SIAM J. NUMER. ANAL.
Vol. 13, No. 4, September 1976

RATES OF CONVERGENCE FOR DISCRETE APPROXIMATIONS TO UNCONSTRAINED CONTROL PROBLEMS*

WILLIAM W. HAGER†

Abstract. Convergence rates for the error between the solution to a discrete approximation of a fixed time, unconstrained control problem and the corresponding continuous optimal control are derived for one-step and multistep integration schemes. The convergence rate for multistep schemes depends on the order of the integration scheme and the approximation properties of the discrete costate equation at the right endpoint. Furthermore, the order is $\leqq 3$ and the error in the optimal discrete control exhibits a boundary layer with most of the error concentrated at the right endpoint. For a class of one-step integration schemes satisfying a symmetry condition, second order convergence of the optimal discrete control is both proved and observed experimentally. The computations also indicate that the convergence rate of the optimal discrete state and costate variables equals the order of the integration scheme. By an auxiliary computation, this order can also be recovered for the control approximation. Some numerical examples indicate that the convergence estimates are tight. The question of the best integration scheme for achieving a given order of convergence is examined, and the modified Euler, the 3-point Adams–Moulton and a Runge–Kutta scheme appear to be optimal for orders 2, 3 and 4.

1. Introduction. Discrete approximations of the following control problem are studied:

$$\text{Minimize } \{x_0(1) : \dot{x}(t) = f(x(t), u(t)), x(0) = z\}, \tag{1}$$

where the minimization in (1) is over $u(\cdot)$ and where $x : R \to R^{n+1}$, $u : R \to R^m$, $f : R^{n+1} \times R^m \to R^{n+1}$, and x_0 is the zeroth component of x. The variable x is called the *state variable* or *trajectory* generated by the *control* u. It is assumed that f is not a function of x_0 and $(z)_0 = 0$ so that the cost functional to be minimized in (1) is the following:

$$x_0(1) = \int_0^1 f_0(x(t), u(t))\, dt. \tag{2}$$

The differential equations describing the remaining components of x are called the *system dynamics*. By a discrete approximation to the control problem, we mean that the differential equation is to be replaced by an integration procedure such as modified Euler's method, fourth order Runge–Kutta, or the fifth order Milne scheme. This paper estimates the convergence rate as a function of the grid interval for the error between the solution to the discrete optimization problem and the continuous optimal control.

A number of papers have developed conditions under which the solution to a discrete approximation to (1) will converge to the optimal continuous control. For example, see Cullum [6], [7], [8], Budak, Berkovich and Solov'eva [2], [3], [4], and Klessig and Polak [13]. The question of convergence rates, however, has not been considered although these rates are very important in making judgements concerning the efficiency of various discrete approximations; in

* Received by the editors July 19, 1974, and in final revised form October 24, 1975.
† Department of Mathematics, University of South Florida, Tampa, Florida 33620.

fact, the simplest scheme that achieves a given convergence rate is generally the most efficient scheme.

The results that follow are based on considering the relationship between the *discrete necessary conditions* or *Kuhn–Tucker* conditions and the *continuous necessary conditions* or *Pontryagin minimum principle* (see L. S. Pontryagin et al. [16]). The continuous necessary conditions are the following: Suppose u^* solves (1) and x^* is the corresponding trajectory; whenever $f(\cdot, \cdot)$ is sufficiently smooth (see [16]), there exists a function p^* called the *costate variable* such that the following conditions hold for $(x, u, p) = (x^*, u^*, p^*)$:

$$\dot{x}(t) = f(x(t), u(t)), \qquad x(0) = z, \tag{3}$$

$$\dot{p}(t) = -f_x(x(t), u(t))^T p(t), \quad p(1)^T = (1, 0, 0, \cdots, 0)^T, \tag{4}$$

$$f_u(x(t), u(t))^T p(t) = 0. \tag{5}$$

The conditions (3), (4) and (5) are called the *state equation*, *costate equation* and *control minimum principle*, respectively.

Notice that (3)–(5) define a two-point boundary value problem since (5) defines $u(t)$ as a function of $x(t)$ and $p(t)$ and hence (3) and (4) are differential equations in x and p with half the boundary conditions specified at $t = 0$ and the other half at $t = 1$. In fact, one method for solving (1) is to use any of the standard two-point boundary value techniques (see [14], [17]) to compute x^* and p^* and then to determine u^* by (5).

This approach, however, ignores much of the basic structure of the control problem. It has already been observed with the finite element method for solving elliptic partial differential equations, that many of the stability problems associated with approximations to the necessary conditions (or the partial differential equation) are eliminated by approximating the variational formulation of the problem. The finite element method generates very unconventional difference approximations to the partial differential equation with very good stability properties.

Similar behavior is observed for discrete approximations to the unconstrained control problem (1). If standard two-point boundary value approximations are used to solve (3)–(5), stability problems arise since most techniques for solving the two-point boundary value problem involve the inversion of a transition matrix associated with the linearized problem and the condition number of this matrix appears to be very large for many problems. On the other hand, if the original optimization problem (1) is discretized, then an algorithm such as conjugate gradients with both guaranteed convergence and a high asymptotic convergence rate can be utilized. As with the finite element method, the necessary conditions for the discrete approximating problem produce some very unconventional approximations to the two-point boundary value problem.

As stated earlier, this paper studies the convergence rate of the optimal discrete control as a function of the grid interval h. If the differential equation in (1) is replaced by a multistep integration scheme, then the convergence rate is $\leqq 3$

and depends on the order of the integration scheme and the approximating properties of the discrete costate equation at the right endpoint, $t = 1$.

Although the Milne schemes possess the highest possible order among all integration schemes involving a given number of points, the approximation properties of the discrete costate variable are very bad near $t = 1$ so that the optimal discrete control does not converge to the optimal continuous control. The errors in the modified Euler and 3-point Adams–Moulton schemes are second and third order, respectively, except for a boundary layer at $t = 1$ where the error is, in general, $O(h)$. With the 4-point Adams–Moulton scheme, however, the convergence rate bound of 3, mentioned above, intervenes to keep the convergence rate at third order.

For one-step approximations to the differential equation in (1), the mathematical theory is much less complete. Using the same techniques employed for the multistep schemes, it is shown that the error in the optimal discrete control is at most $O(h^2)$ for a class of schemes satisfying a symmetry condition. However, numerically it is observed that even though the discrete controls are only accurate to order 2, the optimal discrete state and costate variables associated with a bth order one-step scheme satisfying the symmetry condition are accurate to order b, although the proof of this result in general is still an open question. Using these good approximation properties for the state and costate variable, a bth order estimate of u^* can be obtained. The possibility of choosing the integration parameters so that the optimal discrete control is accurate to order b was also examined. For third and fourth order integration schemes, it was found that the discrete control is accurate at best to second and third order, respectively.

Some numerical examples in the last section indicate that the convergence estimates are tight. Also the question of the best integration scheme for obtaining a given convergence rate is examined, and the modified Euler, 3-point Adams–Moulton and a Runge–Kutta scheme appear to be optimal for orders 2, 3 and 4, respectively.

2. Multistep schemes. Multistep schemes of the following form are studied:

$$\sum_{j=0}^{r} a(j)y(j+k) = h \sum_{j=0}^{r} b(j) f(y(j+k), v(j+k)) \tag{6}$$

for $k = 0, 1, \cdots, N - r$. Above $h = 1/N$, and $(y(k), v(k))$ are approximations to $(x(t_k), u(t_k))$ where $t_k = kh$. The $\{a(j)\}$ and $\{b(j)\}$ are fixed constants that characterize the multistep scheme and Table 1 gives some typical choices for these parameters.

Notice that once the control sequence $\{v(k)\}$ and the r initial conditions $\{y(0), y(1), \cdots, y(r-1)\}$ are given, one can often solve for $\{y(r), y(r+1), \cdots, y(N)\}$; for example, (i) the scheme is explicit ($b(r) = 0$) or (ii) $f(\cdot, \cdot)$ is Lipschitz continuous in its first argument and h is sufficiently small. Normally, the r starting conditions must be determined using a one-step scheme; however, in order to isolate the effect of the multistep scheme on the discrete control error, it will be assumed that the initial conditions are known exactly. Henrici [11, Chap. 5] gives an analysis of error propagation for multistep integration procedures.

It is assumed that the integration procedure (6) converges to the solution of the corresponding differential equation with order $b \geqq 1$, that there exists a

TABLE 1
Multistep schemes

Name of Scheme	$a(i)$	$b(i)$	Order of Scheme	Accuracy of Terminal Condition	Computed Convergence Rate in (P1)
Modified Euler	$(1, -1)$	$(\frac{1}{2}, \frac{1}{2})$	2	2	$1.95 \to 2$
3-point Milne	$(1, 0, -1)$	$(\frac{1}{3}, \frac{4}{3}, \frac{1}{3})$	4	0	0
4-point Adams–Moulton	$(1, -1, 0, 0)$	$(\frac{9}{24}, \frac{19}{24}, -\frac{5}{24}, \frac{1}{24})$	4	4	$3.05 \to 3$
5-point Milne	$(1, 0, 0, 0, -1)$	$(\frac{14}{45}, \frac{64}{45}, \frac{24}{45}, \frac{64}{45}, \frac{14}{45})$	6	0	0

solution u^* to (1) and a corresponding trajectory x^*, and that the minimum principle (3)–(5) holds. Similarly assume that there exists a solution $\{v^h(k)\}$ and a corresponding trajectory $\{y^h(k)\}$ for the discrete optimization problem and that the discrete necessary conditions hold.

