JOURNAL OF GUIDANCE, CONTROL, AND DYNAMICS
Vol. 18, No. 5, September–October 1995

Finite Element Method for the Solution of State-Constrained Optimal Control Problems

Robert R. Bless*
Lockheed Engineering and Sciences Company, Inc., Hampton, Virginia 23666
Dewey H. Hodges†
Georgia Institute of Technology, Atlanta, Georgia 30332-0150
and
Hans Seywald‡
Analytical Mechanics Associates, Inc., Hampton, Virginia 23666

This paper presents an extension of a finite element formulation based on a weak form of the necessary conditions to solve optimal control problems. First, a general formulation for handling internal boundary conditions and discontinuities in the state equations is presented. Then, the general formulation is modified for optimal control problems subject to state-variable inequality constraints. Solutions with touch points and solutions with state-constrained arcs are considered. After the formulations are developed, suitable shape and test functions are chosen for a finite element discretization. It is shown that all element quadrature (equivalent to one-point Gaussian quadrature over each element) may be done in closed form, yielding a set of algebraic equations. To demonstrate and analyze the accuracy of the finite element method, a simple state-constrained problem is solved. Then, for a more practical application of the use of this method, a launch vehicle ascent problem subject to a dynamic pressure constraint is solved. The paper also demonstrates that the finite element results can be used to determine switching structures and initial guesses for a shooting code.

Introduction

APPROACHES to the numerical solution of optimal control problems may be classified in two categories: direct methods and indirect methods. Among the direct methods are those that transcribe the (infinite-dimensional) continuous problem to a finite-dimensional nonlinear programming problem (NLP) by some parameterization of the control histories and (possibly) the state histories. Continuing advances in NLP algorithms and related software have made these the methods of choice in many applications.[1,2]

Indirect methods are based on finding the solution of a boundary-value problem that results from the first-order necessary conditions of optimal control.[3,4] For many practical optimization problems the boundary-value problems are quite difficult, but here, again, modern numerical algorithms and associated software have enlarged the class of solvable problems significantly.[5–7] The virtue of the indirect methods is the high precision they offer and their rapid convergence in the immediate neighborhood of the optimal solution. In addition, optimal trajectories are often composed of sequences of arcs along which various state or control constraints are alternately active and inactive (referred to as the switching structure).[8,9] NLP-based approaches can have difficulties in correctly identifying some switching structures.

Shooting methods[10] are indirect methods that yield numerically exact solutions. Unfortunately, shooting methods are, in general, very sensitive, and it may be difficult to obtain a converged solution without good initial guesses for the costates. Therefore, another type of indirect method may be needed to 1) produce initial guesses for a shooting method and 2) identify the optimal switching structure. A finite difference method[10] is one candidate method. This paper presents another alternative, namely, a finite element method based on a weak formulation of the first-order necessary conditions of optimality.

Finite element methods, including some based on Rayleigh–Ritz and Galerkin methods, have been used to solve optimal control problems (see, for example, Refs. 11 and 12). All these methods suffer from some computational challenges. For example, the approximating functions must satisfy all strong boundary conditions, which means that one set of functions will not suffice for all types of problems. Also element quadrature must be done by numerical means, which greatly increases the computational effort. The method presented in this paper avoids these two problems by reformulating the variational problem such that all boundary conditions appear as natural boundary conditions. This way, the simplest possible approximating functions are allowed for all cases, and all element quadrature may be done by inspection. This paper presents an extension of the theory developed in Refs. 13 and 14 to include optimal control problems subject to state-variable inequality constraints.

A treatment of optimal control problems with internal boundary conditions was given in Ref. 14. The formulation was explicitly derived for a problem with only one internal boundary condition. The present paper develops the algebraic equations for an arbitrary number of internal boundary conditions, and these equations are more directly obtained than those in Ref. 14. Additionally, the treatment of state constraints will be handled. Finally, two examples are included to demonstrate some of the features of the finite element method, particularly the accuracy and use of the approximate answers in obtaining the switching structure and initial guesses for a shooting method.

Problem Definition

Consider a system defined over a time interval from t_0 (initial time) to t_f (final time) by a set of n states, x, and a set of m controls, u. The states of the system are governed by a set of first-order differential equations referred to as state equations. During the time interval from t_0 to t_f, there may be state constraints to satisfy (control constraints are discussed in Ref. 15), discontinuities in the states, or discontinuities in the state equations at interior points (times

Received April 15, 1994; revision received March 10, 1995; accepted for publication March 17, 1995.

*Senior Engineer, 144 Research Drive; currently working under contract at the Guidance and Controls Branch, NASA Langley Research Center, Hampton, VA 23681-0001. Member AIAA.

†Professor, School of Aerospace Engineering. Fellow AIAA.

‡Supervising Engineer, 17 Research Drive; currently working under contract at the Guidance and Controls Branch, NASA Langley Research Center, Hampton, VA 23681-0001. Member AIAA.

between t_0 and t_f). These interior points, along with the initial and final points, will be referred to as events (the initial time is event 0), and the time interval between events will be referred to as phases. The time of event i will be denoted by t_i and times just before and after t_i will be denoted by t_i^- and t_i^+, respectively. All event times will be considered unknown except t_0.

Elements of a performance index may be denoted with an integrand $L^{(i)}[x(t), u(t)]$ and a discrete function ϕ of the states and/or times at any of the events. A general class of such problems with N phases involves choosing $u(t)$ to minimize

$$J = \phi\left[x\left(t_0^+\right), x\left(t_1^-\right), x\left(t_1^+\right), \ldots, x\left(t_N^-\right); t_0, t_1, \ldots, t_N\right] + \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} L^{(i)}[x(t), u(t)]\, \mathrm{d}t \tag{1}$$

subject to the state equation constraints

$$\dot{x} = f^{(i)}[x(t), u(t)] \qquad t_{i-1}^+ < t < t_i^-, \qquad i = 1, \ldots, N \tag{2}$$

with q boundary conditions specified as

$$\psi\left[x\left(t_0^+\right), x\left(t_1^-\right), x\left(t_1^+\right), \ldots, x\left(t_N^-\right); t_0, t_1, \ldots, t_N\right] = 0 \tag{3}$$

Introducing Lagrange multiplier functions $\lambda(t)$, referred to as costates, and discrete Lagrange multipliers ν, for convenience, we define Φ and H

$$\Phi \equiv \phi + \nu^T \psi \qquad H^{(i)} \equiv L^{(i)} + \lambda^T f^{(i)} \qquad i = 1, \ldots, N \tag{4}$$

A general multiplier theorem states[16] that the optimal solution furnishes a stationary point to the augmented cost function

$$J_a = \Phi + \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} \left(H^{(i)} - \lambda^T \dot{x}\right) \mathrm{d}t \tag{5}$$

Consider now the case where there is a state constraint of the form $S(x) \leq 0$ imposed on the problem. To define the order of the constraint S, successive total time derivatives of S are taken and $f(x, u)$ is substituted for $\dot{x}$ until an expression explicitly dependent on the control u is obtained. If p total-time derivatives are required, then S is called a pth-order state-variable inequality constraint. With the imposed constraint, the optimal trajectory may 1) never violate the constraint, 2) touch the state constraint at one or more instances in time (touch points), 3) remain on the curve $S = 0$ for one or more finite amounts of time (a boundary arc), or 4) have some combination of boundary arcs and touch points. The sequence of constrained arcs, unconstrained arcs, and touch points will be referred to as the switching structure.

