}\|_0+\|d^{(i)}\|\leq \delta.$$

Then,

$$\|q^{(i+1)}\|_0+\|d^{(i+1)}\|\leq K[\|q^{(i)}\|_0+\|d^{(i)}\|]$$

for

$$\|q^{(0)}\|_0+\|d^{(0)}\|\leq \delta. \qquad \square$$

Theorem 5.3 shows that Algorithm 5.1 can be used to solve the optimal control problem. However, this algorithm is based on the minimization of J

for fixed values of the multipliers. This is not a trivial problem, even if it is simpler than the original optimization problem, because the differential equation and terminal constraints are eliminated. Di Pillo *et al.* (Ref. 4) have studied this problem and shown that a conjugate gradient method can be used. The optimization problem can then be solved using only quadratures and without the solution of any differential equations.

6. Conclusions

The results of Sections 3 and 4 shed some light on how the constant c affects the existence of a minimum for J. In particular, it is interesting to note that, in some cases, the result can be seen at a glance, without any computations, as shown in Example 4.2. It is also worth noting that, for problems where

$$H_{xx}-H_{xu}H_{uu}^{-1}H_{ux}>0, \qquad 0\leq t\leq T,$$
$$F_{xx}>0,$$

with the above expressions evaluated along $\bar{x}$ and $\bar{u}$, the solution to (7) can be extended over the whole interval. In this case, any c greater than zero is sufficient.

The results in Section 4 are analogous to results known from the finite-dimensional case. Theorem 5.3 suggests that high values of c are always good, since they give a high linear convergence rate. In practice, it is also necessary to consider the fact that a high value of c gives an ill-conditioned problem when minimizing J.

7. Appendix A: Properties of the Riccati Equation

Here, some basic properties of the Riccati equation that are needed in the proofs on the preceding pages are collected. Most of them can be found in Refs. 7, 14, 15, but not necessarily in the form given here.

We will write the Riccati equation in the form

$$-S(t)=A^T(t)S(t)+S(t)A(t)+Q(t)-S(t)P(t)S(t), \qquad S(T)=Q_0,$$

where A, Q, P are matrices whose elements are continuous functions of t and Q_0, Q, P are symmetric. It follows from standard theorems for differential equations that $S(t)$ exists at least on a sufficiently small interval $t_0\leq t\leq T$. Moreover, the only way in which S can fail to exist is by having some element which becomes unbounded. In what follows, $M\geq N$, where M and N are symmetric matrices, means that $M-N$ is nonnegative definite and $M>N$ means that $M-N$ is positive definite.

It is useful to rewrite the Riccati equation as an integral equation. Introduce the fundamtneal matrix $\phi(t, T)$ satisfying

$$(d/dt)\phi(t, T)=[A(t)-\tfrac{1}{2}P(t)S(t)]\phi(t, T), \qquad \phi(T, T)=I.$$

Then, we have

$$S(t)=\int_t^T \phi^T(s, t)Q(s)\phi(s, t)\, ds+\phi^T(T, t)Q_0\phi(T, t).$$

Lemma 7.1. For the Riccati equation

$$-\dot{S}=A^TS+SA+Q-SPS,$$

let S_1 and S_2 be the solutions corresponding to

$$S(T)=Q_0^1 \qquad \text{and} \qquad S(T)=Q_0^2,$$

respectively. Then, if $Q_0^2\geq Q_0^1$, it follows that $S_2(t)\geq S_1(t)$ for all $t\in[t_0, T]$, where $[t_0, T]$ is an interval on which both solutions exist.