In the Kuhn–Tucker conditions below, the following notation is used: $\lambda(k)$ denotes the dual multiplier corresponding to (6), f_x and f_u are gradients of f with respect to x and u, respectively, and $G(k) = f_x(y(k), v(k))^T$. To generate the discrete necessary conditions, multiply (6) by $\lambda(k)$, sum over k, add the result to the cost functional to form the Lagrangian, and equate to zero the gradient of the Lagrangian with respect to $\{y(k)\}$ and $\{v(k)\}$. The equations (7) and (8) below correspond to the state and control gradients, respectively:

$$\sum_{j=0}^{r} [a(j)\lambda(k-j)] = e(k) + h \sum_{j=0}^{r} [G(k)b(j)\lambda(k-j)], \tag{7}$$

$$h \sum_{j=0}^{r} [f_u(y(k), v(k))^T b(j)\lambda(k-j)] = 0 \tag{8}$$

for $k = r, r+1, \cdots, N$, where $\lambda(k) = 0$ for $k > N - r$, $e(k) = 0$ for $k \neq N$, and $e(N)^T = (1, 0, 0, \cdots, 0)^T$. Note that (7) and (8) only hold for $v(k) = v^h(k)$ and $y(k) = y^h(k)$; however, the "h" superscript is omitted below and all variables are understood to be optimal. Now change from the variable λ to the variable q given by

$$q(k) = \sum_{j=0}^{r} [b(j)\lambda(k-j)]. \tag{9}$$

After multiplying (7) by $b(m)$, replacing k with $k - m$ and summing from $m = 0$ to $m = r$, we obtain

$$\begin{aligned} &\sum_{m=0}^{r} \sum_{j=0}^{r} [a(j)b(m)\lambda(k-m-j)] \\ &\quad = \sum_{m=0}^{r} \left\{ h \sum_{j=0}^{r} [b(j)b(m)G(k-m)\lambda(k-j-m)] + b(m)e(k-m) \right\}. \end{aligned} \tag{10}$$

Interchanging the order of summation on the left side of (10), and using (9) in both (10) and (8) leads to the following formulation for the discrete necessary conditions:

$$\sum_{j=0}^{r} [a(j)q(k-j)] = \sum_{j=0}^{r} b(j)[hG(k-j)q(k-j) + e(k-j)] \quad \text{for } k = N+r, N+r-1, \cdots, 2r, \tag{11}$$

$$q(k) = 0 \quad \text{for } k > N, \tag{12}$$

$$hf_u(y(k), v(k))^T q(k) = 0 \quad \text{for } k = r, r+1, \cdots, N. \tag{13}$$

The sequence $\{q(k)\}$ and the equation (11) will be called the *discrete costate variable* and the *discrete costate equation*, respectively.

Since the $e(k-j)$ term in (11) vanishes for $k < N$, the discrete costate equation is the same multistep scheme as (6), except that (4) is integrated in the backward direction; that is, the discrete costate is computed in the order $q(N), q(N-1), \cdots$. Also note the difference in the terminal conditions for the discrete and continuous costate equations: $p(1)^T = (1, 0, 0, \cdots, 0)^T$ in (4) and $q(k) = 0$ for $k > N$ in (12). As will be seen below, the $e(k-j)$ term in (11) corrects for the discrepancy in the terminal conditions. Thus the discrete approximation to the optimization problem gives rise to a very unconventional starting procedure for the costate multistep integration scheme.

A summary is now given of the error analysis. The cost functional (2) is an unconstrained function of the control; that is, given $u(\cdot)$, the corresponding cost $x_0(1)$ can be computed. Thus the discrete cost functional, denoted $J(v(r), v(r+1), \cdots, v(N))$, is an unconstrained function of the discrete controls $\{v(k)\}$. Hence the optimal discrete control satisfies $DJ(v^h) = 0$, where D is the derivative operator. Expanding $DJ(v)$ in a Taylor series, $DJ(v) = DJ(v^h) + D^2J(\mathbf{v})(v - v^h)$ where $\mathbf{v}$ lies on the line segment between v and v^h. Defining v^* by $v^*(k) = u^*(t_k)$, $v^* - v^h = D^2J(\mathbf{v})^{-1}DJ(v^*)$ from the expansion above. Thus the error $v^* - v^h$ depends on the properties of the Hessian $D^2J(\mathbf{v})$ and the gradient $DJ(v^*)$.

Let $\{y^*(k)\}$ denote the discrete state generated by (6) using the discrete controls $\{v^*(k)\}$, and let $\{q^*(k)\}$ denote the discrete costate generated by (11) and (12) using $y(k) = y^*(k)$ and $v(k) = v^*(k)$. We show that $DJ(v^*)$ is given by (13) where $(y, v, q) = (y^*, v^*, q^*)$. Furthermore, $DJ(v^*)_k = O(h^2)$ for $k \approx N$ and $DJ(v^*)_k = O(h^{m+1})$ for $k \ll N$ where $m \leqq b$ and m depends on the effectiveness of the $e(k-j)$ term in (11) in approximating the continuous costate variable near $t = 1$. After estimating $D^2J(v)$ in an example, it is proved that the error $v^h(k) - v^*(k)$ is $O(h)$ for $k \approx N$ and $O(h^l)$ for $k \ll N$ where $l = \min\{3, m\}$.

Step 1. *An expression for* $DJ(v^*)$. Recall the implicit function theorem: If $g: R^n \times R^m \to R^n$, $(\mathbf{a}, \mathbf{b})$ satisfy the relation $g(\mathbf{a}, \mathbf{b}) = 0$, g is continuously differentiable in a neighborhood of $(\mathbf{a}, \mathbf{b})$, and the matrix $\partial g(\mathbf{a}, \mathbf{b})/\partial a$ is nonsingular, then there exists a neighborhood $\mathcal{N}$ of $\mathbf{b}$ such that the equation $g(a, b) = 0$ has a solution $a(b)$ for $b \in \mathcal{N}$, the derivative $\partial a(\mathbf{b})/\partial b$ exists, and $\partial a(\mathbf{b})/\partial b = -(\partial g(\mathbf{a}, \mathbf{b})/\partial a)^{-1} \cdot (\partial g(\mathbf{a}, \mathbf{b})/\partial b)$. Hence if $L: R^n \times R^m \to R$ is a real-valued function differentiable at $(\mathbf{a}, \mathbf{b})$, then by the chain rule, $\partial L(a(b), b)/\partial b|_{b=\mathbf{b}} = \partial L(\mathbf{a}, \mathbf{b})/\partial b + \lambda^T(\partial g(\mathbf{a}, \mathbf{b})/\partial b)$ where λ^T is the solution to the linear system $\partial L(\mathbf{a}, \mathbf{b})/\partial a + \lambda^T(\partial g(\mathbf{a}, \mathbf{b})/\partial a) = 0$.

Restating these last results, $\partial L(a(b), b)/\partial b = \partial \mathscr{L}(a, b, \lambda)/\partial b$ where $\mathscr{L}(a, b, \lambda) = L(a, b) + \lambda^T g(a, b)$ and λ is determined from $\partial \mathscr{L}(a, b, \lambda)/\partial b = 0$.

For the control problem, the constraints $g(\cdot, \cdot)$ are given by (6) where $a_k = y(k)$ and $b_k = v(k)$. Hence $\partial \mathscr{L}(a, b, \lambda)/\partial a$ yields (7) or equivalently (11) and, as stated earlier, $\partial L(a(b), b)/\partial b$ reduces to:

$$DJ(v^*)_k = hf_u(y^*(k), v^*(k))^T q^*(k). \tag{14}$$

By the necessary condition (5) above, $hf_u(x^*(t_k), v^*(k))^T p^*(t_k) = 0$, and hence

$$|DJ(v^*)_k| \leqq ch|q^*(k) - p^*(t_k)| + ch|y^*(k) - x^*(t_k)|, \tag{15}$$

where c depends on the second partial derivatives of $f(\cdot, \cdot)$ near $(x^*(t), u^*(t))$ for $t \in [0, 1]$. Since the multistep scheme (6) is of order b, $|y^*(k) - x^*(t_k)| = O(h^b)$. The next step examines the error in q^*.

Step 2. *The error* $q^*(k) - p^*(t_k)$. First consider the zeroth component of q^*. Since f is not a function of x_0, the first row of f_x^T is identically zero and hence by (4), $\dot{p}_0^*(t) = 0$, $p_0^*(1) = 1$, or equivalently $p_0^* \equiv 1$. Similarly using (11), q_0^* can be computed from the initial condition $q^*(k) = 0$ for $k > N$, and these values are given in Table 2 for the schemes in Table 1.

TABLE 2

Scheme	N	$N-1$	$N-2$	$N-3$	$N-4$	$N-5$	$N-6$	$N-7$	$N-8$
Modified Euler	$\frac{1}{2}$	1	1	1	1	1	1	1	1
Milne 3-point	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{4}{3}$	$\frac{2}{3}$
Adams–Moulton 4-point	$\frac{9}{24}$	$\frac{28}{24}$	$\frac{23}{24}$	1	1	1	1	1	1
Milne 5-point	$\frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$	$\frac{64}{45}$	$\frac{28}{45}$	$\frac{64}{45}$	$\frac{24}{45}$	$\frac{64}{45}$	$\frac{28}{45}$

Notice that for both the modified Euler and Adams–Moulton 4-point scheme, the error $|q_0^*(k) - p_0^*(t_k)|$ vanishes within r grid intervals while the accuracy of the two Milne schemes is $O(1)$ for all k.

Now consider the remaining components of the discrete costate variable. Let $\hat{q}(k)$ denote the solution to (11) when $G(k) = f_x(y(k), v(k))^T$ is replaced by $H(k) = f_x(x^*(t_k), u^*(t_k))^T$. The error $|q^*(k) - p^*(t_k)|$ is estimated by computing $|\hat{q}(k) - p^*(t_k)|$ and $|\hat{q}(k) - q^*(k)|$.