One problem facing the analyst trying to solve a state-constrained problem is finding the switching structure of the constrained arcs. Some general guidelines are given in Refs. 17 and 18. Assuming that the hodograph of the problem is strictly convex[19] (also referred to as a regular Hamiltonian), it has been shown that the control must be continuous across a junction point. In this case, for a scalar state constraint, Ref. 17 proves that only touch-point solutions are possible for odd-order constraints of order greater than 2. In Ref. 18, it is shown that first-order constraints have boundary arc solutions except under three uncommon conditions. Second-order constraints can have touch-point or boundary arc solutions depending on the constraint limit.

Touch-Point Cases

For the special case where the trajectory only touches the curve $S = 0$ at discrete instances of time, the problem defined in Eqs. (1–5) is of the proper form. For example, if there were only one touch point occurring in the solution (at some unknown time), then the conditions that $S(t_1) = 0$ and the continuity of all states at time t_1 [i.e., $x(t_1^-) - x(t_1^+) = 0$] would be added to the vector ψ of Eq. (3). Note also that the problem would now have one additional phase due to the extra event at t_1.

Boundary Arc Case

Consider the situation where the solution has one unconstrained arc between t_0 and t_1 followed by a constrained arc between t_1 and t_2 and then another unconstrained arc between t_2 and t_f (or t_3). The augmented cost function defined in Eq. (5) is modified in the second phase to include the state constraint. As is done in Refs. 20 and 21, successive time derivatives of the state constraint are taken until the control appears explicitly, say in the pth derivative. Then, to enforce the condition $\mathrm{d}^p S/\mathrm{d}t^p = 0$ for time between t_1 and t_2, the term

$$\eta \frac{\mathrm{d}^p S}{\mathrm{d}t^p} \tag{6}$$

is added to the integrand of Eq. (1) and the augmented cost function of Eq. (5) takes the form

$$J_a = \Phi + \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} \left(H^{(i)} - \lambda^T \dot{x} + \eta \frac{\mathrm{d}^p S}{\mathrm{d}t^p}\right) \mathrm{d}t \tag{7}$$

Here, η is a new multiplier function of time that is nonnegative for t between t_1 and t_2 and is identically zero for t between t_0 and t_1 and t between t_2 and t_f. The continuity of the states at t_1 and t_2, namely,

$$x\left(t_1^-\right) - x\left(t_1^+\right) = 0 \qquad x\left(t_2^-\right) - x\left(t_2^+\right) = 0 \tag{8}$$

and the tangency conditions on the state constraint S,

$$T\left[x\left(t_1^+\right)\right] = \left[S \quad \frac{\mathrm{d}S}{\mathrm{d}t} \quad \cdots \quad \frac{\mathrm{d}^{p-1}S}{\mathrm{d}t^{p-1}}\right]^T = 0 \tag{9}$$

are enforced by adding Eqs. (8) and (9) to the constraint vector ψ of Eq. (3).

Derivation of the Weak Formulation

In this section, a variational formulation for solving optimal control problems subject to state-variable inequality constraints is developed. It will be shown that the formulation has no strong boundary conditions (i.e., boundary conditions requiring the virtual, or variational, quantities to be zero) but only natural, or weak, boundary conditions (i.e., those determined by setting the coefficient of a virtual quantity to zero). This allows the test functions (which approximate the virtual quantities) and shape functions (which approximate the dependent variables) to be chosen from a less restrictive class of functions; indeed, the same test and shape functions may be chosen for every optimal control problem. Thus, a general set of algebraic equations will be developed that will not have to be altered (in terms of numbers of equations) depending on the type of boundary conditions. Hence, this formulation is also referred to as a weak formulation.

The derivation of the weak formulation begins by expanding the augmented cost function of Eq. (5) into a Taylor series about the optimal solution. When the higher order terms are neglected and an integration by parts is done, this yields (see Ref. 20)

$$\begin{aligned}
\delta J_a = {} & \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} \left\{\delta\lambda^T\left(f^{(i)} - \dot{x}\right) + \delta x^T\left[\left(\frac{\partial H^{(i)}}{\partial x}\right)^T + \dot{\lambda}\right] + \delta u^T\left(\frac{\partial H^{(i)}}{\partial u}\right)^T\right\} \mathrm{d}t \\
& + \sum_{i=1}^{N}\left[\frac{\partial \Phi}{\partial x\left(t_i^-\right)} - \lambda^T\left(t_i^-\right)\right] \mathrm{d}x\left(t_i^-\right) \\
& + \sum_{i=0}^{N-1}\left[\frac{\partial \Phi}{\partial x\left(t_i^+\right)} + \lambda^T\left(t_i^+\right)\right] \mathrm{d}x\left(t_i^+\right) + \mathrm{d}\nu^T \psi \\
& + \sum_{i=1}^{N-1}\left[\frac{\partial \Phi}{\partial t_i} + H^{(i)}\left(t_i^-\right) - H^{(i+1)}\left(t_i^+\right)\right] \mathrm{d}t_i \\
& + \left[\frac{\partial \Phi}{\partial t_N} + H^{(N)}\left(t_N^-\right)\right] \mathrm{d}t_N
\end{aligned} \tag{10}$$

where $\delta x(t)$ signifies the perturbation of x when t is considered fixed and $\mathrm{d}x(t)$ signifies the perturbation of x due to changes in x and t.

The first-order necessary conditions of optimal control require that $\delta J_a = 0$ for all independently chosen perturbations $\mathrm{d}t_1, \ldots, \mathrm{d}t_N, \mathrm{d}\nu, \mathrm{d}x(t_0^+), \mathrm{d}x(t_1^-), \mathrm{d}x(t_1^+), \ldots, \mathrm{d}x(t_N^-), \delta x(t), \delta\lambda(t)$, and $\delta u(t)$. Hence, the discrete boundary condition terms are

$$\frac{\partial \Phi}{\partial t_i} + H^{(i)}\left(t_i^-\right) - H^{(i+1)}\left(t_i^+\right) = 0 \quad i = 1, 2, \ldots, N-1 \tag{11}$$

$$\frac{\partial \Phi}{\partial t_N} + H^{(N)}\left(t_N^-\right) = 0 \tag{12}$$

$$\psi\left[x\left(t_0^+\right), x\left(t_1^-\right), \ldots, x\left(t_N^-\right); t_0, \ldots, t_N\right] = 0 \tag{13}$$

$$\frac{\partial \Phi}{\partial x\left(t_i^-\right)} - \lambda^T\left(t_i^-\right) = 0 \qquad i = 1, 2, \ldots, N \tag{14}$$

$$\frac{\partial \Phi}{\partial x\left(t_i^+\right)} + \lambda^T\left(t_i^+\right) = 0 \qquad i = 0, 1, \ldots, N-1 \tag{15}$$

where Eq. (12), for instance, is obtained by setting all variations equal zero except for $\mathrm{d}t_N$. It is straightforward to put Eqs. (11–15) into discretized form.