Proof. We have

$$\begin{aligned}-(d/dt)(S_2-S_1)&=(A-PS_1)^T(S_2-S_1)+(S_2-S_1)(A-PS_1)\\&\quad-(S_2-S_1)P(S_2-S_1),\\ S_2(T)-S_1(T)&=Q_0^2-Q_0^1.\end{aligned}$$

Regarding this as a Riccati equation in S_2-S_1 we get, using the integral equation representation above,

$$S_2-S_1=\phi^T(T, t)(Q_0^2-Q_0^1)\phi(T, t),$$

where $\phi(t, T)$ now is the fundamental matrix corresponding to

$$A-PS_1-\tfrac{1}{2}P(S_2-S_1).$$

Lemma 7.2. Let S_1 and S_2 be the solutions of the Riccati equations

$$-\dot{S}=A^TS+SA+Q-SP_1S, \qquad S(T)=Q_0,$$
$$-\dot{S}=A^TS+SA+Q-SP_2S, \qquad S(T)=Q_0,$$

respectively. If $P_1\geq P_2$, then

$$S_2(t)\geq S_1(t), \qquad t\in[t_0, T],$$

where $[t_0, T]$ is any interval on which both solutions exist.

Proof. We have

$$-(d/dt)(S_2-S_1)=(A-P_2S_1)^T(S_2-S_1)+(S_2-S_1)(A-P_2S_1)$$
$$-(S_2-S_1)P_2(S_2-S_1)+S_1(P_1-P_2)S_1,$$
$$S_2(T)-S_1(T)=0.$$

Using the integral equation form, this can be written as

$$S_2(t)-S_1(t)=\int_t^T \phi^T(s,t)S_1(P_1-P_2)S_1\phi(s,t)\,ds.$$

Lemma 7.3. Let S_1 and S_2 be the solutions of the Riccati equations

$$-\dot{S}=A^TS+SA+Q_1-SPS, \qquad S(T)=Q_0,$$
$$-\dot{S}=A^TS+SA+Q_2-SPS, \qquad S(T)=Q_0,$$

respectively. Then, if $Q_2\geq Q_1$, it follows that

$$S_2(t)\geq S_1(t), \qquad t\in[t_0,T],$$

where $[t_0, T]$ is any interval on which both solutions exist.

Proof. We have

$$S_2(t)-S_1(t)=\int_t^T \phi^T(s,t)(Q_2-Q_1)\phi(s,t)\,ds,$$

where ϕ is the fundamental matrix corresponding to

$$A-PS_1-\tfrac{1}{2}P(S_2-S_1).$$

We can now deduce the following result.

Lemma 7.4. If $P>0$, then there exists a continuous matrix $R(t)$ such that $S(t)\leq R(t)$ on any interval $[t_0, T]$ where S exists.

Proof. From Lemma 7.2, it follows that $S(t)\leq R(t)$, where R is the solution to the linear differential equation

$$-\dot{R}=A^TR+RA+Q, \qquad R(T)=Q_0.$$

From this lemma, it follows that, to prove existence of $S(t)$ on some interval, all that is needed is a lower bound on S on that interval.

Lemma 7.5. Let S be the solution of the Riccati equation

$$-\dot{S}=A^TS+SA+Q+SPS, \qquad S(T)=Q_0,$$

and assume that S exists on the interval $[t_0, T]$. Let $\tilde{S}$ be the solution to the Riccati equation where $\tilde{A}, \tilde{Q}, \tilde{P}$ have replaced A, Q, P. Then, there exists an $\varepsilon > 0$ such that $\tilde{S}$ also exists on $[t_0, T]$ if

$$\|\tilde{A}-A\| \leq \varepsilon, \qquad \|\tilde{Q}-Q\| \leq \varepsilon, \qquad \|\tilde{P}-P\| \leq \varepsilon.$$

Proof. Since the right-hand side of the Riccati equation is a continuous function of S, A, Q, P, the result follows from general results for nonlinear differential equations (see Ref. 12).

8. Appendix B: Two-Point Boundary-Value Problem

A linear two-point boundary-value problem can be written as

$$\dot{y} = V(t)y + f(t), \qquad My(0) + Ny(1) = c,$$

where V, M, N are $p \times p$ matrices and f, c, y are p-vectors.

Definition 8.1. (*See Ref. 13*). The set $\{V, M, N\}$ is called boundary compatible if (i) $V(t)$ is measurable with $\|V(t)\| < m(t)$ for an integrable $m(t)$, and (ii) $\det[M + N\phi(1, 0)] \neq 0$, where $\Phi(t, s)$ is the fundamental matrix of $y = V(t)y$.