Using a Taylor series expansion, it is possible to express $\hat{q}(k)$ in terms of $f(x^*(1), u^*(1))$ and derivatives of $f(\cdot, \cdot)$ evaluated at $(x^*(1), u^*(1))$. For example, consider the modified Euler scheme in Table 1:

$$\begin{aligned} \hat{q}(N) &= \tfrac{1}{2}(I - \tfrac{1}{2}hH(N))^{-1} e(N) \\ &= [\tfrac{1}{2}I + \tfrac{1}{4}hH(N) + O(h^2)]\, e(N), \end{aligned} \tag{16}$$

$$\begin{aligned} \hat{q}(N-1) &= (I - \tfrac{1}{2}hH(N-1))^{-1}[\tfrac{1}{2}(I + \tfrac{1}{2}hH(N))(I - \tfrac{1}{2}hH(N))^{-1} + \tfrac{1}{2}I]\, e(N) \\ &= [I + hH(N) + O(h^2)]\, e(N). \end{aligned} \tag{17}$$

Thus $|\hat{q}(N) - p^*(t_N)| = O(1)$ and $|\hat{q}(N-1) - p^*(t_{N-1})| = O(h^2)$. As Henrici [11] shows, the order of approximation for a multistep integration scheme is the minimum of the order of initial condition error and integration error. Since the modified Euler scheme is second order and $\hat{q}(N-1)$ is accurate to second order, $|\hat{q}(k) - p^*(t_k)| = O(h^2)$ for $k \leqq N-1$.

In general, the Taylor expansion yields $|\hat{q}(k) - p^*(t_k)| = O(h^s)$ for $k \approx N - \Delta$. The value of Δ is r for the Adams–Moulton and modified Euler scheme. As noted above, the value of s is 2 for the modified Euler scheme and 0 for the Milne schemes. For the 4-point Adams–Moulton scheme, $s = 4$. Since the multistep scheme has order b, $|\hat{q}(k) - p^*(t_k)| = O(h^m)$, where $m = \min\{s, b\}$ except for k satisfying $N - \Delta < k \leqq N$ where the error may be larger.

Now consider the difference $\delta(k) = \hat{q}(k) - q^*(k)$. Subtracting the equations for $\hat{q}$ and q^* yields

$$\sum_{j=0}^{r} a(j)\delta(k-j) = \sum_{j=0}^{r} b(j)\{hH(k-j)\delta(k-j) + h[f_x(x^*(t_{k-j}), u^*(t_{k-j}))^T - f_x(y^*(k-j), u^*(t_{k-j}))^T]q^*(k-j)\}. \tag{18}$$

Assuming (x^*, u^*) are bounded and $f_x(\cdot,\cdot)$ is continuous, the condition $|y^*(k) - x^*(t_k)| = O(h^b)$ implies that both $f_x(y^*(k), v^*(k))$ and $q^*(k)$ are bounded uniformly in h and k. Furthermore, if $f(\cdot,\cdot)$ has two continuous derivatives, then the last term in (18) is $hO(h^b) = O(h^{b+1})$. This last result together with the condition $\delta(k) = 0$ for $k > N$ and Lemma 5.6 in Henrici [11] proves that $|\delta(k)| = O(h^b)$ for $k \leqq N$; that is, adding a forcing term of size $O(h^{b+1})$ to the convergent difference scheme (18) contributes at most $O(h^b)$ to the solution $\delta(k)$ for $0 \leqq k \leqq N$.

Finally, $|q^*(k) - p^*(t_k)| \leqq |\hat{q}(k) - q^*(k)| + |\hat{q}(k) - p^*(t_k)| = O(h^b) + O(h^m) = O(h^m)$.

Step 3. *A bound for* $DJ(v^*)_k$. Two regions are considered: $N - \Delta < k \leqq N$ and $k \leqq N - \Delta$. In the second region, $|q^*(k) - p^*(t_k)| = O(h^m)$ by the results above. Since $|y^*(k) - x^*(t_k)| = O(h^b)$, the relation (15) implies that

$$|DJ(v^*)_k| \leqq hO(h^m) + hO(h^b) = O(h^{m+1}).$$

In the region $N - \Delta < k \leqq N$, we show that $DJ(v^*)_k = O(h^2)$. Let $(\mathbf{f}, \mathbf{p}, \mathbf{q})$ denote the vectors formed after removal of the zeroth components of (f, p, q). First note the following: (i) $|q_0^*(k)| = O(1)$ for $k \leqq N$ and (ii) $\mathbf{p}^*(t_k) = O(h) = \mathbf{q}^*(k)$ for $N - \Delta < k \leqq N$. The relation (i) follows since the $e(k-j)$ term in (11) makes only a nonzero contribution to $q(k)$ for $N - r \leqq k \leqq N$ while (ii) follows since $\mathbf{q}^*(k) = 0$ for $k > N$ and $\mathbf{p}^*(1) = 0$.

Inserting $p_0^*(t) = 1$ in (5), moving the p_0 term to the other side of the equation, and multiplying by $q_0^*(k)$, we get the following equalities:

$$\begin{aligned} q_0^*(k)(f_0)_u(x^*(t_k), u^*(t_k)) &= q_0^*(k)\mathbf{f}_u(x^*(t_k), u^*(t_k))^T\mathbf{p}^*(t_k) \\ &= O(h) \quad \text{for } N - \Delta < k \leqq N. \end{aligned} \tag{19}$$

The last equality in (19) follows by (i) and (ii) above and the bound on the gradient can now be established:

$$
\begin{aligned}
DJ(v^*)_k &= hf_u(y^*(k), v^*(k))^T q^*(k) \\
&= hq_0^*(k)(f_0)_u(x^*(t_k), u^*(t_k)) \\
&\quad + h\mathbf{f}_u(y^*(k), v^*(k))^T \mathbf{q}^*(k) + O(h^2) \\
&= O(h^2) \quad \text{for } N - \Delta < k \leqq N,
\end{aligned} \tag{20}
$$

where the second equality follows from the condition $|x^*(t_k) - y^*(k)| = O(h^b) \leqq O(h)$ and third equality is a consequence of (19) and (ii) above.

Step 4. *The Hessian* $D^2J(v)$. In many cases, the matrix $D^2J(v)$ is observed to have the following properties: (i) the diagonal entries are at least $O(h)$ in magnitude, (ii) the off-diagonal entries are at most $O(h^2)$, (iii) the off-diagonal entries in the last Δ rows and columns are $O(h^3)$, and (iv) $h|D^2J(v)^{-1}|$ is bounded uniformly in h.

These properties are illustrated for the following control problem in one dimension:

$$
\begin{aligned}
&\text{minimize} \int_0^1 [cx(t)^2 + u(t)^2]\, dt \\
&\text{subject to } \dot{x}(t) = ax(t) + bu(t), \qquad x(0) = z,
\end{aligned} \tag{21}
$$

where $c \geqq 0$, a and b are scalar constants. We consider only the Euler one-step method; however, the analysis is very similar for $r + 1$ point schemes since they can be expressed as one-step schemes involving vectors with r components.

For Euler's method, $J(v) = \sum_{k=0}^{N} h(cy(k)^2 + v(k)^2)$ where

$$
y(k+1) = y(k) + h(ay(k) + bv(k)) = (1 + ah)^k z + h \sum_{j=0}^{k} (1 + ah)^{k-j} bv(j).
$$

Since $y(k)$ is an affine function of $\{v(k)\}$, $J(v)$ is a quadratic in v, and hence $D^2J(v)$ is a constant matrix denoted H.

The diagonal of H is at least $2h$ since the $cy(k)^2$ term in $J(v)$ can only lead to a positive contribution to the diagonal while the $v(k)^2$ term contributes $2h$ to the kth diagonal entry. The (i, j)th off-diagonal entry in H arises from differentiation of $y(k)^2$ in the cost function:

$$
H_{ij} = \sum_{k > \max(i,j)} \frac{ch\partial^2 y(k)^2}{\partial v(i)\partial v(j)}. \tag{22}
$$

Since $y(k)$ is an affine function of the controls,

$$
\frac{\partial^2 y(k)^2}{\partial v(i)\partial v(j)} = \frac{2\partial y(k)}{\partial v(i)} \frac{\partial y(k)}{\partial v(j)} = 2h^2b^2(1 + ha)^{2k-j-i-2} \leqq 2e^{2a}h^2b^2 \tag{23}
$$

for $k > \max(i, j)$, where the last inequality in (23) follows from the relation $1 + ah \leqq e^{ah}$. Inserting the bound (23) into (22) yields

$$|H_{ij}| \leqq [N - \max(i,j)]ch2e^{2a}h^2b^2 \leqq O(h^2). \tag{24}$$

For either the last Δ rows or columns of H, $|N - \max(i, j)| \leqq \Delta$ and hence the expression in (24) is $O(h^3)$. This completes the proof of (i)–(iii) above; next the bound on $|H^{-1}|$ is proved.

Note that as $c \to 0$, the off-diagonal entries in H converge to zero while the diagonal entries are at least $2h$. Thus for c sufficiently small, H will be a diagonally dominant matrix and the ratio of the absolute sum of the off-diagonal entries in any row to the corresponding diagonal entry will be bounded by a constant $\rho < 1$ that is independent of N and the row number. The ℓ_∞- or maximum vector norm, $|y|$, generates a matrix norm $|M| = \max\{|My| : |y| = 1\}$. Since $1/|H^{-1}| = \min\{|Hy| : |y| = 1\}$, it can be proved that $|H^{-1}| \leqq 1/(1 - \rho)\delta h$, where $\delta = \min\{|H_{kk}|/h\}$: Suppose that z satisfies the conditions $|z| = 1$ and $|Hz| = 1/|H^{-1}|$. Let z_k be any component of z satisfying $z_k = 1$. Then

$$|(Hz)_k| \geqq |H_{kk}z_k| - |\sum_{j \neq k} H_{kj}z_j| \geqq |H_{kk}|(1 - \rho), \tag{25}$$

where the last inequality follows by the condition $|z_j| \leqq 1$ for all j.

For the control problem (21) above, $|H^{-1}| \leqq 1/2(1 - \rho)h$ since $\delta > 2$. The principal theorem of this section is now presented.