Using the conditions in Eqs. (11–15) to simplify the expression $\delta J_a = 0$ of Eq. (10) and then integrating by parts yield

$$\begin{aligned} 0 = \sum_{i=1}^{N} \int_{t_{i-1}^+}^{t_i^-} \Bigg[& \delta\lambda^T f^{(i)} + \delta\dot{\lambda}^T x + \delta x^T \left(\frac{\partial H^{(i)}}{\partial x}\right)^T \\ & - \delta\dot{x}^T\lambda + \delta u^T \left(\frac{\partial H^{(i)}}{\partial u}\right)^T \Bigg] \mathrm{d}t \\ & + \sum_{i=1}^{N} \left[\delta x^T\left(t_i^-\right)\lambda\left(t_i^-\right) - \delta\lambda^T\left(t_i^-\right)x\left(t_i^-\right)\right] \\ & + \sum_{i=0}^{N-1} \left[\delta\lambda^T\left(t_i^+\right)x\left(t_i^+\right) - \delta x^T\left(t_i^+\right)\lambda\left(t_i^+\right)\right] \end{aligned} \tag{16}$$

Equation (16), when taken with Eqs. (11–15), is called the weak formulation of the first-order necessary conditions for the optimal control problem defined in Eqs. (1–3). Equations (11–16) will be used for the finite element discretization scheme described in the next section. Note that after the integration by parts no time derivatives of x or λ appear in Eq. (16).

If the optimal solution contains a touch point of a constraint, then the weak form defined in Eqs. (11–16) is still valid. As stated earlier, the effect of the touch point is reflected as a change in the number of phases and in the boundary conditions listed in ψ.

If the optimal solution contains a boundary arc of a constraint, then Eqs. (11–15) remain valid. However, $H^{(i)}$ is redefined to be

$$H^{(i)} = L^{(i)} + \lambda^T f^{(i)} + \eta \frac{\mathrm{d}^p S}{\mathrm{d}t^p} \tag{17}$$

for phase(s) i, which have an active state constraint. Additionally, in Eqs. (10) and (16), the term $\delta\eta(\mathrm{d}^p S/\mathrm{d}t^p)$ is added to the integrand.

Finite Element Discretization

Let the time interval from t_{i-1}^+ to t_i^- be broken into M_i elements, where $i = 1, 2, \ldots, N$. For convenience, define

$$\bar{M}_i = \sum_{j=1}^{i} M_j \qquad \text{for} \qquad i = 1, 2, \ldots, N \tag{18}$$

and define $\bar{M}_0 = 0$. This yields a subdivision of the original time interval from t_0 to t_f into $\bar{M}_N$ subintervals. The boundaries of these subintervals are called nodes and are denoted by $t^{(i)}$ for $i = 1, 2, \ldots, \bar{M}_N + 1$. Note that $t_0 = t^{(1)}$, $t_i = t^{(\bar{M}_i+1)}$, and $t_f = t_N = t^{(\bar{M}_N+1)}$. A nondimensional elemental time τ is defined as

$$\tau = \frac{t - t^{(i)}}{t^{(i+1)} - t^{(i)}} = \frac{t - t^{(i)}}{\Delta t^{(i)}} \qquad \text{so that} \qquad 0 \le \tau \le 1 \tag{19}$$

Since first-order time derivatives of δx and $\delta\lambda$ appear in Eq. (16), linear shape functions are chosen for δx and $\delta\lambda$ along the ith interval $t^{(i)} \le t \le t^{(i+1)}$ for $i = 1, 2, \ldots, \bar{M}_N$. Since no time derivatives of δu appear (or $\delta\eta$ if there were a state constraint active in a particular phase), then piecewise constant shape functions may be chosen along each interval. The following functions were chosen:

$$\delta x = \delta x_i^+(1 - \tau) + \delta x_{i+1}^- \tau \tag{20}$$

$$\delta\lambda = \delta\lambda_i^+(1 - \tau) + \delta\lambda_{i+1}^- \tau \tag{21}$$

and (with δ_D defined as the Dirac delta function)

$$\delta u = \delta\hat{u}_i^+ \delta_D(\tau) + \delta\bar{u}_i + \delta\hat{u}_{i+1}^- \delta_D(\tau - 1) \tag{22}$$

$$\delta\eta = \delta\hat{\eta}_i^+ \delta_D(\tau) + \delta\bar{\eta}_i + \delta\hat{\eta}_{i+1}^- \delta_D(\tau - 1) \tag{23}$$

The Dirac delta functions appearing in the discretization of δu and $\delta\eta$ have the effect that the associated parts of the integrand of Eq. (16) are forced to zero pointwise wherever the delta functions have nonzero values.

Since no derivatives of x, λ, u, or η appear in Eq. (16), piecewise constant shape functions are chosen to be

$$x = \begin{cases} \hat{x}_i^+ & \text{if} \quad \tau = 0 \\ \bar{x}_i & \text{if} \quad 0 < \tau < 1 \\ \hat{x}_{i+1}^- & \text{if} \quad \tau = 1 \end{cases} \tag{24}$$

$$\lambda = \begin{cases} \hat{\lambda}_i^+ & \text{if} \quad \tau = 0 \\ \bar{\lambda}_i & \text{if} \quad 0 < \tau < 1 \\ \hat{\lambda}_{i+1}^- & \text{if} \quad \tau = 1 \end{cases} \tag{25}$$

$$u = \begin{cases} \hat{u}_i^+ & \text{if} \quad \tau = 0 \\ \bar{u}_i & \text{if} \quad 0 < \tau < 1 \\ \hat{u}_{i+1}^- & \text{if} \quad \tau = 1 \end{cases} \tag{26}$$

$$\eta = \begin{cases} \hat{\eta}_i^+ & \text{if} \quad \tau = 0 \\ \bar{\eta}_i & \text{if} \quad 0 < \tau < 1 \\ \hat{\eta}_{i+1}^- & \text{if} \quad \tau = 1 \end{cases} \tag{27}$$

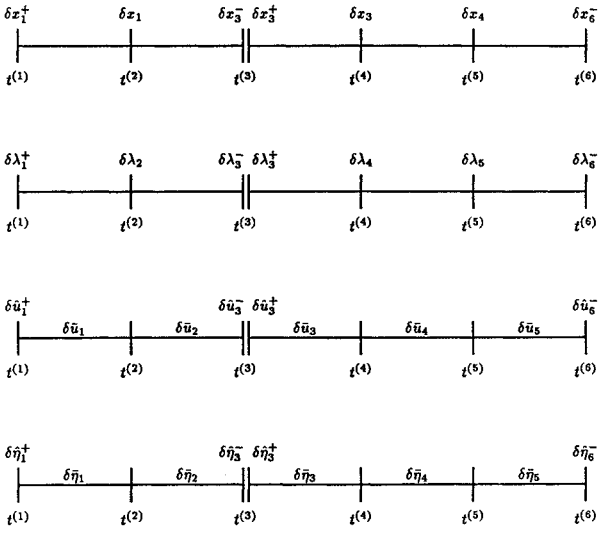

Fig. 1 Location of independent variational quantities for δx, $\delta\lambda$, δu, and $\delta\eta$.

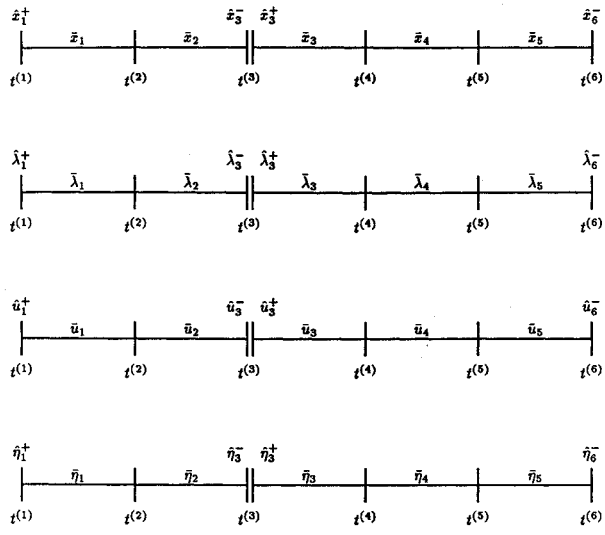

Fig. 2 Location of unknowns for x, λ, u, and η.