$\{V, M, N\}$ is a boundary compatible set iff the linear two-point boundary-value problem has a solution for all f and c.

Lemma 8.1. Let D be an open set in R^p, and let I be an open set in R containing $[0, 1]$. Assume the following: (i) $F(y, t)$ is a map of $D \times I$ into D which is measurable in t for each fixed y and continuous in y for each fixed t; (ii) there is an integrable function $m(t)$ such that $\|F(y, t)\| < m(t)$ on $D \times I$; (iii) $g(y)$ and $h(y)$ are maps of D into D; and (iv) $\{V(t), M, N\}$ is a boundary compatible set. Then, the boundary-value problem

$$\dot{y} = F(y, t), \qquad g(y(0)) + h(y(1)) = c$$

has the equivalent representation

$$y(t) = H(t)\{c - g(y(0)) - h(y(1)) + My(0) + Ny(1)\}$$
$$+ \int_0^1 G(t, s)\{F(y(s), s) - V(s)y(s)\}\, ds,$$

where the Green's functions $H(t)$ and $G(t, s)$ are given by

$$H(t) = \Phi(t, 0)(M + N\Phi(1, 0))^{-1},$$

$$G(t, s) = \begin{cases} \Phi(t, 0)(M + N\Phi(1, 0))^{-1}M\Phi(0, s), & 0 < s < t, \\ -\Phi(t, 0)(M + N\Phi(1, 0))^{-1}N\Phi(1, s), & t < s < 1, \end{cases}$$

where $\Phi(t, s)$ is the fundamental matrix of the linear system $y = V(t)y$.

Proof. See Falb and de Jong (Ref. 13).

References

1. Hestenes, M. R., *An Indirect Sufficiency Proof for the Problem of Bolza in Nonparametric Form*, Transactions of the American Mathematical Society, Vol. 62, pp. 509–535, 1947.
2. Hestenes, M. R., *Multiplier and Gradient Methods*, Journal of Optimization Theory and Applications, Vol. 4, pp. 303–320, 1969.
3. Rupp, R. D., *A Method for Solving a Quadratic Optimal Control Problem*, Journal of Optimization Theory and Applications, Vol. 9, pp. 238–250, 1972.
4. Di Pillo, G., Grippo, L., and Lampariello, F., *The Multiplier Method for Optimal Control Problems*, Ricerche di Automatica, Vol. 5, No. 2–3, pp. 133–157, 1974.
5. Rupp, R. D., *Approximation of the Classical Isoperimetric Problem*, Journal of Optimization Theory and Applications, Vol. 9, pp. 251–264, 1972.
6. Nahra, J. E., *Balance Function for the Optimal Control Problem*, Journal of Optimization Theory and Applications, Vol. 8, pp. 35–48, 1971.
7. Mårtensson, K., *New Approaches to the Numerical Solution of Optimal Control Problems*, Lund University, Department of Automatic Control, Report No. 7206, 1972.
8. O'Doherty, R. J., and Pierson, B. L., *A Numerical Study of Augmented Penalty function Algorithms for Terminally Constrained Optimal Control Problems*, Journal of Optimization Theory and Applications, Vol. 14, pp. 393–403, 1974.
9. Luenberger, D., *Optimization by Vector Space Methods*, John Wiley and Sons, New York, New York, 1968.
10. Bryson, A. E., and Ho, Y. C., *Applied Optimal Control*, Ginn and Company, Waltham, Massachusetts, 1969.
11. Gelfand, I. M., and Fomin, S. V., *Calculus of Variations*, Prentice-Hall, London, England, 1963.
12. Coddington, E. A., and Levinson, N., *Theory of Ordinary Differential Equations*, McGraw-Hill Book Company, New York, New York, 1955.
13. Falb, P. L. and De Jong, J. L., *Some Successive Approximation Methods in Control and Oscillation Theory*, Academic Press, New York, New York, 1969.
14. Brockett, R. W., *Finite Dimensional Linear Systems*, John Wiley and Sons, New York, New York, 1970.
15. Anderson, B. D. O., and Moore, J. B., *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.