THEOREM 2.1. *Assume the following*: (A1) *The multistep scheme* (6) *is of order b*, (A2) *The costate integration scheme* (11) *is accurate to order* $m = \min\{s, b\}$ *for* $k \leqq N - \Delta$, *and* (A3) $D^2J(v)$ *satisfies the conditions* (i)–(iv) *in Step* 4 *uniformly in v. Then the solution* v^h *corresponding to the discrete approximation* (6) *satisfies* $|(v^h - v^*)_k| = O(h^l)$ *for* $k \leqq N - \Delta$ *and* $|(v^h - v^*)_k| = O(h)$ *for* $N - \Delta < k \leqq N$ *where* $\ell = \min\{3, s, b\}$.

Proof. By the development above, $v^* - v^h = D^2J(\mathbf{v})^{-1}DJ(v^*)$. By Steps 1, 2 and 3, $DJ(v^*) = f + g$ where $g_k = DJ(v^*)_k = O(h^{m+1})$ for $k \leqq N - \Delta$ and $f_k = DJ(v^*)_k = O(h^2)$ for $N - \Delta < k \leqq N$. Letting H denote the Hessian matrix, $|H^{-1}g| \leqq |H^{-1}|\,|g| = O(h^m)$ by (iv) above.

Now define $\hat{f}$ by $\hat{f}_k = f_k/H_{kk}$ and note that $H\hat{f} = f + \delta$ where $|\delta| = O(h^4)$ since the last Δ columns of H are $O(h^3)$. Also observe that $|\hat{f}| = O(h)$ and $\hat{f}_j = 0$ for $j \leqq N - \Delta$. Thus $|\hat{f} - H^{-1}f| = |H^{-1}\delta| \leqq |H^{-1}|\,|\delta| = O(h^3)$ by condition (iv) in Step 4. Since $\hat{f}_j = 0$ for $j \leqq N - \Delta$, the last inequality implies that $(H^{-1}f)_j = O(h^3)$ for $j \leqq N - \Delta$ while $|H^{-1}f| \leqq |H^{-1}|\,|f| = O(1/h)O(h^2 = O(h)$. Combining these bounds on $|H^{-1}g|$ and $|H^{-1}f|$ completes the proof of the theorem. □

The numerical experiments presented in § 4 indicate that the convergence rates in Theorem 2.1 are tight. To estimate u^* more accurately near $t = 1$, ignore the discrete controls near the right endpoint where the error is large, and pass a polynomial through the more accurate control parameters to approximate $u^*(1)$ by extrapolation.

Also in another paper of the author [10], a large truncation error is shown to produce a local error in the discrete approximation for discretizations yielding diagonally dominant, banded, linear systems.

3. One-step integration schemes. One-step integration schemes of the following form are studied:

$$y(j,k) = y(0,k) + h\sum_{m=0}^{j-1} a(j,m) f(y(m,k), v(m,k)) \quad \text{for } j = 1, 2, \cdots, r, \tag{26}$$

$$y(0, k+1) = y(r,k) \tag{27}$$

for $k = 0, 1, \cdots, N-1$. In (26), $y(j,k)$ and $v(j,k)$ are the discrete approximations to $x(t_k + h\xi_j)$ and $u(t_k + h\xi_j)$, respectively, where $h = 1/N$, $t_k = kh$, and the parameters $\{\xi_j\}$ and $\{a(j,m)\}$ are fixed constants that characterize the integration scheme. Some common one-step methods are given in Table 3, and are discussed in both Isaacson and Keller [12] and Gear [9]. Note that always $\xi_0 = 0$ and $\xi_r = 1$ so that $y(r,k) = y(0,k+1)$ is an approximation to $x(t_{k+1})$. The variables $y(j,k)$ for $0 < j < r$ are intermediate variables used to generate $y(r,k)$ and are usually of lower order accuracy than the final variable $y(r,k)$.

The following development is identical with the earlier development for multistep schemes; however, only second order convergence of the optimal discrete control is proved. For the examples of § 4, this is the convergence rate in the ℓ_∞-norm of the optimal discrete controls. These examples also indicate that the discrete costate and state convergence rate is higher than that for the discrete control, although the proof of this result in general is still an open question. Using the more accurate discrete state and costate, a higher order approximation to the optimal control u^* can be computed.

The following properties of the differential equation in (1) will be required: (i) there exists an optimal control u^*, a corresponding trajectory x^*, and, furthermore, the differential equation in (1) can be integrated for all controls u in some neighborhood of u^*, (ii) a solution v^h and a corresponding trajectory y^h exists to the discrete optimization problem, (iii) both the discrete and continuous necessary conditions hold, (iv) the integration scheme (26) approximates the solution to the corresponding differential equation to order b, and (v) the differential equation (3) obeys the standard theorem describing the effect of a perturbation in the initial condition on the trajectory; namely, if $x^1(\cdot)$ and $x^2(\cdot)$ are solutions to $\dot{x}(t) = f(x(t), u(t))$ that satisfy the conditions $x^1(s) = \rho^1$ and $x^2(s) = \rho^2$, respectively, for some $s \in [0,1]$, then $|x^1(t) - x^2(t)| = O(|\rho^1 - \rho^2|)$ for $t \in [0,1]$. If the cost functional is quadratic and strictly convex and the system dynamics are linear, then all the assumptions above are satisfied.

The discrete necessary conditions are given using the following notation: $\lambda(j,k)$ denotes the dual multiplier associated with (26) and $G(j,k) = f_x(y(j,k), v(j,k))^T$. In deriving these necessary conditions, $y(r,k)$ in (26) is replaced by $y(0,k+1)$ and thus (27) is eliminated along with any associated dual multiplier. Equations (28), (29), (30) and (31) below correspond to differentiating the Lagrangian with respect to $y(0,k+1)$, the intermediate variable $y(j,k)$ for $0 < j < r$, $v(j,k)$, and $y(0,N)$, respectively.

$$\sum_{j=1}^{r} \{[I + ha(j,0)G(r,k)]\lambda(j,k+1)\} - \lambda(r,k) = 0, \tag{28}$$

TABLE 3
One-step schemes

Name of Scheme	$a(i,j)$	Order of Scheme	Convergence Rate for $y^*(j,k)$	Convergence Rate for $q^*(j,k)$	Minimum of Rates for $y^*(j,k)$ and $q^*(j,k)$	Computed Convergence Rate for $u^h(r-1,k)$ (P1)	Computed Convergence Rate for $u^h(r-1,k)$ (P2)
Modified Euler	$1 \quad 0$ $\frac{1}{2} \quad \frac{1}{2}$	2	2 2	2 2	2 2	$1.73 \to 2$	$1.73 \to 2$
Kutta	$\frac{1}{2}$ $-1 \quad 2$ $\frac{1}{6} \quad \frac{2}{3} \quad \frac{1}{6}$	3	2 2 3	2 3 2	2 2 2	$2.90 \to 3$	$2.02 \to 2$
Runge–Kutta 1	$\frac{1}{2}$ $0 \quad \frac{1}{2}$ $0 \quad 0 \quad 1$ $\frac{1}{6} \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{6}$	4	2 2 3 4	2 2 4 3	2 2 3 3	$3.95 \to 4$	$2.97 \to 3$
Runge–Kutta 2	$\frac{1}{3}$ $-\frac{1}{3} \quad 1$ $1 \quad -1 \quad 1$ $\frac{1}{8} \quad \frac{3}{8} \quad \frac{3}{8} \quad \frac{1}{8}$	4	2 2 2 4	2 2 4 2	2 2 2 2	$4.06 \to 4$	$1.94 \to 2$

$$hG(j,k)\sum_{m=j+1}^{r}[a(m,j)\lambda(m,k)]-\lambda(j,k)=0 \quad \text{for } 1\leqq j<r,\quad 1\leqq k<N, \tag{29}$$

$$hf_u(y(j,k),v(j,k))^T\sum_{m=j+1}^{r}[a(m,j)\lambda(m,k)]=0 \quad \text{for } 0\leqq j<r,\quad 0\leqq k<N, \tag{30}$$

$$\lambda(r,N-1)^T=(1,0,0,\cdots,0)^T. \tag{31}$$

It should be remembered that (28)–(31) above only hold when $v(j,k)=v^h(j,k)$ and $y(j,k)=y^h(j,k)$; however, the "h" superscript is omitted below and all variables are understood to be optimal unless stated otherwise. Now change from the variable λ to the variable q given by

$$q(j,k)=\sum_{m=j+1}^{r}\frac{a(m,j)}{a(r,j)}\lambda(m,k) \quad \text{for } 0\leqq j<r. \tag{32}$$

It is assumed that $a(r,j)\neq 0$ for $0\leqq j<r$ so that $1/a(r,j)$ in (32) is defined. A case where $a(r,j)=0$ for some j is considered in the Appendix. It was discovered that as $a(r,j)\to 0$ for some j, the error in the optimal discrete control becomes infinite.

Replacing the summation in (29) by $a(r,j)q(j,k)$, multiplying the resulting equation by $a(j,l)/a(r,l)$, and summing from $j=l+1$ to $r-1$ yields:

$$\begin{aligned} h\sum_{j=l+1}^{r-1}a(j,l)\frac{a(r,j)}{a(r,l)}G(j,k)q(j,k)&=\sum_{j=l+1}^{r-1}\frac{a(j,l)}{a(r,l)}\lambda(j,k)\\ &=q(l,k)-\lambda(r,k)=q(l,k)-q(r-1,k)\end{aligned} \tag{33}$$

for $l=r-2,r-3,\cdots,0$. Summing (29) from $j=1$ to $r-1$ leads to

$$\sum_{j=1}^{r-1}\lambda(j,k)=h\sum_{j=1}^{r-1}[a(r,j)G(j,k)q(j,k)]. \tag{34}$$

Substituting (34) into (28) yields (36) below, while (35) is obtained by changing the indices in (33):

$$q(j,k)=q(r-1,k)+h\sum_{m=j+1}^{r-1}a(m,j)\frac{a(r,m)}{a(r,j)}G(m,k)q(m,k) \quad \text{for } j=r-2,r-3,\cdots,0, \tag{35}$$

$$q(r-1,k-1)=q(r-1,k)+h\sum_{m=0}^{r-1}[a(r,m)G(m,k)q(m,k)], \tag{36}$$

$$q(r-1,N-1)^T=(1,0,0,\cdots,0)^T. \tag{37}$$

The variable q and the equations (35)–(37) will be called the *discrete costate* and the *discrete costate equations*, respectively.