The superscript minus and plus signs in Eqs. (20–27) signify values just before and after the nodal times $t^{(i)}$. For all nodes except event nodes the values for x and λ, as well as for δx and $\delta\lambda$, are equal on either side of the node. In other words, only at an event are the states, costates, and their variations allowed to jump in value. For an example case with $M_1 = 2$ and $M_2 = 3$, the quantities defined in Eqs. (20–23) are shown on a timeline in Fig. 1, and the quantities defined in Eqs. (24–27) are shown in Fig. 2.

After substituting τ for t and inserting the above shape functions into Eq. (16), the element quadratures can be carried out analytically, yielding an algebraic equation. Since Eq. (16) must be satisfied for all independently chosen perturbations δx, $\delta\lambda$, and δu ($\delta\eta$ also if the optimal solution involves a state-constrained arc), Eq. (16) is transformed into the discretized condition that

$$
\begin{aligned}
&\sum_{j=1}^{N}\Bigg\langle \delta\lambda^{+T}_{\bar{M}_{j-1}+1}\left(-\hat{x}^{+}_{\bar{M}_{j-1}+1}+\bar{x}_{\bar{M}_{j-1}+1}-\frac{\Delta t_j}{2}\bar{f}_{\bar{M}_{j-1}+1}\right)\\
&+\delta x^{+T}_{\bar{M}_{j-1}+1}\left[\hat{\lambda}^{+}_{\bar{M}_{j-1}+1}-\bar{\lambda}_{\bar{M}_{j-1}+1}-\frac{\Delta t_j}{2}\left(\frac{\partial\bar{H}}{\partial\bar{x}}\right)^{T}_{\bar{M}_{j-1}+1}\right]\\
&+\delta\hat{u}^{+T}_{\bar{M}_{j-1}+1}\left[\left(\frac{\partial\hat{H}}{\partial\hat{u}}\right)^{+T}_{\bar{M}_{j-1}+1}\right]+\delta\hat{\eta}^{+}_{\bar{M}_{j-1}+1}\left[\left(\frac{\mathrm{d}^p\hat{S}}{\mathrm{d}t^p}\right)^{+}_{\bar{M}_{j-1}+1}\right]\\
&+\sum_{i=\bar{M}_{j-1}+2}^{\bar{M}_j}\Bigg\{\delta\lambda_i^{T}\left(-\bar{x}_{i-1}-\frac{\Delta t_j}{2}\bar{f}_{i-1}+\bar{x}_i-\frac{\Delta t_j}{2}\bar{f}_i\right)\\
&+\delta x_i^{T}\left[\bar{\lambda}_{i-1}-\frac{\Delta t_j}{2}\left(\frac{\partial\bar{H}}{\partial\bar{x}}\right)^{T}_{i-1}-\bar{\lambda}_i-\frac{\Delta t_j}{2}\left(\frac{\partial\bar{H}}{\partial\bar{x}}\right)^{T}_{i}\right]\Bigg\}\\
&+\sum_{i=\bar{M}_{j-1}+1}^{\bar{M}_j}\left\{\delta\bar{u}_i^{T}\left[\left(\frac{\partial\bar{H}}{\partial\bar{u}}\right)^{T}_{i}\right]+\delta\bar{\eta}_i\left[\left(\frac{\mathrm{d}^p\bar{S}}{\mathrm{d}t^p}\right)_{i}\right]\right\}\\
&+\delta\lambda^{-T}_{\bar{M}_j+1}\left(-\bar{x}_{\bar{M}_j}-\frac{\Delta t_j}{2}\bar{f}_{\bar{M}_j}+\hat{x}^{-}_{\bar{M}_j+1}\right)\\
&+\delta x^{-T}_{\bar{M}_j+1}\left[\bar{\lambda}_{\bar{M}_j}-\frac{\Delta t_j}{2}\left(\frac{\partial\bar{H}}{\partial\bar{x}}\right)^{T}_{\bar{M}_j}-\hat{\lambda}^{-}_{\bar{M}_j+1}\right]\\
&+\delta\hat{u}^{-T}_{\bar{M}_j+1}\left[\left(\frac{\partial\hat{H}}{\partial\hat{u}}\right)^{-T}_{\bar{M}_j+1}\right]+\delta\hat{\eta}^{-}_{\bar{M}_j+1}\left[\left(\frac{\mathrm{d}^p\hat{S}}{\mathrm{d}t^p}\right)^{-}_{\bar{M}_j+1}\right]\Bigg\rangle=0 \qquad (28)
\end{aligned}
$$

In Eq. (28), note that the superscript minus and plus signs are dropped except at the event nodes where discontinuities may take place in both the state and costate quantities and the variational state and variational costate quantities. Also, $\bar{H} = H(\bar{x}, \bar{u}, \bar{\lambda}, \bar{\eta})$, $\bar{f} = f(\bar{x}, \bar{u})$, $\bar{S} = S(\bar{x})$, $\hat{H} = H(\hat{x}, \hat{u}, \hat{\lambda}, \hat{\eta})$, and $\hat{S} = S(\hat{x})$. Additionally, we have taken the elements within each phase j to be of constant width Δt_j. Furthermore, note that there would be no $\delta\eta$ equations in phases that have no active state constraints. Finally, in the above equation, $\delta\hat{u}$ terms should appear at every node; however, the $\delta\hat{u}$ terms at internal nodes are decoupled from the other ones and may be used to solve for internal nodal values of the control after the other unknowns have been solved.

When the shape functions for x, λ, and u as defined in Eqs. (24–26) are substituted into the boundary conditions of Eqs. (11–15), the following equations are obtained:

$$\frac{\partial\Phi}{\partial t_i}+\hat{H}^{-}_{\bar{M}_i+1}-\hat{H}^{+}_{\bar{M}_i+1}=0 \qquad i=1,2,\ldots,N-1 \qquad (29)$$

$$\frac{\partial\Phi}{\partial t_N}+\hat{H}^{-}_{\bar{M}_N+1}=0 \qquad (30)$$

$$\psi\left[\hat{x}^{+}_1,\hat{x}^{-}_{\bar{M}_1+1},\ldots,\hat{x}^{-}_{\bar{M}_N+1};t_0,\ldots,t_N\right]=0 \qquad (31)$$

$$\frac{\partial\Phi}{\partial\hat{x}\left(t_i^{-}\right)}-\hat{\lambda}^{-T}_{\bar{M}_i+1}=0 \qquad i=1,2,\ldots,N \qquad (32)$$

$$\frac{\partial\Phi}{\partial\hat{x}\left(t_i^{+}\right)}+\hat{\lambda}^{+T}_{\bar{M}_i+1}=0 \qquad i=0,1,\ldots,N-1 \qquad (33)$$

Now, in Eq. (28), the coefficient of each arbitrary virtual quantity (δx, $\delta\lambda$, δu, and $\delta\eta$) must be set equal to zero in order to satisfy the equation. When the coefficients are set to zero and these equations are combined with Eqs. (29–33), the result is a sparse system of nonlinear equations whose size depends on the number of elements. A summary of the number of equations and unknowns is now given.

Summary of Equations and Unknowns

Assuming there are no state constraints for the moment, for any given phase i, the unknowns appearing in Eqs. (28–33) are the barred quantities for each element for x, λ, and u, along with the hatted (nodal) quantities at the beginning and end of each phase for x, λ, and u (see Fig. 2). Thus, there are a total of $(2n+m)(M_i+2)$ unknowns for the ith phase. Summing these over N phases results in $(2n+m)(\bar{M}_N+2N)$ total barred or hatted quantities for x, λ, and u. In addition to these unknowns, there are q unknown multipliers ν and the N unknown times $t_1, t_2, \ldots, t_N$. Therefore, the total number of variables to solve for is $(2n+m)(\bar{M}_N+2N)+q+N$.

For any given phase i, the equations appearing in Eq. (28) are the coefficients of δx, $\delta\lambda$, and δu when set equal to zero. Since the δx and $\delta\lambda$ quantities appear at each node (see Fig. 1), there are $2n(M_i+1)$ equations for the ith phase. Also, Eqs. (32) and (33) provide $2n$ additional equations for each phase. Since the δu quantities appear at the end nodes of each phase, and also at the midpoints of each element (see Fig. 1), there are $m(M_i+2)$ equations. The total number of equations, then, from phase i is $(2n+m)(M_i+2)$. Summing these over N phases results in $(2n+m)(\bar{M}_N+2N)$ total equations. Additionally, Eq. (31) provides q equations, and Eqs. (29) and (30) provide N equations. Thus the total number of equations is $(2n+m)(\bar{M}_N+2N)+q+N$, which agrees with the number of unknowns.

If there was one active state constraint on a particular phase i, then there would be some additional unknowns and equations. The unknowns would be the barred quantities for η at the midpoint of each element and the hatted quantities for η at the end nodes of the phase, for a total of M_i+2 unknowns. Additional equations are available to solve for these unknowns from setting the coefficients of $\delta\eta$ equal to zero. Since the $\delta\eta$ quantities appear at the end nodes of each phase, and also at the midpoints of each element, there are M_i+2 equations. Thus, there are still the same number of equations as unknowns.

Example 1: A Second-Order Problem

Consider the double-integrator problem in Sec. 3.11 of Ref. 20. Let x and v define the position and velocity of a particle in rectilinear motion. The problem is to choose the control force u to minimize

$$J = \int_0^{t_1} \frac{1}{2}u^2 \, dt$$

subject to $\dot{x} = v$, $\dot{v} = u$ and the state and time boundary conditions

$$\psi = \begin{Bmatrix} x(t_0) \\ v(t_0) - 1 \\ x(t_1) \\ v(t_1) + 1 \\ t_1 - 1 \end{Bmatrix} = 0 \tag{34}$$

The problem is complicated when a second-order state inequality constraint is added of the form $S = x - L$, where L is a constant. The first and second total time derivatives of S yield $\dot{S} = \dot{x} = v$ and $\ddot{S} = \dot{v} = u$. The state constraint is not active for $L \geq 0.25$. For certain lower choices of L, the solution exhibits a touch-point behavior, and for other L the solution rides a boundary arc.

The Variational Trajectory Optimization Tool Set program described in Ref. 7 was used to obtain all the finite element results presented in this paper. This code attempts to find a solution of the algebraic equations in Eqs. (28–33). Figure 3 shows the displacement of the particle and Fig. 4 shows the control history.

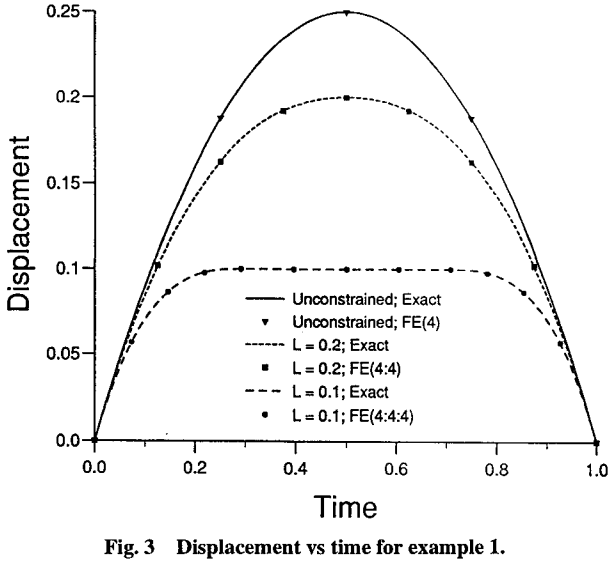

Fig. 3 Displacement vs time for example 1.

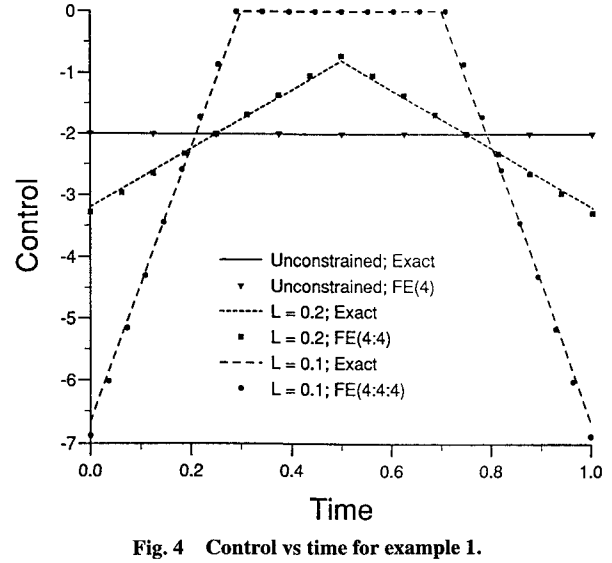

Fig. 4 Control vs time for example 1.

Table 1 Allowable values of constraint limit for touch-point solution vs number of elements

Elements per phase	Minimum value of L for touch-point solution
2:2	0.125 (0.124)
4:4	0.146 (0.145)
8:8	0.157 (0.156)
16:16	0.162 (0.161)
32:32	0.165 (0.164)
Exact	$0.16666666\overline{6}$

Note: Numbers in parentheses denote values of L where constraint is violated.

Table 2 Allowable values of constraint limit for boundary arc solution vs number of elements

Elements per phase	Maximum value of L for boundary arc solution
2:2:2	0.1875 (0.1876)
4:4:4	0.1718 (0.1719)
8:8:8	0.1679 (0.1680)
16:16:16	0.1669 (0.1670)
32:32:32	0.1667 (0.1668)
Exact	$0.16666666\overline{6}$

Note: Numbers in parentheses denote values of L where $t_2 < t_1$.

In the figures, the exact unconstrained solution is graphed versus a four-element finite element solution. Also, a touch-point case (with $L = 0.2$) is graphed versus a finite element run with four elements in each phase, denoted by FE(4:4). Finally, a value of $L = 0.1$ was chosen to demonstrate the exact solution versus the finite element solution (four elements per phase) for a boundary arc case. It is seen for this simple example that there is excellent agreement between all results for both the displacement and the control.

For states and costates it is appropriate to plot the nodal values since the midpoint, or bar values, are just an average of the nodal values. Hence, for a four-element solution, there will be five distinct points. For the control, however, repeated use of the optimality condition $\partial H/\partial u = 0$ supplies information about the control at the nodes and midpoints; thus, there are nine data points for a four-element solution. Note, however, that only the midpoint values are used in the actual discretized state and costate equations.