Note that the difference relations (35)–(37) run backward in time; that is, (37) gives the terminal condition for q while (35) specifies the intermediate variables $\{q(r-2,k),q(r-3,k),\cdots,q(0,k)\}$, and (36) gives the final variable $q(r-1,k-1)$.

Also observe that the one-step schemes in Table 3 satisfy the following identities:

$$a(m,j)a(r,m)/a(r,j) = a(r-1-j, r-1-m) \quad \text{for } 0 \leqq j \leqq r-2,\ j+1 \leqq m \leqq r-1, \tag{38}$$

$$a(r,j) = a(r, r-1-j) \quad \text{for } 0 \leqq j \leqq r-1. \tag{39}$$

Only a small subset of the class of all one-step schemes satisfy the symmetry conditions above; however, results given below indicate that when these symmetry conditions are violated, either there is a reduction in the anticipated convergence rate or the constant c involved in the error bound ch^{α} is so large, that the scheme is of little practical interest. (See § 4 and the Appendix.)

An alternative statement of (38) and (39) is the following: The difference relation describing the discrete costate is the same as the difference relation (26) except that the scheme (35)–(37) integrates the costate equation backward in time while (26) integrates the state equation forward in time. Substituting (38) and (39) into (35) and (36) yields the following expression for the discrete costate equations:

$$q(j,k) = q(r-1,k) + h \sum_{m=j+1}^{r-1} a(r-1-j, r-1-m)G(m,k)q(m,k) \quad \text{for } j = r-2, r-3, \cdots, -1, \tag{40}$$

$$q(r-1, k-1) = q(-1,k). \tag{41}$$

Also inserting (32) into (30) yields the discrete control minimum principle:

$$ha(r,j)f_u(y(j,k), v(j,k))^T q(j,k) = 0. \tag{42}$$

Repeating the multistep development, Steps 1 and 4 are essentially unaltered, and all that remains is to bound the gradient,

$$DJ(v^*)_{jk} = ha(r,j)f_u(y^*(j,k), v^*(j,k))^T q^*(j,k), \tag{43}$$

where v^* is defined by $v^*(j,k) = u^*(t_k + h\xi_j)$, y^* is the solution to (26) with $v = v^*$, and q^* is given by (35)–(37) with $(y, v) = (y^*, v^*)$. Since

$$f_u(x^*(t_k + h\xi_j), u^*(t_k + h\xi_j))^T p^*(t_k + h\xi_j) = 0,$$

$$|DJ(v^*)_{jk}| \leqq ch|y^*(j,k) - x^*(t_k + h\xi_j)| + ch|q^*(j,k) - p^*(t_k + h\xi_j)|. \tag{44}$$

The error in the discrete state $y^*(j,k)$ is known for the standard integration schemes and is given in Table 3, while the error in the discrete costate $q^*(j,k)$ follows from Lemma 3.1 below.

The analysis of the error in the discrete costate variable is complicated by the presence of $y(j,k)$ in $G(j,k) = f_x(y(j,k), v(j,k))^T$; thus the discrete costate equation is implicit. First observe that if the one-step procedure (26) is applied to the linear differential equation $\dot{z}(t) = A(t)z(t)$, then the intermediate variables $z(1,k), \cdots, z(r-1,k)$ can be eliminated, and the scheme can be expressed as $z(0, k+1) =$

$F(A(0,k),\cdots,A(r-1,k))z(0,k)$ where $A(j,k)=A(t_k+h\xi_j)$. Another interesting property of the schemes in Table 3 is the following:

$$F(B_1,B_2,\cdots,B_r)=F(B_r^T,B_{r-1}^T,\cdots,B_1^T)^T. \tag{45}$$

LEMMA 3.1. *Suppose that* (38), (39) *and* (45) *hold and that the one-step scheme* (26) *is of order b. Let u be some fixed control (not necessarily optimal) and let x and p be the corresponding state and costate variables generated by* (3) *and* (4); *let* $\{y(j,k)\}$ *be the discrete state generated by* (26) *using the discrete controls* $v(j,k)=u(t_k+h\xi_j)$, *and let* $\{q(j,k)\}$ *be the corresponding discrete costate generated by* (40) *and* (41). *Then* $|p(t_{k+1})-q(r-1,k)|=O(h^b)$.

Proof. It will be shown that the matrix $F(G(r-1,k),G(r-2,k),\cdots,G(0,k))$ approximates the transition matrix associated with the linear costate equation on the interval $[t_k,t_{k+1}]$ to $O(h^{b+1})$. The global error in the discrete costate then follows by adding up the local errors.

Consider the coupled system of differential equations

$$\dot{x}^k(t)=f(x^k(t),u(t)),\qquad x^k(t_k)=\rho_x, \tag{46}$$

$$\dot{\pi}^k(t)=f_x(x^k(t),u(t))\pi^k(t),\qquad \pi^k(t_k)=\rho_\pi. \tag{47}$$

Use the one-step scheme given in (26) to approximate the solution to this system, and let $x^h(\cdot,\cdot)$ and $\pi^h(\cdot,\cdot)$ denote the discrete variables corresponding to the continuous variables $x^k(\cdot)$ and $\pi^k(\cdot)$, respectively. Since π^k does not appear in (46), $x^h(j,k)$ is the solution $y(j,k)$ of (26) corresponding to the initial condition $y(0,k)=\rho_x$. The discrete approximation to (47), on the other hand, can be expressed as

$$\pi^h(0,k+1)=F(G^h(0,k)^T,\cdots,G^h(r-1,k)^T)\rho_\pi, \tag{48}$$

where $G^h(j,k)=f_x(x^h(j,k),v(j,k))^T$. Since the one-step scheme (26) is of order b, $|\pi^k(t_{k+1})-\pi^h(0,k+1)|\leqq ch^{b+1}$, where c depends on the derivatives of $f(\cdot,\cdot)$ and $u(\cdot)$ to order $b+1$, $|\rho_x|$ and $|\rho_\pi|$. Choose c large enough that the error in $\pi^h(0,k+1)$ is bounded by ch^{b+1} for all ρ_x and ρ_π satisfying the following bounds: $|\rho_\pi|\leqq 1$ and $|\rho_x|\leqq\max\{|x(t)|+\varepsilon: t\in[0,1]\}$ for some fixed $\varepsilon>0$ where $x(\cdot)$ is the trajectory corresponding to the control $u(\cdot)$ in the lemma's statement.

Now let $H(A,t,s)$ denote the transition matrix or fundamental matrix for the linear system $\dot{\eta}(t)=A(t)\eta(t)$. That is, $H(A,t,s)$ is a matrix satisfying $(d/dt)H(A,t,s)=A(t)H(A,t,s)$ and $H(A,s,s)=I$, and $H(A,\cdot,s)$ has the property that the solution to the linear differential equation satisfies $\eta(t)=H(A,t,s)\eta(s)$. Thus by (47),

$$\pi^k(t_{k+1})=H(f_x(x^k,u),t_k+h,t_k)\rho_\pi. \tag{49}$$

By comparing (49) with (48), we see

$$F(G^h(0,k)^T,\cdots,G^h(r-1,k)^T)=H(f_x(x^k,u),t_k+h,t_k)+O(h^{b+1}) \tag{50}$$

whenever the initial condition ρ_x satisfies the bound given above.

Taking the transpose of (50) and applying the symmetry condition (45), we see

$$F(G^h(r-1,k),\cdots,G^h(0,k)) = H(f_x(x^k,u),t_k+h,t_k)^T + O(h^{b+1}). \tag{51}$$

Transition matrices obey the identity $H(A,t,s)^T = H(-A^T,s,t)$, which can be proved by verifying that both matrices satisfy the differential equation $[\dot{M}(t) = M(t)A(t)^T, M(s) = I]$ whose solution is unique. Thus (51) becomes

$$F(G^h(r-1,k),\cdots,G^h(0,k)) = H(-f_x(x^k,u)^T,t_k,t_k+h) + O(h^{b+1}). \tag{52}$$

Let $\hat{x}^k(\cdot)$ denote the solution to (46) for the initial condition $\rho_x = y(0,k)$ where $y(0,k)$ is generated by (26) with the initial condition $y(0,0) = z$. Recall the assumption (v) of initial condition stability for the state equation; hence $|\hat{x}^k(t) - x(t)| = O(|y(0,k) - x(t_k)|)$ for $t \in [t_k, 1]$. Since $y(0,k)$ is accurate to $O(h^b)$, $|\hat{x}^k(t) - x(t)| = O(h^b)$ and, consequently, $|f_x(\hat{x}^k(t),u(t)) - f_x(x(t),u(t))| = O(h^b)$.

It can be shown that transition matrices possess the following property: $|H(A_1,t,s) - H(A_2,t,s)| \leqq c\|A_1 - A_2\|\,|t-s|$ where $\|A\| = \max\{|A(\tau)| : \tau \in [s,t]\}$ and c depends on $\|A_1\| + \|A_2\|$. Thus

$$|H(-f_x(\hat{x}^k,u)^T,t_k,t_k+h) - H(-f_x(x,u)^T,t_k,t_k+h| = O(h^{b+1}). \tag{53}$$

If $\rho_x = y(0,k)$, then $G^h(j,k) = G(j,k) = f_x(y(j,k),v(j,k))^T$ and hence (52) and (53) imply that

$$F(G(r-1,k),\cdots,G(0,k)) = H(-f_x(x,u)^T,t_k,t_k+h) + O(h^{b+1}). \tag{54}$$

The continuous and the discrete costate variables satisfy the following relations:

$$p(t_k) = H(-f_x(x,u)^T,t_k,t_k+h)p(t_{k+1}), \tag{55}$$

$$q(r-1,k-1) = F(G(r-1,k),\cdots,G(0,k))q(r-1,k) \tag{56}$$

$$= H(-f_x(x,u)^T,t_k,t_k+h)q(r-1,k) + O(h^{b+1}), \tag{57}$$

where (57) follows by (54). Subtracting (56) from (55), taking the absolute value of both sides, defining $e(k) = |p(t_k) - q(r-1,k-1)|$, and using the relation $H(A,t,s) = I + O(|t-s|)$ yields the following:

$$e(k) \leqq (1 + O(h))e(k+1) + O(h^{b+1}) \tag{58}$$

$$\leqq (1 + O(h))^{N-k}e(N) + \sum_{j=k+1}^{N} (1 + O(h))^{N-j}O(h^{b+1}). \tag{59}$$

Since $(1 + O(h))^{N-j} \leqq \exp[O(h)(N-j)] = O(1)$ and $e(N) = 0$, (59) implies that $e(k) = O(h^b)$ as stated in the lemma. □

Using the result that $q^*(r-1,k)$ is accurate to order b, it can be shown for the schemes in Table 3 that the order of the error $|p^*(t_{k+1} - h\xi_{r-j-1}) - q^*(j,k)| =$ the order of the error $|x^*(t_k + h\xi_{r-j-1}) - y^*(r-j-1,k)|$ for $0 \leqq j \leqq r-1$. That is, the jth intermediate variables in the discrete costate and state equations have the same order of accuracy. (Recall that $q(r-2,k)$ and $y(1,k)$ are the first

intermediate variables computed in the discrete costate equation and state equation, respectively.)

By (44), the convergence rate of $DJ(v^*)_{jk}/h$ equals the minimum of the convergence rates of $y^*(j,k)$ and $q^*(j,k)$, and this minimum is given in Table 3. The analogue of our earlier theorem for the convergence of multistep schemes follows:

THEOREM 3.1. *If* $h|D^2J(v)^{-1}|$ *is bounded uniformly in* h *and* v *and* $|DJ(v^*)_{jk}| = O(h^3)$ *for all* j, k, *then* $|v^*(j,k) - v^h(j,k)| = O(h^2)$ *for all* j, k.

Proof. $|v^* - v^h| \leqq |D^2J(v)^{-1}|\,|DJ(v^*)| = O(1/h)O(h^3) = O(h^3)$. □

For the schemes in Table 3, the numerical examples in the next section indicate that the convergence rate for the optimal discrete control is second order. Hence the convergence rate given in Theorem 3.1 appears to be tight. However, closer inspection of the numerical results revealed that $\{y^h(r,k)\}$ and $\{q^h(r-1,k)\}$ are accurate to order b for a bth order integration scheme, and if $v(j,k)$ is eliminated from the discrete state and costate equations by using (42) to express the control in terms of $y(j,k)$ and $q(j,k)$, then the resulting system is a bth order implicit approximation to the two-point boundary value problem described after (3)–(5).

This last result seems very intuitive, but is currently an unsolved problem; Lemma 3.1 proved that if u is *any* control, then the discrete state and costate equations represent a bth order implicit scheme for integrating (3) and (4). It must now be proved that if $u(t)$ and $v(j,k)$ are required to satisfy the minimum principles (5) and (42), respectively, then the resulting discrete state and costate equations are bth order implicit schemes for integrating the continuous two-point boundary value problem.

Assuming that $y^h(r,k)$ and $q^k(r-1,k)$ are accurate to $O(h^b)$, a bth order estimate of $u^*(t_{k+1})$ is given by the solution to the equation

$$f_u(y^h(r,k),u)^T q^h(r-1,k) = 0.$$

4. Numerical examples and discussion of results. The convergence properties of the integration schemes given in Tables 1 and 3 were studied for the following linear one-dimensional control problems with quadratic cost functionals:

$$\text{(P1)} \qquad \text{minimize}\left\{\int_0^1 [.5u(t)^2 + x(t)^2]\,dt : \dot{x}(t) = .5x(t) + u(t),\ x(0) = 1\right\},$$

$$\text{(P2)} \qquad \text{minimize}\left\{\int_0^1 [.625x(t)^2 + .5x(t)u(t) + .5u(t)^2]\,dt : \dot{x}(t) = .5x(t) + u(t),\ x(0) = 1\right\}.$$

The optimal control for these regulator problems can be expressed in the feedback form $u^*(t) = -K_1(t)x^*(t)$ for (P1) and $u^*(t) = -(K_2(t) + .5)x^*(t)$ for (P2) where K_1 and K_2 are solutions to the corresponding Riccati equations (see Brockett [1]):

$$\dot{K}_1(t) = -K_1(t) + K_1(t)^2 - 2, \qquad K_1(1) = 0,$$

$$\dot{K}_2(t) = K_2(t)^2 - 1, \qquad K_2(1) = 0.$$

These solutions are $K_1(t) = (2 + ae^{3t})/(1 - ae^{3t})$ where $a = -2/e^3$ and $K_2(t) = \tanh(1 - t)$. Inserting the feedback expression for u^* into the system dynamics

yields an equation for x^* while u^* is computed from x^* by the feedback law; the optimal solutions to (P1) and (P2) are the following:

$$\text{(S1)} \qquad x^*(t) = (ae^{3t} - 1)/e^{3t/2}(a - 1), \qquad u^*(t) = (2 + ae^{3t})/e^{3t/2}(a - 1),$$

$$\text{(S2)} \qquad x^*(t) = \cosh(1)/\cosh(1 - t),$$

$$u^*(t) = -[\tanh(1 - t) + .5]\cosh(1)/\cosh(1 - t).$$

The control problems (P1) and (P2) are quite similar—their system dynamics are identical, and the only major difference in their cost functionals is the $.5x(t)u(t)$ term present in (P2). The optimal controls are plotted in Fig. 1 and also appear very similar. However, as will be seen below, the convergence rate for the optimal discrete control is quite different for one-step approximations to (P1) and (P2).

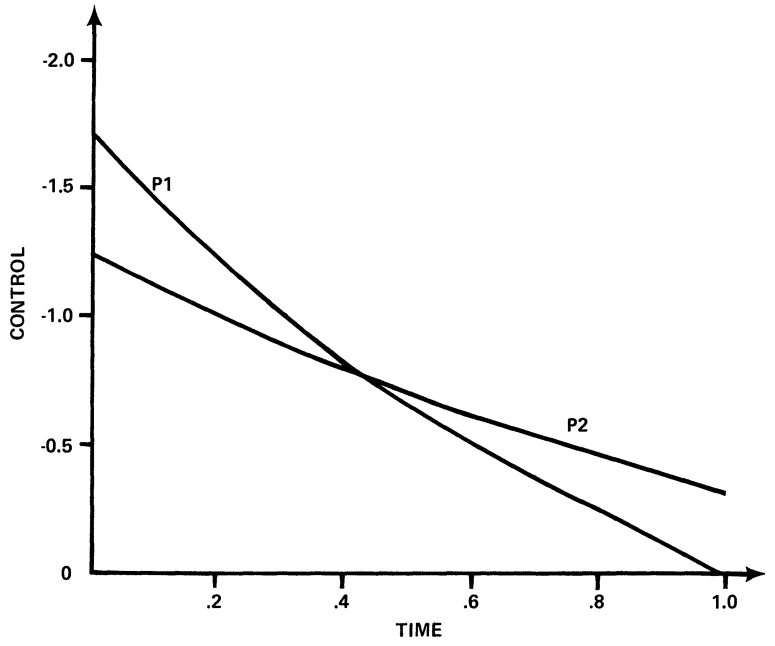

FIG. 1. *The optimal controls for the test problems*

One-step schemes. The optimal solution to the discrete one-step approximation was computed as follows: Using (42), we expressed $v(j, k)$ in terms of $y(j, k)$ and $q(j, k)$ and the result was inserted into the discrete state and costate equations (26) and (35)–(36) to obtain $2r$ relations for

$$\{y(0, k), \cdots, y(r, k), q(r - 1, k - 1), q(0, k), \cdots, q(r - 1, k)\}.$$

Then the computer program determined the matrix A such that $z(k) = Az(k - 1)$ where $z(k)^T = [y(r, k)^T, q(r - 1, k)^T]$. Since $q(r - 1, N - 1) = 0$ and $y(r, -1) = y(0, 0) = 1$ are known, $q(r - 1, -1)$ can be computed from the relation $z(N - 1) = A^N z(-1)$. Since $z(-1)$ is now known, $z(k) = A^{k+1}z(-1)$ is also determined and the optimal discrete controls can be computed from (42).

For more complicated nonlinear problems, the discrete optimization problem can be solved by the conjugate gradient method where the gradient with respect to the controls $\{v(j,k)\}$ is given by (42).

The convergence rate for the ℓ_∞-error in the optimal discrete controls $\{v^h(j,k)\}$ is second order for all the schemes studied so the rate given in Theorem 3.1 appears to be tight. However, it was observed that the controls $\{v^h(r-1,k)\}$ were accurate to a much higher order as shown by the convergence rates in Table 3 and Figs. 2 and 3. These rates were determined by computing the ℓ_∞-error of $\{v^h(r-1,k)\}$ using $1/h = 10, 20, 40, 80, 120, 160$ and then fitting a least squares line to the graph of log (error) versus log (h). The slope of this line shown in Figs. 2 and 3 is the convergence rate.