Of greater interest in this problem is the accuracy with which the finite elements approximate the switching structures (i.e., whether the solution is unconstrained or has a touch-point or boundary arc solution). For this particular example, the finite element approach identifies the exact unconstrained value of $L = 0.25$ for the displacement. However, the value of L for which there is a touch-point solution or a boundary arc solution is a function of the number of elements. The numerical values displayed in Table 1 were generated by assuming a touch-point solution. Table 1 shows the smallest value that L can take and still yield satisfactory results. For instance, when running just two elements per phase (the first row of the table), L could be set as low as 0.125 and still yield an approximate answer that satisfies all the discretized necessary conditions. If, however, $L = 0.124$, then the discretized answer violates the constraint limit before and after the touch point (which always remains at $t = 0.5$). As the number of elements is increased, the allowable values for L increases toward the exact value of $1/6$.

Table 2 was generated under the assumption that a boundary arc case existed. This table displays the maximum value of L versus the number of elements where a boundary arc solution exists. For example, with two elements per phase, we could set $L = 0.1875$. At this value of L, the entry and exit times are almost equal (approximately 0.5 for each). At $L = 0.1876$, the code produces a solution such that the exit time t_2 is less than the entry time t_1, a nonsense solution.

By comparing Tables 1 and 2, it appears that fewer elements per phase are required to approximate the limiting value between a touch-point solution and a boundary arc solution ($L = 1/6$) when assuming the boundary arc solution. It is not known if this is true in general.

Example 2: Constrained Rocket Trajectory Problem

A constrained launch vehicle will now be examined to evaluate the usefulness of the finite element method in obtaining optimal trajectories for a more difficult problem. Consider the following model of a two-stage rocket with the states chosen to be mass m, altitude h, velocity V, and flight-path angle γ and the control to be the angle of attack α. The problem is to choose α to minimize $J = \phi = -m(t_f)$ subject to the dynamic equations

$$\begin{aligned} \dot{m} &= -\frac{T_{\text{vac}}}{g I_{\text{sp}}} \\ \dot{h} &= V \sin\gamma \\ \dot{V} &= \frac{T\cos\alpha - D}{m} - \frac{\mu \sin\gamma}{r^2} \\ \dot{\gamma} &= \frac{T\sin\alpha + L}{mV} + \left(\frac{V}{r} - \frac{\mu}{r^2 V}\right)\cos\gamma \end{aligned} \tag{35}$$

where $T = T_{\text{vac}} - A_e p$, T_{vac} is the thrust in a vacuum, A_e is the exit area of the nozzles, p is the local atmospheric pressure, I_{sp} is the specific impulse, g is acceleration due to gravity at sea level, μ is Earth's gravitational constant, and r is the distance from the center of Earth to the vehicle (radius of Earth R_e plus altitude h). Defining ρ as the density, S as the reference area, and C_a and C_N as the axial and normal force coefficients, respectively, the dynamic pressure q, the drag D, and the lift L are defined as

$$\begin{aligned} q &= \tfrac{1}{2}\rho V^2 \\ D &= qS\left(C_a + C_N\alpha^2\right) \\ L &= qS(C_N - C_a)\alpha \end{aligned} \tag{36}$$

The numerical values for the physical constants are $\mu = 3.986 \times 10^{14}$ m^3/s^2, $R_e = 6.378 \times 10^6$ m, and $g = 9.81$ m/s^2. The propulsive model used was

$$\begin{aligned} T_{\text{vac}} &= 1.2975 \times 10^7 \text{ N} \quad \text{before staging} \\ T_{\text{vac}} &= 2.5950 \times 10^6 \text{ N} \quad \text{after staging} \\ A_e &= 19.115 \text{ m}^2 \quad \text{before staging} \\ A_e &= 3.823 \text{ m}^2 \quad \text{after staging} \\ I_{\text{sp}} &= 430.55 \text{ s} \end{aligned} \tag{37}$$

and the atmospheric and aerodynamic model was

$$\begin{aligned} \rho &= 1.225\exp(-h/8600) \text{ kg/m}^3 \\ p &= 101{,}325\exp(-h/7600) \text{ Pa} \\ S &= 55.18 \text{ m}^2 \\ C_a &= 0.3 \qquad C_N = 3.1 \end{aligned} \tag{38}$$

We note that constant values for the aerodynamic coefficients are unrealistic; however, this assumption does simplify the numerical operations needed to solve the problem. Also, for numerical simplicity, all aerodynamic and atmospheric terms were neglected after staging, since they are extremely small beyond this point in the trajectory.

Defining $m_0 = 890{,}150$ kg, $h_f = 148{,}160$ m, and $V_f = 7854$ m/s, all the state and time boundary conditions required for the finite element formulation are listed below for the unconstrained and constrained cases. For the unconstrained case, this is a two-phase problem where the change in phase is dictated by a known staging time. The boundary conditions are

$$\psi = \left\{ \begin{array}{c} m(t_0) - m_0 \\ h(t_0) - 60 \\ V(t_0) - 25 \\ \gamma(t_0) - 1.5 \\ m(t_1^+) - m(t_1^-) + 29{,}920 \\ h(t_1^+) - h(t_1^-) \\ V(t_1^+) - V(t_1^-) \\ \gamma(t_1^+) - \gamma(t_1^-) \\ h(t_2) - h_f \\ V(t_2) - V_f \\ \gamma(t_2) \\ t_1 - 195 \end{array} \right\} = 0$$

A dynamic pressure constraint is now added to the problem of the form $S(h, V) = q - q_{\text{lim}} \leq 0$. This is a first-order state constraint, as is seen by taking the first total time derivative, substituting in values for ρ, $\dot{h}$, and $\dot{V}$ and simplifying to obtain the following control-dependent algebraic constraint that must be enforced during the time interval that $q = q_{\text{lim}}$:

$$\frac{T\cos\alpha - D(\alpha)}{m} - \frac{\mu\sin\gamma}{r^2} - \frac{V^2\sin\gamma}{2 \times 8600} = 0 \tag{39}$$

This constraint will be enforced as one boundary arc occurring before the staging time. Thus the first phase of the unconstrained problem will be broken into three phases consisting of an unconstrained, constrained, and unconstrained arc, followed by the final phase from staging to the final time. The boundary condition vector must also be modified to include the tangency conditions and extra internal boundary conditions. The vector is

$$\psi = \left\{ \begin{array}{c} m(t_0) - m_0 \\ h(t_0) - 60 \\ V(t_0) - 25 \\ \gamma(t_0) - 1.5 \\ q(t_1^+) - q_{\text{lim}} \\ x(t_1^+) - x(t_1^-) \\ x(t_2^+) - x(t_2^-) \\ m(t_3^+) - m(t_3^-) + 29{,}920 \\ h(t_3^+) - h(t_3^-) \\ V(t_3^+) - V(t_3^-) \\ \gamma(t_3^+) - \gamma(t_3^-) \\ h(t_4) - h_f \\ V(t_4) - V_f \\ \gamma(t_4) \\ t_3 - 195 \end{array} \right\} = 0$$

where $x = [m \;\; h \;\; V \;\; \gamma]^T$. For the generation of the numerical results that follow, the states were normalized in a simple manner by defining $\bar{m} = m/m_0$, $\bar{h} = h/h_f$, and $\bar{V} = V/V_f$, where m_0 is the initial mass and h_f and V_f are the final altitude and velocity, respectively. Normalizing the states automatically scales the costates.

Obtaining guesses for a small number of elements is generally easier than for a large number of elements. Additionally, once a solution has been found for a few elements, more elements may be added readily in many cases. For example, Table 3 was produced after a two-element-per-phase, unconstrained solution was found (a very crude approximation). By linearly interpolating the solution,

Table 3 Buildup of finite element values

Elements per phase	Average initial error	Number of iterations
4:4	2.60015×10^{-2}	5
8:8	9.93419×10^{-3}	4
16:16	2.35328×10^{-3}	3
32:32	4.81494×10^{-4}	3

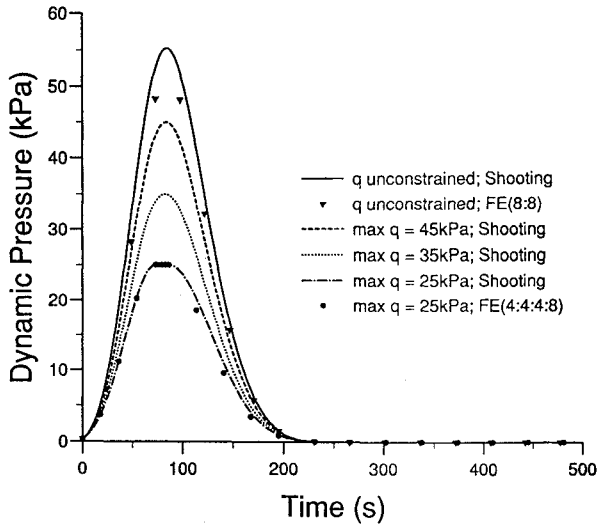

Fig. 5 Dynamic pressure history.

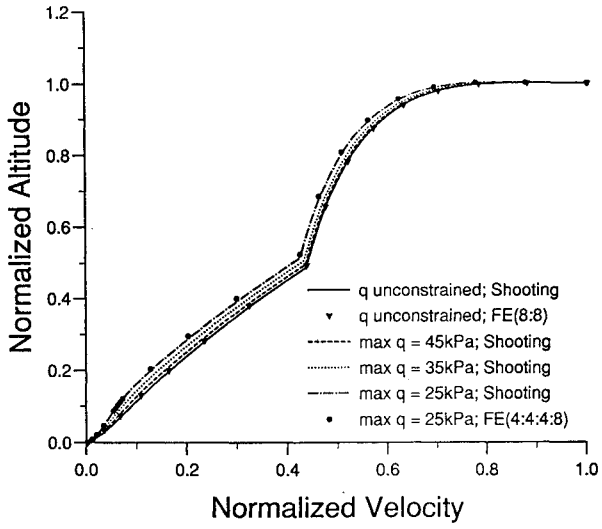

Fig. 6 Normalized altitude vs normalized velocity profile.

new initial guesses were computed for a four-element-per-phase solution. These new guesses yielded an average initial error of the algebraic equations of 2.60015×10^{-2}, as displayed in the second column. The last column of Table 3 displays the total number of iterations, using a Newton method, required to obtain a converged solution. The 4:4 solution was then used to obtain initial guesses for the 8:8 case, and similarly, the table was completed. It is seen that the initial error continually drops, as does the number of iterations required for convergence. Doubling the number of elements has proven to be a quick and easy way of obtaining reasonably accurate answers that may be used in a shooting code as initial guesses.

Some of the results are shown in Figs. 5–10. In all of the figures, the unconstrained shooting results are shown, along with the unconstrained finite element results taken with eight elements per phase and denoted by FE(8:8) in the figures. The maximum value of the dynamic pressure q is approximately 55.3 kPa. Also in these figures shooting results are given for the constrained cases where the maximum allowable value for the dynamic pressure is 45, 35, and 25 kPa. Finite element results were obtained for all the constraint

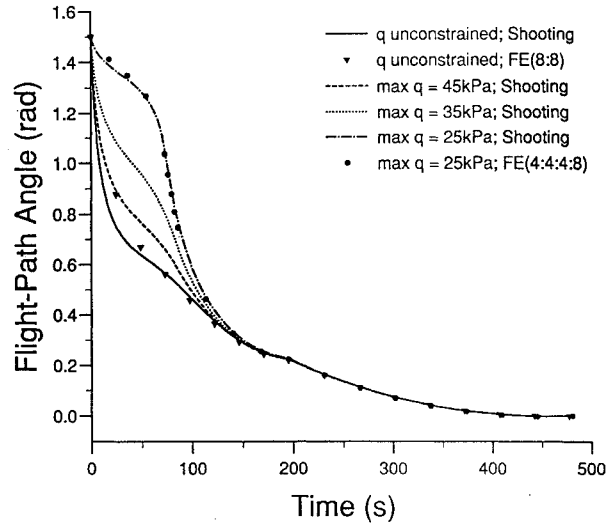

Fig. 7 Flight-path angle history.

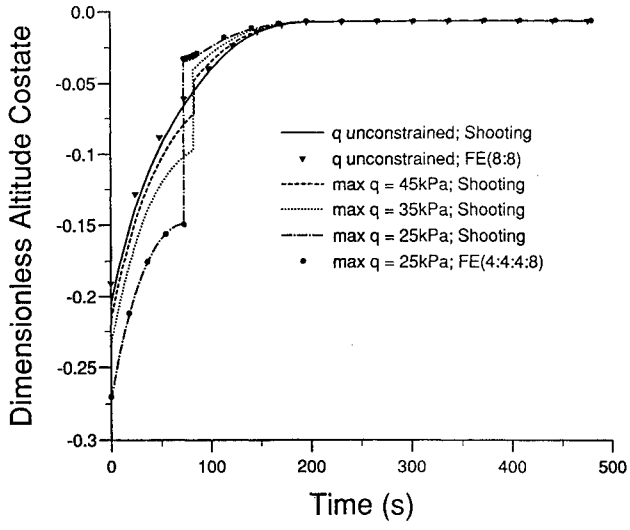

Fig. 8 Dimensionless altitude costate history.

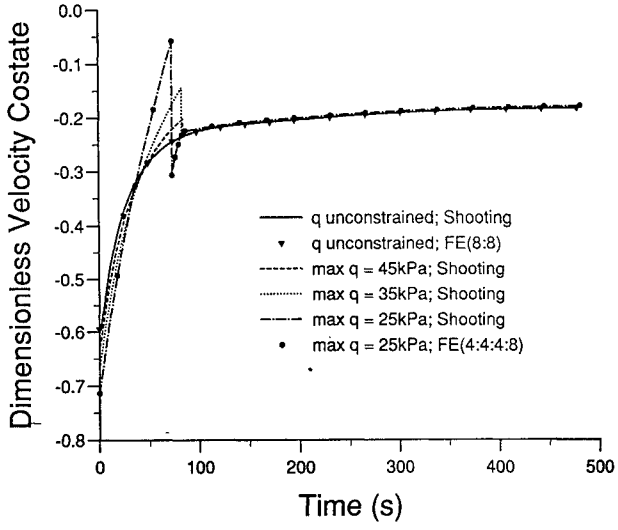

Fig. 9 Dimensionless velocity costate history.

values; however, for clarity in the figures, results will only be presented for the 25-kPa case. The finite element results are for four elements in each of the first three phases and eight elements in the last phase (the one after staging) and are denoted with FE(4:4:4:8). The dynamic pressure histories are shown in Fig. 5.