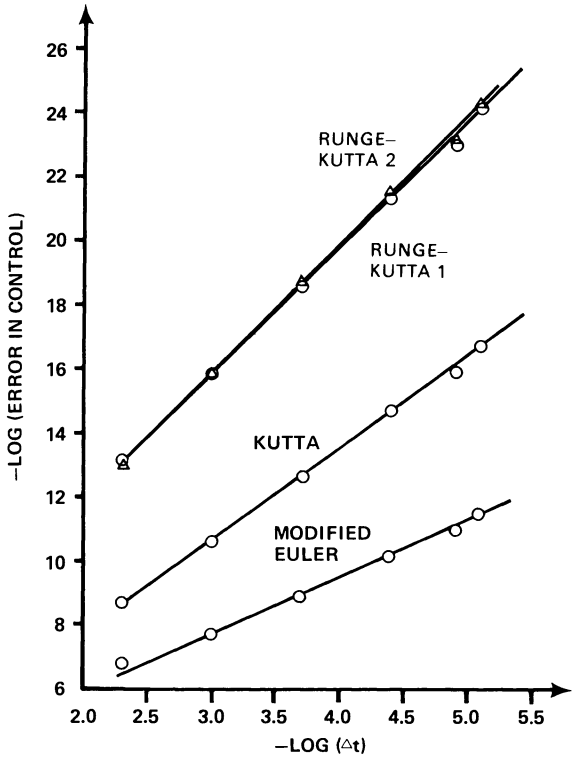

FIG. 2. *Maximum norm error in the discrete controls* $\{v^h(r-1,k)\}$ *for* (P1)*—one-step methods*

Notice that for (P1), the order of accuracy of the controls $\{v^h(r-1,k)\}$ is the same as that of the corresponding integration scheme while the accuracy of the controls decreases in (P2). Closer study of the discrete variables reveals that in both (P1) and (P2), $\{y^h(r,k)\}$ and $\{q^h(r-1,k)\}$ converge at the same rate as the order of the integration scheme. Hence the intermediate variables $y^h(j,k)$ and $q^h(j,k)$ converge at the same rate as $y^*(j,k)$ and $q^*(j,k)$, respectively. Furthermore, using relation (44), the convergence rate of $v^h(j,k)$ must equal the minimum of the orders of $y^h(j,k)$ and $q^h(j,k)$.

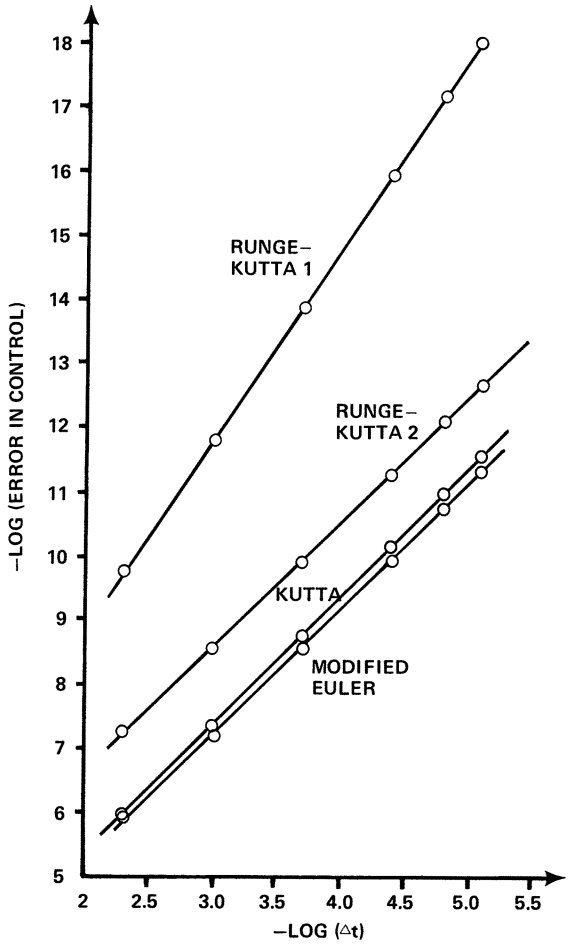

FIG. 3. *Maximum norm error in the discrete controls* $\{v^h(r-1,k)\}$ *for* (P2)*—one-step methods*

The difference in the convergence rate of $\{v^h(r-1,k)\}$ for (P1) and (P2) can now be explained. In (P1) f_u is not a function of the state variable so that the state term in (44) is not present and hence $|v^h(r-1,k) - u^*(t_{k+1})| \leqq c|q^h(r-1,k) - p^*(t_{k+1})|$. In (P2), however, f_u depends on the state and both terms in (44) are present. Since the convergence rate for $y^h(r-1,k)$ is less than the rate for the costate $q^h(r-1,k)$, a reduction in the control accuracy should be expected in (P2).

As noted earlier, a higher order estimate of $u^*(t_{k+1})$ in (P2) is given by the solution u to $f_u(y(r,k),u)^T q(r-1,k) = 0$. For (P2) this reduces to $u = -q(r-1,k) - y(r,k)/2$.

Since the optimal discrete state and costate converge at a higher rate than the discrete control for the schemes in Table 3, the possibility of obtaining a more accurate discrete control by a judicious choice of the integration coefficients was examined; hence the auxiliary computation given above to obtain an improved estimate of u^* is unnecessary.

Note that for the modified Euler scheme, the discrete parameters $\{y^h(r,k)\}$, $\{v^h(r-1,k)\}$ and $\{q^h(r-1,k)\}$ are all second order accurate so that the discrete control accuracy is the best that could be expected.

For the Kutta scheme, however, the optimal discrete control is only second order accurate while the discrete state and costate are third order accurate. If $y^h(r-1,k)$ can be made third order accurate by an appropriate choice of the integration coefficients and the third order accuracy of $q^h(r-1,k)$ maintained, then $v^h(r-1,k)$ will also be third order accurate.

There is a two parameter class of third order one-step schemes of the form (26) with $r=3$. These are developed by Gear [9, p. 34] and in terms of the parameters β and γ are given by:

$$
\begin{aligned}
&a(3,2) = (3\gamma-2)/(6\beta(\gamma-\beta)), && a(2,1) = 1/(6\gamma a(3,2)),\\
&a(1,0) = \gamma, && a(3,1) = (3\beta-2)/(6\gamma(\beta-\gamma)),\\
&a(3,0) = 1 - a(3,2) - a(3,1), && a(2,0) = \beta - 1/(6\gamma a(3,2)).
\end{aligned} \tag{60}
$$

Since the error in $y(2,k)$ equals $(\beta-\frac{1}{2})O(h^2)+O(h^3)$, $y(2,k)$ is accurate to $O(h^3)$ when $\beta=\frac{1}{2}$. For all choices of the parameter γ, however, it can be shown that the symmetry conditions (38)–(39) are violated. Also by numerical experiments using problems (P1) and (P2), the discrete state and costate are second order accurate for all choices of γ. Examining the coefficients describing the discrete costate integration scheme (35)–(36), we find that they yield a second order integration method. These observations seem to indicate the following: (i) the order of accuracy of the optimal discrete state and costate variables is the minimum of the order associated with the state and costate integration coefficients, and (ii) if $r=3$, no choice of the integration coefficients will make $v^h(r-1,k)$ third order accurate.

Similarly for fourth order schemes with $r=4$, requiring $y(r-1,k)$ to be fourth order accurate leads to an inconsistent system of 11 equations in 10 unknowns. Thus the scheme Runge–Kutta 1 in Table 3 appears to have the best possible accuracy for the discrete control $v^h(r-1,k)$.

Multistep schemes. Next the test problems were approximated by the multistep schemes of Table 1, and the discrete optimization problem was solved by the conjugate gradient algorithm where the gradient was given by the left side of (13). The convergence rates reported in Table 1 were determined by plotting log (error in optimal discrete control at $t=.5$) versus log (h) using $1/h = 10, 20, 40, 80, 120, 160$ and approximating the graph using a least squares line. As noted in Theorem 2.1, the error in the discrete control is concentrated at $t=1$ so the point $t=.5$ was chosen for measuring the convergence rates to avoid the boundary layer in the error at the endpoint.

Let s and b be as defined in § 2. For both the Milne schemes in Table 1, $s=0$, so by Theorem 2.1, the error in v^h is at most $O(1)$. Numerically it was observed that v^h neither converged nor diverged, but oscillated about u^*. For the modified Euler scheme, $s=b=2<3$, so by Theorem 2.1, v^h converges to v^* at rate 2 except near $t=1$. This was exactly the convergence rate observed numerically and also shown

in Fig. 4. Finally for the Adams–Moulton scheme, $s = b = 4 > 3$, so Theorem 2.1 implies third order convergence of v^h and again this was exactly the convergence rate observed numerically and shown in Fig. 4.

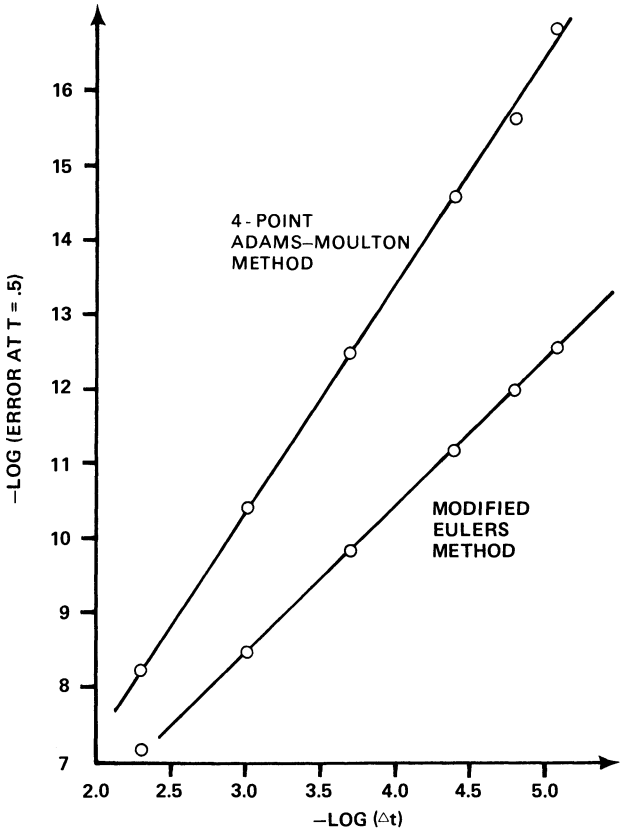

FIG. 4. *Error in the discrete control* $\{v^h(N/2)\}$ *for* (P1)*—multistep methods*

Next the problem of finding the most efficient integration scheme for achieving the maximum convergence rate of 3 was examined. For any value of the parameter γ satisfying $-1 \leqq \gamma < 1$, the following multistep scheme is at least third order (for $\gamma = -1$, the scheme is fourth order):

$$\begin{aligned} &a(2) = 1, && a(1) = -1 - \gamma, && a(0) = \gamma, \\ &b(2) = (5 + \gamma)/12, && b(1) = 2(1 - \gamma)/3, && b(0) = -(1 + 5\gamma)/12. \end{aligned} \tag{61}$$

This class of third order schemes is given in Lambert [15, p. 42].