Figure 6 shows the normalized altitude profile vs the normalized velocity. As expected, the vehicle climbs higher more quickly as

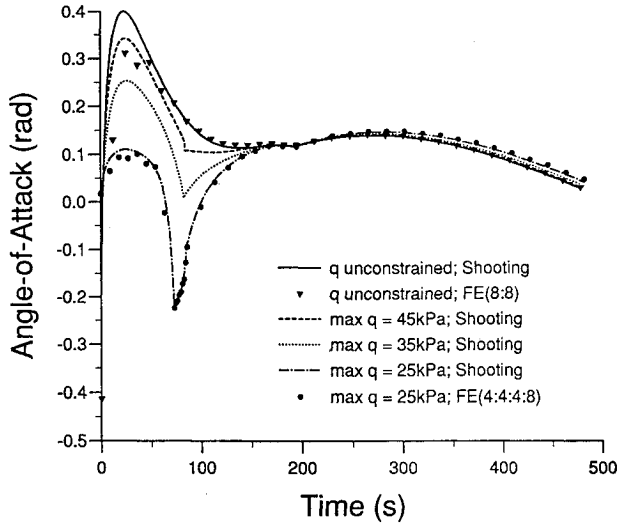

Fig. 10 Angle-of-attack (control) history.

the constraint becomes stricter. Note also that even with the coarse discretization presented, the finite element results lie essentially on top of the exact curves. Figure 7 shows the flight-path angle γ versus time. The vehicle is seen to roll over more slowly with the strict q limit in order to climb out of the atmosphere more quickly. Notice that the finite element results show some error for the unconstrained results but appear much more accurate for the constrained case. The error in the unconstrained case is due to the steep gradient in the γ profile at the initial time. A finer mesh increases the accuracy in this area.

Figures 8 and 9 give two costate histories for the vehicle and demonstrate the jump that occurs in these multipliers due to the internal boundary condition that $q = q_{\text{lim}}$. Note also that the finite elements have solved for this jump and the initial costate values accurately. These values are, of course, critical to the shooting code.

The control history is shown in Fig. 10. The control experiences large variations due to the changing constraint limits. Once again the finite element results are less accurate at the initial time due to the steep gradients there.

Conclusions

In this paper, it has been shown that the finite element formulation is useful for the solution of optimal control problems with inequality constraints that are functions of the state variables. The algebraic equations may be defined prior to specifying the dynamics of the problem, and the solution of these equations may be obtained quickly when the sparsity of the Jacobian is exploited. The following has been observed:

1) It is helpful to obtain a converged answer for a small number of elements and then add elements using guesses obtained from a previously converged solution.

2) This formulation can be useful in determining the switching structure for an optimal trajectory.

3) The solution of the finite element equations can provide initial guesses for a shooting code in order to generate numerically exact answers.

The application of this method to a constrained booster trajectory problem has allowed for a family of solutions to be generated and shooting results to be obtained.

Acknowledgments

This work was supported by NASA Contracts NAS1-18935 and NAS1-19000 and NASA Grant NAG-1-1435. Technical discussions with Daniel D. Moerder are gratefully acknowledged.

References

[1]Betts, J., and Huffman, W., "The Application of Sparse Nonlinear Programming to Trajectory Optimization," *Proceedings of the 1990 AIAA Guidance, Navigation, and Control Conference*, AIAA, Washington, DC, pp. 1198–1218 (AIAA Paper 90-3448).

[2]Gill, P. E., Murray, W., and Wright, M. H., *Practical Optimization*, Academic, New York, 1981.

[3]Lee, E. B., and Markus, L., *Foundations of Optimal Control Theory*, Wiley, New York, 1967.

[4]Berkovitz, L. D., *Optimal Control Theory*, Springer–Verlag, New York, 1974.

[5]Stoer, J., and Bulirsch, R., *Introduction to Numerical Analysis*, Springer–Verlag, New York, 1980.

[6]Oberle, H. J., and Grimm, W., "BNDSCO: A Program for the Numerical Solution of Optimal Control Problems," English Translation of DLR-Mitt. 85-05, Oberpfaffenhofen, Germany.

[7]Bless, R. R., Queen, E. M., Cavanaugh, M. D., Wetzel, T. A., and Moerder, D. D., "Variational Trajectory Optimization Tool Set; Technical Description and User's Manual," NASA TM 4442, July 1993.

[8]Seywald, H., "Optimal Control Problems with Switching Points," NASA CR-4393, Sept. 1991.

[9]Seywald, H., and Cliff, E. M., "Goddard Problem in Presence of a Dynamic Pressure Limit," *Journal of Guidance, Control, and Dynamics*, Vol. 16, No. 4, 1993, pp. 776–781.

[10]Roberts, S. M., and Shipman, J. S., *Two-Point Boundary Value Problems: Shooting Methods*, American Elsevier, New York, 1972.

[11]Bosarge, W. E., Jr., and Johnson, O. G., "Error Bounds of High Order Accuracy for the State Regulator Problem Via Piecewise Polynomial Approximations," *SIAM Journal on Control*, Vol. 9, No. 1, 1971, pp. 15–28.

[12]Chen, G., and Mills, W. H., Jr., "Finite Elements and Terminal Penalization for Quadratic Cost Optimal Control Problems Governed by Ordinary Differential Equations," *SIAM Journal on Control and Optimization*, Vol. 19, No. 6, 1981, pp. 744–764.

[13]Hodges, D. H., and Bless, R. R., "Weak Hamiltonian Finite Element Method for Optimal Control Problems," *Journal of Guidance, Control, and Dynamics*, Vol. 14, No. 1, 1991, pp. 148–156.

[14]Hodges, D. H., Bless, R. R., Calise, A. J., and Leung, M., "Finite Element Method for Optimal Guidance of an Advanced Launch Vehicle," *Journal of Guidance, Control, and Dynamics*, Vol. 15, No. 3, 1992, pp. 664–671.

[15]Bless, R. R., and Hodges, D. H., "Finite Element Solution of Optimal Control Problems with State-Control Inequality Constraints," *Journal of Guidance, Control, and Dynamics*, Vol. 15, No. 4, 1992, pp. 1029–1032.

[16]Wouk, A., *A Course of Applied Functional Analysis*, Wiley, New York, 1979.

[17]Jacobson, D. H., Lele, M. M., and Speyer, J. L., "New Necessary Conditions of Optimality for Control Problems with State-Variable Inequality Constraints," *Journal of Mathematical Analysis and Applications*, Vol. 35, 1971, pp. 255–284.

[18]Seywald, H., and Cliff, E. M., "On the Existence of Touch Points for First-Order State Inequality Constraints," *AIAA Guidance, Navigation, and Control Conference*, AIAA, Washington, DC, 1993, pp. 372–376 (AIAA Paper 93-3743).

[19]Cliff, E. M., Seywald, H., and Bless, R. R., "Hodograph Analysis in Aircraft Trajectory Optimization," *AIAA Guidance, Navigation, and Control Conference*, AIAA, Washington, DC, 1993, pp. 363–371 (AIAA Paper 93-3742).

[20]Bryson, A. E., Jr., and Ho, Y.-C., *Applied Optimal Control*, Hemisphere, New York, 1975.

[21]Bryson, A. E., Jr., Denham, W. F., and Dreyfus, S. E., "Optimal Programming Problems with Inequality Constraints, I: Necessary Conditions for Extremal Solutions," *AIAA Journal*, Vol. 1, No. 11, 1963, pp. 2544–2550.