Convergence of v^h to v^* requires convergence of $q_0^h(k)$ to $p_0^* \equiv 1$. Solving the difference equation for $q_0(\cdot)$ yields:

$$q_0(N) = \frac{(5 + \gamma)}{12}, \qquad q_0(k) = 1 + \frac{(\gamma - 1)^2}{12}\gamma^{N-1-k}. \tag{62}$$

(Note that when $\gamma = 0$, define $0^0 = 1$ for the solution above to be correct.) For $-1 < \gamma < 1$, the error $|p_0^*(t_k) - q_0^h(k)| = (\gamma - 1)^2\gamma^{N-1-k}/12$ decays geometrically as k decreases.

Figure 5 plots log (error in $v^h(k)$) as a function of k for various choices for γ. Note that the discrete control error possesses two distinct characteristics: near $t = 1$ the error decays geometrically (the graph is linear) and far from $t = 1$, the $O(h^3)$ error predominates and the discrete control error levels off after some oscillations. As $\gamma \to \pm 1$, the integration schemes approach the unstable region $\gamma \geqq 1$ or $\gamma < -1$ and the oscillations at the interface of the $O(h^3)$ error and the geometrically decaying error increase. The case $\gamma = -1$ corresponds to a Milne scheme and since the γ^{N-1-k} term in (62) does not decay, the error in $q_0(k)$ is always $O(1)$ and v^h does not converge to v^*.

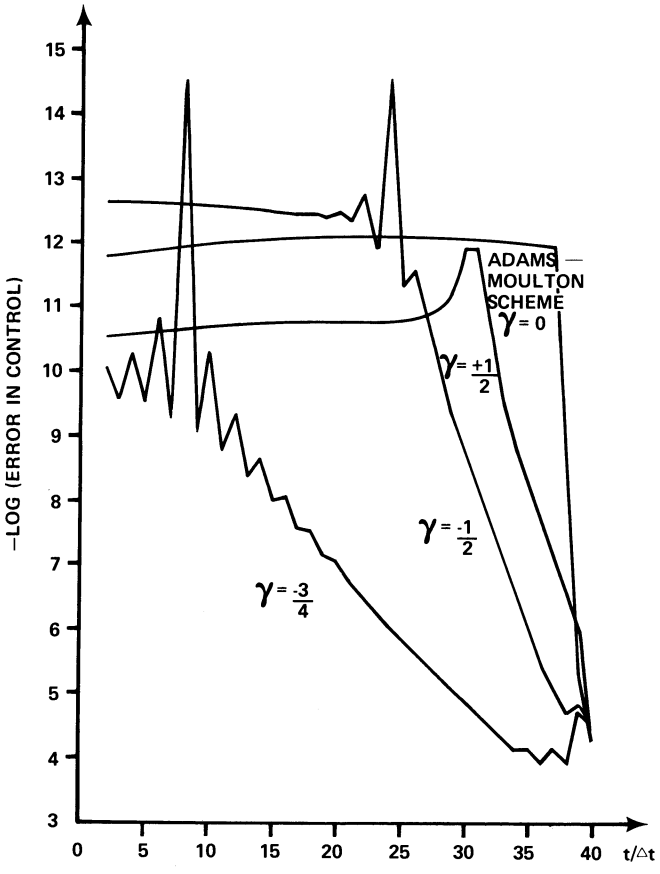

FIG. 5. *Pointwise error in the discrete controls for* (P1) *using several* 3*-point schemes*

Note that as γ approaches -1, the order of the integration scheme approaches 4, and hence the contribution to the discrete control error arising from the integration error decreases. On the other hand, as γ approaches -1, the decay rate for the geometric error decreases, and hence the width of the boundary layer near

$t = 1$ increases. Thus the most accurate scheme corresponds to either $\gamma = 0$ (3-point Adams–Moulton) or γ slightly negative. The choice of $\gamma = 0$ reduces the boundary layer at $t = 1$ to only a few grid intervals while the choice of γ slightly negative widens the boundary layer but reduces the error far from $t = 1$. (See Fig. 5.)

Comparison of one-step and multistep schemes. Based on the results in the Appendix and an analysis of truncation errors, the modified Euler scheme in Table 3 is both the only symmetric and the most accurate one-step method. Similarly the Euler scheme in Table 1 is the unique second order multistep scheme with $r = 1$. Comparing the two Euler schemes, the one-step method has the advantage of being explicit, while the multistep scheme involves half the number of unknown controls per grid interval.

In the class of third order schemes, the one-step procedures involve three unknown controls per grid interval compared to only one unknown control for the multistep procedures. If the implicit multistep equations are not too difficult to solve, then the Adams–Moulton scheme is probably the most efficient third order method. Since there appear to be no multistep methods higher than third order, the one-step schemes win by default assuming the conjecture on the accuracy of $y^h(r, k)$ and $q^h(r - 1, k)$ holds in general.

One case where the one-step procedures would have an advantage over multistep schemes are situations where there is a discontinuity in the data defining the control problem at some fixed times. The error in one-step integration methods depends on the derivatives of u^* between the grid points while the error in multistep integration procedures depends on the global derivatives of u^*. Hence by placing grid points wherever there is a discontinuity in the data, the error in the discrete approximation caused by the discontinuity can be eliminated with a one-step scheme. Similarly the multistep method can be restarted at these points of discontinuity, but this increases programming complexity.

Appendix. The case $a(r, j) = 0$. If $a(r, j) = 0$ for some j, then $q(j, k)$ in (32) is no longer defined. To study the effect of the vanishing of an integration coefficient, consider the class of second order one-step schemes with $r = 2$. In terms of the parameter γ, the coefficients for these second order schemes are the following:

$$\text{(A.1)} \qquad a(1,0) = 1/(2\gamma), \qquad a(2,0) = 1 - \gamma, \qquad a(2,1) = \gamma.$$

The value $\gamma = \frac{1}{2}$ gives the modified Euler scheme in Table 3 while the value $\gamma = 1$ corresponds to a common Euler method with $a(2, 0) = 0$.

Using (35) and (36), it is possible to compute the integration coefficients for the costate equation:

$$\text{(A.2)} \qquad a(1,0) = 1/(2(1 - \gamma)), \qquad a(2,0) = \gamma, \qquad a(2,1) = 1 - \gamma.$$

Note that $\gamma = \frac{1}{2}$ is the only case where the schemes (A.1) and (A.2) are identical and hence satisfy the symmetry conditions (38)–(39); nonetheless, the scheme (A.2) is accurate to second order for all values of γ, although the $O(h^3)$ truncation error involves a $1/(1 - \gamma)$ term. Hence as $\gamma \to 1$ and h *remains fixed*, the error of the

integration scheme (A.2) becomes infinite. Numerically it is found that the error in the optimal solution to the discrete approximation to (P2) is 1000 times bigger for $\gamma = .999$ than the error for $\gamma = .5$. Thus when $r = 2$, the effect of an integration coefficient $a(r, j)$ approaching zero is disastrous.

REFERENCES

[1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[2] B. M. BUDAK AND E. M. BERKOVICH, *On the approximation of extremal problems*, Ž. Vyčisl. Mat. i Mat. Fiz., 11 (1971), pp. 580–596.

[3] B. M. BUDAK, E. M. BERKOVICH AND E. N. SOLOV'EVA, *Difference approximations in optimal control problems*, SIAM J. Control, 7 (1969), pp. 18–31.

[4] ———, *The convergence of difference approximations in optimal control problems*, Ž. Vyčisl. Mat. i Mat. Fiz., 9 (1969), pp. 522–547.

[5] M. D. CANNON, C. D. CULLUM AND E. POLAK, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.

[6] J. CULLUM, *Finite-dimensional approximations of state-constrained continuous optimal control problems*, SIAM J. Control, 10 (1972), pp. 649–670.

[7] ———, *Discrete approximations to continuous optimal control problems*, Ibid., 7 (1969), pp. 32–49.

[8] ———, *An explicit procedure for discretizing continuous optimal control problems*, J. Optimization Theory Appl., 8 (1971), pp. 15–34.

[9] C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[10] W. W. HAGER AND G. STRANG, *Free boundaries and finite elements in one dimension*. Math. Comp., 29 (1975), pp. 1020–1031.

[11] P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.

[12] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[13] R. KLESSIG AND E. POLAK, *An adaptive precision gradient method for optimal control*, SIAM J. Control, 11 (1973), pp. 80–93.

[14] H. B. KELLER, *Numerical Methods for Two-Point Boundary-Value Problems*, Blaisdell, Waltham, Mass., 1968.

[15] J. D. LAMBERT, *Computational Methods in Ordinary Differential Equations*, John Wiley, London, 1973.

[16] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1965.

[17] S. M. ROBERTS AND J. S. SHIPMAN, *Two-Point Boundary-Value Problems: Shooting Methods*, American Elsevier, New York, 1972.